A Framework for User-Centered Evaluation for a Visual Analytics Contest

Sharon Laskowski, Theresa O'Connell, Yee-Yin Choong

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

ABSTRACT

There are many challenges in developing an evaluation methodology for a visual analytics (VA) contest. In this position paper we focus on contest scenarios and how they provide the underpinning for evaluation metrics. Our discussion is based on expanding metrics under iterative development for the Visual Analytics Science and Technology (VAST) contest and metrics used at the National Institute of Standards and Technology (NIST) for evaluating information analysts' interaction with visualization tools supporting analysis. We propose broadening the scope of scenarios to shed more light on the analytic processes used to solve contest "puzzles".

1 INTRODUCTION

The Visual Analytics Science and Technology (VAST) contest was established "to promote the development of benchmark data sets and metrics for visual analytics, and to establish a forum to advance visual analytics evaluation methods [1]." The contests in 2006 and 2007 have provided visual analytics (VA) developers with interesting datasets, and an engaging problem to solve, as stated in a scenario. The datasets contain subplots and some misleading data. Entries were judged for the correctness of their solution to a problem with some metrics that look at accuracy of the solution and the usability and utility of the VA used in the analysis. However, utility has been difficult to characterize and currently is qualitative, using metrics such as interaction quality, insights, and "utility".

It is a goal of the VAST contest to continue to refine the metrics for VA evaluation. Instead of system performance, it stresses correctness and utility with a user-centered-focus. The longer term goal is a library of scenarios, metrics, and evaluation methodologies that can be used to benchmark VA tools and challenge researchers to develop better tools. While there has been work on evaluation methods and benchmarks for information visualizations such as the Infovis Contests [2, 3], VA evaluation needs to go beyond the current metrics. While information visualization tools can be evaluated in terms of how well various low-level exploratory tasks can be executed, they do not address how well the visualizations support analysis. To go to the next step, we propose that these metrics be viewed in the context of a framework for evaluations that is driven by analysis scenarios and associated datasets. The metrics, then, represent the utility of a VA tool to support a particular kind of analysis required by the scenario. The National Institute of Standards and Technology (NIST) regularly performs analyst-centered formative evaluations of VA systems developed for intelligence analysts. We draw on that work to propose metrics and scenarios within the context of a framework for user-centered information visualization evaluations driven by scenarios.

2 SCENARIOS AND EVALUATION IN THE VAST CONTEST

In the first two VAST contests, scenarios fell into the category of "Who Done It" mystery stories. The scenario set out one big question: "What is the situation and what is your assessment of the situation, including possible next steps?" In order to answer the big question, participants had to uncover plots and subplots, naming principal players and identifying the relationships among them. They needed to uncover time lines and to identify places that were important to the plots and subplots. The mystery story can be viewed as a type of "sensemaking".

The current evaluation process coordinates subjective judgments on the quality of the tools and objective measures such as the number of principal players identified. Judges include experts in information visualization and user interface evaluations as well as analysts who either use VA tools in their work or who are the intended users of research VA systems.

For this type of judging, it is important for submitted solutions to include not only major people, places, events, and plots, but also screen shots and a video to illustrate the use of the tool(s).

Judges examine the submissions independently and then meet to discuss their findings. Judges have their own scoring sheets to record their ratings. To support discussion when the judges meet to assess the submissions, judges write short rationales for their ratings. The method is similar to an expert review in which usability experts evaluate the quality of an interface against usability principles. The primary difference is that the VAST contest judging panel includes analysts as well as members versed in usability engineering principles. Another difference is that the judges do not have an opportunity to interact with the VA tools; their findings derive from their subjective assessment of the developers' video and an objective assessment of the answer's correctness.

Given ground truth, judges score the solutions for accuracy, clarity and overall understanding of the situation. The judges base ratings of utility and the visualization itself on what they have seen in the video and in screen shots in the debriefing. The measures require expert judgment, but are quantified. The judges rate the overall utility of each component of the VA. Using a scale of 0 (poor) to 5 (excellent), judges assess overall utility; insights provided by each visualiztion component; and interaction quality over three dimensions: ease of use; consistency of interactions; and accommodation of appropriate interactions by the interface.

Using the 0 to 5 scale, judges rate the visualization itself for overall quality and five dimensions of quality: layout; color use; clarity of symbols; clarity of labels; and saliency of the information. Judges can also add up to five bonus points for scalability; versatility; handling of missing data and uncertainty; collaboration support; creativity/innovative features; learnability; or other usability and utility features not covered elsewhere.

3 THE FRAMEWORK

Frameworks are structures for best practices. Their goal is to empower practitioners to apply repeatable best practices consistently and with rigor. The framework we propose is a structure for the evaluation of the correctness of submissions as well as the usability and utility of VA systems used for

information analyses. The framework attempts to broaden as well as organize the scope of evaluation of contest entries.

3.1 Scenarios and Analysis Metrics

Scenarios and datasets are both artificially created. They are highly interdependent because the scenario is built on top of the dataset. While current scenarios and datasets represent a valid scope for a VA contest, there are other possibilities, derived from commonly used analysis strategies. These include sensemaking, information seeking, prediction, network analysis, and hypothesis generation. We would expect that a particular VA tool or component would address some scenarios more effectively than others. We can measure this using the metrics already developed in the VAST at a low level. For example, The VAST contest dataset has ground truth. With ground truth we can know the extent of correctness of analysts' findings. With a library of different scenarios, we can begin to develop utility metrics tied to the tasks associated with different scenarios. We now describe some initial approaches for utility metric development.

3.1.1 Addressing Analytical Processes

Understanding the processes that analysts use can help us judge a VA tool's ability to empower those processes. Hypothesis formation, consideration, elimination, and validation are important steps in intelligence analysis processes. Instead of simply asking participants to identify players, the scenario can require them to report the hypotheses they investigated and the processes they used to do this. The goal of the *hypothesis scenario* would be to see if the VA systems support processes used by analysts, e.g. analysis of competing hypotheses [4]. Metrics can include the number of hypotheses generated; hypotheses eliminated; red herring hypotheses followed; and red herring hypotheses eliminated.

Analysts also perform predictive analyses, providing projections of future events and their potential impacts. If each event in the data has a date and/or time associated with it, and if events are consequential, such that one event necessarily causes one or more outcomes, we could measure the tool's ability to empower analysts to do predictive analyses. The *predictive scenario* would have to require participants to provide a timeline or some indicator of the sequence of events. The VAST contest already evaluates accuracy in terms of plots, times, people and events. The *predictive scenario* requires more specificity with respect to time. It requires a timeline. It adds the need for a statement of relationships among events.

Sensemaking is an integral step in analysis. It depends upon information finding and organization. After becoming familiar with the interface by seeing the film and screen shots in the debrief, analyst judges can rate the tool's ability to assist an analyst in information finding and information organization to promote sensemaking.

The culmination of the analytical process is often a report. The debrief section of submissions is a surrogate report. The quality of the reports sheds light on the quality of the VA tools. Brei [5] cites four qualities of an intelligence report: accuracy, objectivity, usability, and relevance. Given ground truth, we can rate the debriefs against these qualities.

We propose to do further work on characterizing scenarios and their associated tasks and metrics for future VA contests. In the next section we discuss some relatively simple ideas for improving the next VAST contest that will lead to further development of scenarios and metrics in the long term.

3.2 Utility/Usability Metrics

We propose that for the next contest, the metrics be revisited to include visualization metrics currently used at NIST that were developed specifically for evaluating VA tools. To mitigate the fact that adding dimensions to the judging criteria increases the complexity of the judging process and causes participants to have to supply more information about their systems, the additional dimensions could be graded as bonus questions.

There is an opportunity to extend the range of the usability principles that underlie the utility/usability aspects of the evaluation. The contest does not currently include user effort in its judging criteria. Using the usability principle that an interface should require the minimal number of actions to accomplish a goal, we can count the number of steps taken to accomplish basic tasks during the film.

There is also the opportunity to make the measures more specific. For example, judges rate screen layout and information saliency, but for visualizations, there are more specific aspects of screen layout that are integral to user efficiency and effectiveness. One of these aspects is the absence of occlusions (occasions where data an analyst seeks is hidden by another screen element.)

The purpose of VA systems is to support analysis. There is an opportunity to take advantage of the fact that practicing analysts serve on the judging panel. NIST evaluations typically ask analysts if the VA system is appropriate for insertion into their workplaces. The explanations that accompany their answers provide valuable feedback for developers. Ask analyst judges to rate the degree, from their perspective, that the tool will help them do their jobs more efficiently and effectively.

4 Conclusions and Future Work

The framework we propose takes advantage of the experience of analysts on the judging panel and provides valuable feedback to developers. It takes VAST contest metrics to the next level of specificity.

The inherent challenge is that the framework must not be too complex. We believe that building the framework around scenarios and the capabilities of VA tools to address specific scenarios can help focus the metrics and evaluations.

In this workshop we would like to discuss this framework and begin the effort to develop a taxonomy of scenarios.

References

[1] IEEE VAST Contest, http://www.cs.umd.edu/hcil/VASTcontest07/

[2] Information Visualization Benchmark Repository http://www.cs.umd.edu/hcil/InfovisRepository/index.shtml

[3] Plaisant, C., Fekete, J-D. and Grinstein, G. Promoting insight based evaluation of visualizations: From contest to benchmark repository. http://hcil.cs.umd.edu/trs/2004-30/2004-30.htm

[4] Heuer, R.J. Psychology of intelligence analysis. Center for the Study of Intelligence, Central Intelligence Agency. 1999

[5] Brei W. S., *Getting Intelligence Right: The Power of Logical Procedure*, Occasional Paper Number Two (Washington DC: Joint Military Intelligence College, 1996), 6.