Progress and Challenges in Evaluating Tools for Sensemaking

Jean Scholtz
Pacific Northwest National Laboratory
Richland, WA
Jean.scholtz@pnl.gov

ABSTRACT
In this paper we discuss current work and challenges for the development of metrics to evaluate software designed to help analysts with sensemaking activities. While much of the work we describe has been done in the context of intelligence analysis, we are also concerned with the general applicability of metrics and evaluation methodologies for other analytic domains.

Author Keywords
Evaluation, metrics, intelligence analysis, sensemaking.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
The Pacific Northwest National Laboratory has been funded by the Department of Homeland Security (DHS) to run the National Visualization and Analysis Center (NVAC). A number of activities occur in this program. The Regional Visualization and Analysis Centers (RVACs) are engaged in research to help solve problems that analysts in DHS encounter. Other activities within NVAC include: development of visual analysis software, customizing and deploying software at various DHS agencies, and developing metrics and evaluation methodologies for the field of visual analysis. The latter effort includes: designing and conducting the Visual Analytics Science and Technology (VAST) 2006 and 2007 contests, developing metrics for the various software deployments and consulting with the RVACs on user-centered evaluation studies of their research (http://www.cs.umd.edu/hcil/VASTcontest07/).

The goal of the evaluation efforts is to devise a number of user-centered metrics for researchers in visual analysis to use in evaluating their work. We use the term user-centered to describe metrics that are focused on the utility the software provided to the analysts. While performance metrics are certainly necessary, they are not sufficient to guarantee utility. Moreover, user-centered metrics are one way to convey to researchers the priorities of the analysts. This is extremely valuable as academics most often have little access to analysts.

Information analysts carry out complex tasks. Complex tasks differ from well structured tasks in many ways, including [6,7]:

- Information overload is endemic. People have to locate the relevant information/sources and focus only on those.
- Data analysis and recursive decision-making are cognitively very burdensome; people have little cognitive workload available for dealing with unusable interfaces.
- Information is often incomplete or even unreliable.
- In some domains, there may be no way to know if the result one gets is right or wrong.
- In many domains time may be critical. Good decisions made too late are bad decisions.
- Domain experts may not be computer experts. The demands of their work may make it difficult for them to put much time or effort into the learning curve of new programs.
- Software systems for complex problem solving often consist of several components and visualizations. We must evaluate the individual components and/or visualizations as well as the entire system.
- Information for many of these tasks is not static but changes over time. This dynamic nature increases the complexity of the task.
- Complex problems often require multiple domain experts to collaborate. This presents a challenge in combining multiple perspectives.
In the rest of the paper we discuss various activities that are helping to define user-centered metrics and the challenges that are still ahead.

**THE VAST CONTESTS**

We have to date conducted two VAST contests: VAST 2006 and VAST 2007. Material developed for these contests include:

- A data set containing group truth and some background information about the task
- Metrics for evaluation
- Specifications on materials to submit

Participants are given approximately five months to assemble and submit their entry. Then a panel of judges consisting of experts in visualization and usability and experts in intelligence analysis meet and review the submissions. Winning entries are then invited to a closed session held in conjunction with the VAST symposium. In this session, the participants work with a professional intelligence analyst to investigate a data set for possible threats. The data set for the interactive session is smaller than the contest data set to allow for sufficient progress to be made in a two hour session.

The metrics for the VAST contest are refined every year. The different categories of metrics for the VAST 2007 contest were [4]

- Accuracy on the answers to who, what, where and when questions
- Quality of the debrief
- Quality of the visualizations
- Appropriateness of the interactions
- Overall utility of the components of the software

Clearly, quality measures are subjective. Each of the above categories had subcategories. Judges were given criteria and asked to assign points from 1-5 to provide a rationale. It is not surprising that judges with different expertise had difficulty in rating different aspects. The analysts were very comfortable rating the debriefs while the visualization and usability professionals were not. And while the visualization and usability professionals were comfortable in rating the visualizations and interactions, the comments of the analysts were extremely helpful to rate these in the context of use.

The overall utility of the different software components was the most difficult aspect to consider. Ideally these components should be examined for their contributions to different aspects of sensemaking but that was difficult to do from the static submissions (process description and screen shots).

The contest committee is interested in refining the metrics to reduce the ambiguity in judging the subjective measures. Observations of the interactive session help in this respect as do our other interactions with analysts in different NVAC activities.

**SENSEMAKING AND METRICS**

Since sensemaking is an integral part of analysis it is logical that we focus on assessing how tools help with the different aspects of sensemaking. However, not all analytic tasks are the same. What aspects of the task do we need to consider in devising appropriate measures for assessing improvement? Let’s consider three different scenarios.

Scenario 1 involves a law enforcement official who gets a tip that an individual might have been involved in a particular crime. The law enforcement agent will go to a number of databases to put together what is known about this individual including such things as outstanding warrants, past arrests, any information about whereabouts. Using this information the law enforcement official will determine whether or not to further investigate this individual.

Scenario 2 involves an intelligence analyst who has been asked to prepare a report on some suspicious activity in a remote corner of the world. There seems to be some interaction with a foreign chemical plant and construction activities in this region. In addition, the small military in this rather volatile country have been actively moving troops to different parts of their country. This analyst will need to look through many reports and possibly images to understand and report on the situation. It is known that this country has a reputation for hiding activities both from the outside but also from citizens within the country.

Scenario 3 also involves a law enforcement official. This analyst is aware of a particular crime that is being investigated and is looking through other unsolved cases to determine if there are similarities with this particular case.

These three scenarios are only a small representation of the variability of analytic tasks. However, even in these three tasks we can see the activities included in foraging, hypothesis generation, and evidence gathering. In scenario 1 the analyst is looking at a number of data bases but focusing on a particular known individual. In scenario 2 the analyst is looking at heterogeneous data, needs to cover a number of possible scenarios, but has to consider misinformation and deception in interpreting the information as well as missing information. In scenario 3, the analyst is looking through many cases to determine if there are patterns that match a case currently under investigation.

**Analytic Methods**

Moreover, there are a number of analytic methods that analysts employ. These include:

- Problem Restatement
• Link analysis  
• Social network analysis  
• Telephone toll analysis  
• Timelines  
• Decision Tree  
• Analysis of Competing Hypotheses  
• Deception detection  
• Red Cell  
• Devil’s advocate  
• Team A/Team B  
• Alternative futures  
• Role Playing

A number of these are best done as collaborative exercises, such as Team A/Team B, Devil’s advocate and Role Playing. Other methods, such as Red cell, involve looking at the situation from another point of view to gain a different perspective [4].

WORKSHOP RESULTS

A workshop on Evaluation for Visual Analysis Environments was held at InfoVis 2007. Participants were asked to submit papers outlining possible metrics. Prior to the workshop the organizers extracted metrics from these position papers plus added measures used in the VAST 2007 contest. The metrics were grouped into a number of different categories. The category labels were NOT discussed and therefore should only be considered as placeholders; however, it is clear that metrics in a number of different categories should be considered.

Discussions during the workshop lead to the following list of prioritized metrics:

Analysis
- Was the analyst able to explore a comprehensive set of hypotheses?
- Was the analyst able to generate a number of hypotheses; eliminate red herring hypotheses; track the hypotheses being followed?

Collaboration
- Was the analyst able to obtain multiple viewpoints?

Visualization - Data types
- What kinds and amounts of data can be considered by the tool?

Visualization - usability
- Effectiveness, efficiency, user satisfaction BUT in context of user’s work/goals

Visualization - utility
- Flexibility/ adaptability of visualizations

Work Practices
- Productivity measure (Document decisions per hour was suggested in one particular domain)

Quality
- Report quality (accuracy, usability, relevance)

Insertion
- Integration of visualizations into current work practices

Usability
- Number of steps needed to accomplish basic tasks

Participants represented several different domains – intelligence analysis; transportation; and law. It was agreed that productivity measures could differ based on the domain and even on the high level task.

The workshop participants agreed that this is clearly only a beginning. A web site is being established and a list serve will be setup. We envision further workshops as more of us attempt evaluations and can report on our results.

NEXT STEPS

A number of suggestions for evaluating visual analytic environments have been proposed. However, these are more general metrics and do not directly assess the impact the software has on sensemaking activities [1, 8, 12]. A number of different activities currently underway will contribute to better means of assessing new tools for analysis. Sensemaking research and task analysis will improve our understanding and identify measures. Current work [3] has been extremely useful for researchers developing tools to improve sensemaking. Working with different types of analytic tasks, different users, and different application domains will help us understand commonalities as well as differences. We also need to look at other research areas to understand the impact of different application domains and tasks. Human–robot interaction (HRI) is also struggling with this problem [11].

Task analysis has been widely used in the HCI field as the basis for understanding user needs and designing software to integrate well into the current work practice. In information analysis, however, we need to consider that current work practices are NOT (in most cases) satisfactory to deal with the increasing amount of data types and data sources. We must design tools to facilitate transitions to or at least accommodate new work practices. The issue of how various aspects of problem solving should be divided between humans and computers will be of great importance here.

We are working on an approach in the VAST contest to reduce the risk in entering and also to increase the reward. The goal is to have participants learn more about their software. We are investigating ways to use the evaluations to increase communication between participants. We are also investigating ways to increase participation by allowing partial entries. This might be done by setting up tracks for different sensemaking activities.
We are working to develop an infrastructure for the VAST contest to make the data sets and metrics easily available for interested parties to use on their own. This will include ways for users to provide feedback on the results of their evaluations. This, in addition to workshops on evaluation, will help us understand which metrics work in which situations.

We also need to incorporate information from other activities. Badalamente and Greitzer [2] reported on a workshop with analysts identifying the top needs for tools. Included in this list were very practical needs:

- Seamless data access and ingest
- Diverse data ingest and fusion
- Imagery data resources
- Intelligence analysis knowledge base
- Templates for analysis strategies

Hypothesis generation and tracking was also high on the list.

We also need to work with different research programs in the analytic community to understand their goals and to develop ways to measure the impact of the research tools they seek to develop. The Novel Intelligence from Massive Data (NIMD) program was focused on user-centered metrics and provided researchers with a much improved understanding of users and user-centered evaluations [9,10]. Looking at software that has been transitioned into the various analytic domains from these research programs is also helpful. Looking at software that was NOT transitioned into actual use is also useful.

It is important to remember that user-centered evaluation is a communication tool. The metrics and evaluation methods are useful to convey to researchers the priorities and problems of the analysts. This is extremely important in domains where we are designing for complex tasks for expert users.

Evaluation is needed to make progress in a field. In the case of visual analysis environments, user-centered evaluation is also a research topic. It must be investigated in conjunction with sensemaking research and research in visual analysis tools and methods.

ACKNOWLEDGMENTS
This work is supported in part by the National Visualization and Analytics Center (NVAC) led by the Pacific Northwest National Laboratory for the U.S. Department of Homeland Security and by National Science Foundation Collaborative Research IIS-0713198.

REFERENCES
7. Redish, J. C. and Scholtz, J., 2007. Evaluating complex information systems for domain experts, paper presented at a symposium on HCI and Information Design to Communicate Complex Information, Memphis, TN: University of Memphis, February. (paper available by request to the first author at ginny@redish.net)