SAND2009-7914C

Proposed Working Memory Measures for Evaluating Visual Analytics

Laura Matzen, Laura McNamara, Kerstan Cole, Alisa Bandlow, Courtney Dornburg, & Travis Bauer
Sandia National Laboratories
PO Box 5800
Albuquerque, NM 87185
505-844-1505

lematze@sandia.gov, lamcnam@sandia.gov, kscole@sandia.gov, abandlo@sandia.gov, ccdornb@sandia.gov, tlbauer@sandia.gov

ABSTRACT

The current information visualization literature highlights design and evaluation processes that are highly variable and situation dependent, which raises at least two broad challenges. First, lack of a standardized evaluation criterion leads to costly re-designs for each task and specific user community. Second, this inadequacy in criterion validation raises significant uncertainty regarding visualization outputs and their related decisions, which may be especially troubling in high consequence environments like those of intelligence analysts. As an attempt to standardize the "apples and oranges" of the extant situation, we propose the creation of standardized evaluation tools using general principles of human cognition. Theoretically, information visualization tools enable the user to see information in a way that should attenuate the user's memory load and increase the user's task-available cognitive resources. By using general cognitive abilities like available working memory resources as our dependent measures, we propose to develop standardized evaluative capabilities that can be generalized across contexts, tasks, and user communities.

Keywords

Information visualization, Usability evaluation, Cognitive load, Working memory.

1. INTRODUCTION

The current information visualization literature highlights design and evaluation processes that are highly variable and situation dependent [1]. One important reason for this is that the field of information visualization is not focused on a particular domain or type of data. Instead, researchers have identified a range of knowledge tasks that lend themselves to visualization technology, including exposing uncertainty, concretizing relationships, formulating cause and effect, determining domain parameters, multivariate explanation, testing hypothesis, answering previously unforeseen questions, looking at data from different perspectives, and discovering patterns [1]. A wide variety of user communities can benefit from visualization technology that supports these knowledge tasks. We will focus on the Intelligence Community (IC) in our discussion because information visualization tools could be extremely beneficial to this community. In addition, intelligence analysis is a high-consequence domain in which understanding the impact of the tools on analyst performance is particularly important. However, the same ideas apply to any user community that deals with large data sets and could benefit from information visualization tools.

The amount of information that is theoretically available to intelligence analysts is a double-edged sword: although analysts are privy to extensive, often proprietary datasets, it can be difficult to retrieve, assess, aggregate, and interpret the massive amounts of data that such databases contain, particularly considering the tight timelines that analysts frequently face. Ideally, information visualization tools should make more information available, more easily and rapidly, than the current suite of search, notation, and storage tools that analysts typically have on their desktops. Perhaps most importantly, these tools must minimize cognitive load, to ensure that analysts' cognitive resources are fully available for making sense of complex information sources.

We are aware of many visual analytics toolsets aimed at the Intelligence Community. However, formal validation of these tools is rare for a range of reasons. Ideally, the formal validation of information visualization software would involve a controlled, comparative experiment to assess the impact of information visualization tools on intelligence products used for decisions in key real-world events, including the completeness, reliability, and accuracy of the assessments derived with and without visual analytics software. Such formal validation studies are difficult for a variety of reasons. For one thing, intelligence analysis is often task and analyst-specific, involving a wide range problems, datasets, domains, and approaches. This makes experimental controls difficult to attain across a population. Secondly, ground truth can be difficult to come by. Collecting historical user effectiveness data requires accessing and understanding complex data sets that are often sensitive, proprietary, even classified. Additionally, some data sets lack ground truth by their very nature. For example, terrorism threat analysis accuracy may be indeterminate if the threat is never realized.

This is not to say that information visualization researchers have not tried to develop evaluative techniques to assess the usability and usefulness of their software packages. However, as discussed by Plaisant et al. (2008), *in-situ* usability studies and controlled experiment methodologies related to information visualization evaluation are "helpful but take significant time and resources,"

[5] and moreover, do not generalize across conditions and contexts, leading to costly re-designs for each project and specific user community.

A key reason for the cost and difficulty of well-controlled usability studies is the lack of standardized metrics that can be used to comparatively evaluate different information visualization tools across a range of knowledge domains. Given the demands that intelligence analysis and other complex analytic tasks place on cognitive resources, we propose using cognitive workload measures as a source of standardized evaluation metrics for information visualization tools. Measures of cognitive workload have been well-characterized by psychology and human factors research. We suggest that evaluation techniques focused on cognitive processing measures could provide metrics that are applicable across tools, tasks, and datasets. Evaluative principles that rely on cognitive processing, and not on domain-specific analysis outcomes, can lead to more cost-effective design principles for all visual analytics software in the IC. We believe that cognitive load evaluation is a cost-effective start to developing performance metrics for visual analytics packages.

2. COGNITIVE LOAD EVALUATION

Our approach begins with the recognition that intelligence analysis involves many challenges that relate to human cognition. Those include manipulation and comparison of information, remembering relevant information, discrimination of threat from non-threat, and avoiding cognitive bias. These are cognitively intense activities that require maximum attention.

We suggest that *effective information visualization tools should minimize the cognitive demands stemming from finding and manipulating raw data*. Instead, effective tools should lower cognitive burden, freeing the analyst's cognitive resources for making sense of the information. In making this suggestion, we agree with Huang et al. (2008), who point out that typical visualization performance measures compare performance differences in response time and accuracy, but fail to capture the amount of mental effort that might be required to compensate for a poor visualization tool. Thus, while performance measures might be equivocal across two visualizations, users may have to expend greater mental effort to compensate for a bad visualization.

Cognitive load is defined as "the amount of cognitive resources needed to perform a given task" [4]. Cognitive load measures have been used to test performance in a variety of domains [2]. For example, in educational and instructional research, measures of cognitive load are increasingly used in conjunction with outcomes measures (such as the quality and quantity of information that a user acquires) to assess loads on cognitive capacity in multimedia learning environments [3-5]. In general, there are four types of cognitive load measures: primary task measures, such as measures of a person's speed and accuracy when completing their main task; secondary task measures, which measure performance on a concurrent task; physiological measures, where physiological markers of stress or effort are recorded; and subjective measures, such as questionnaires. There is some precedent for using measures of cognitive load to assess software designed for the IC; however, most are limited to subjective questionnaires [5, 8, 9].

We are particularly interested in metrics related to *working memory*, which refers to the brain's ability to acquire and maintain small amounts of information under active processing for short periods of time (cf [6]). Working memory is the "theoretical construct that has come to be used in cognitive psychology to refer to the system or mechanism underlying the maintenance of task-relevant information during the performance of a cognitive task" [7]. Working memory measures seem to map well onto what North (2006) has called for in terms of a new evaluation method to measure visualization "insight" [4]. North describes insight as a process that is complex, deep, qualitative, unexpected, and gives relevance to the data by connecting it to existing domain knowledge. If an analyst is devoting more cognitive resources to navigating a difficult interface, he or she will have fewer resources available for analyzing and understanding the data. This could decrease the likelihood of gaining insight into a visualized data set and increase the analyst's chances of making errors or missing important information.

We propose using working memory metrics to evaluate information visualization tools by designing a methodology to assess what proportion of a user's working memory resources are consumed by using the tool itself as opposed to making sense of the data that the tool is designed to illuminate. This idea is related to previous work by Huang et al. [4] who have suggested that visualization effectiveness relates to how well a person can maintain concentration when working on a complicated task. Their focus was on using general-purpose processing resources to assess the distractive nature of a visualization. We share this focus on utilizing general-processing resources, but our focus is on assessing the usefulness of a tool by how well it allows users to devote their working memory resources to thinking about the data as opposed to wrestling with the visualization tool. We suggest that when an information visualization tool is useful and intuitive for its users, their cognitive resources will be free to support the types of high-level cognitive processes that support understanding and gaining insight into the data.

3. PROPOSED METHODOLOGY FOR WORKING MEMORY MEASURES

To assess the feasibility of working memory-oriented metrics for evaluating induced cognitive load, we are engaging in a series of evaluation studies during this calendar year (2010) in which professional analysts will be asked to solve a series of simple information retrieval problems that are similar to the kinds of tasks they complete in their daily work. In a counterbalanced within-subjects test, the analysts will solve one problem using traditional read-and-search methods with raw text data and another problem using a prototype visual analytics tool built on Sandia's Titan visualization framework (see www.vtk.org). In addition to primary task metrics, such as time to resolution, correctness, and completeness, all participants will be asked to perform a secondary working memory task as part of a dual-task experiment.

In a dual-task experimental paradigm, participants are asked to perform two tasks simultaneously. This type of study design is used frequently in psychology and human factors research. The participants can be directed to focus their effort on both tasks equally or to focus on one task (the primary task) at the expense of the other (the secondary task). Participants can perform well on

a secondary task only when they have excess cognitive resources that are not consumed by their primary task. Secondary task measures have been shown to be more sensitive measures of workload than primary tasks alone [13, 14].

In our proposed evaluation methodology, the participants' primary task will involve interacting with software tools as described above. As a secondary task, participants will complete a well-validated memory task called the Sternberg task [15]. This task, which has been shown to be a sensitive secondary task measure in numerous human factors studies [e.g., 15, 16], will allow us to assess participants' spare working memory capacity during concurrent interaction with the visualization tool. The Sternberg task will require participants to remember and detect a set of targets among a stream of distracter items. Specifically, participants will hear target set such as a set of three random letters that they will have to maintain in working memory. After a delay, participants will hear a series of probe letters and they must indicate whether or not each probe is a member of the target memory set. Their accuracy and reaction time will be recorded and compared to their baseline level of performance (their speed and accuracy when they are completing the task alone, without a concurrent primary task). By measuring participants' accuracy and reaction time to target items during this secondary memory task, we can assess the cognitive load that the visualization tool imposes on participants' working memory.

If the software is difficult to use, navigating it will impose a high burden on the participants' cognitive resources and they will have very little working memory capacity left over for the secondary task. That will degrade their reaction times and accuracy for the Sternberg task. In contrast, if the software is easy to use, participants will perform well on the secondary task, with little or no degradation from their baseline levels of performance. Since the secondary task is continuous, we can examine performance over time to determine if some aspects of interacting with the software are more difficult than others or to compare the effectiveness of different types of tools, processes, or visualizations.

Our primary hypothesis is that good performance on the secondary working memory task indicates that the software does not impose a high burden on cognitive resources. In a real-world analysis task, that would mean that the analyst would have more resources available to support high-level cognitive engagement with the data set. In other words, performance on the secondary working memory task should be a good indicator of the effectiveness of the visualization tool. In the course of our experiment, we will do several comparisons to test this hypothesis. First, as mentioned above, we will have analysts use traditional read-and-search techniques or use a visualization tool to perform the same search and retrieval tasks. We will ask participants to complete the NASA TLX questionnaire to assess subjective workload experience for each method, as has been done in previous evaluations of software for the IC [10]. We will evaluate the results of the secondary working memory task with respect to the analysts' performance on the primary task metrics (time to resolution, correctness, and completeness) and subjective workload measures for each tool.

In additional experiments, we will conduct tests in which groups of analysts interact with visualization tools that have different types of visual representations that are judged subjectively to be easier or more difficult to interpret. We will also compare different versions of the same tool in which a key feature is present or absent, making the software easier or more difficult to use. Once again, we will compare the analysts' performance on the secondary working memory task to their performance on the primary task and their subjective assessments of the workload imposed by interacting with each variant of the software. We expect to report the results of these experiments in late 2010; in the meantime, the authors are happy to provide information on our study design and our progress.

If we are successful in validating the secondary working memory task as a tool for evaluating information visualization software, the same basic techniques could be applied to evaluating software for any domain. Since this method is based on general principles of human cognition, it can be used as a standardized comparison across different user groups, analysis tasks, and data sets. New tools or variants of tools could be tested quickly and easily to determine their effectiveness. In addition to comparing complete software tools, the continuous nature of the working memory load assessment could also be used to identify problematic points within a tool where a user might get bogged down or confused. The basic study design could be applied to any evaluation: a dual-task paradigm in which users interact with the software as a primary task and perform a Sternberg working memory task as their secondary task. The users would not need to be domain experts; they would simply need to walk through the mechanics of working with the visualization tool. This method could provide a fast, cost-effective, and standardized way for assessing the effectiveness of new tools or

4. CONCLUSION

We believe that evaluation approaches that incorporate working memory-derived measures will help researchers assess whether their tools actually enhance users' cognitive processing. Such evaluation approaches will also have reliability and validity across temporal, contextual, and user-related variances. We also believe such measures can be used to evaluate any type of visual analytics software without requiring the costly and time-consuming development of application-specific evaluation metrics. An effective tool should demand few of the user's cognitive resources, enabling the user to perform high-level analysis and sensemaking tasks more effectively. We think this novel approach to evaluating visualization software will allow for: 1) effective comparisons across different users, data sets, and analysis tasks and 2) informative evaluations of visualization tools without the need to develop costly tool- and application-specific evaluation metrics and design.

5. REFERENCES

[1]. Brunken, R., Plass, J.L., and Leutner, D. Assessment of Cognitive Load in Multimedia Learning with Dual-Task Methodology: Auditory Load and Modality Effects, *Instructional Science,* vol. 32, pp. 115-132, 2004.

[2]. Brunken, R., Steinbacher, S., Plass, J.L., and Leutner, D. Assessment of Cognitive Load in Multimedial Learning Using Dual-Task Methodology, *Experimental Psychology,* vol. 49, pp. 109-119, 2002.

[3]. Meneghetti, C., Gyselink, V., Pazzaglia, F., and De Beni, R. Individual Differences in Spatial Text Processing: High Spatial Ability and Compensate for Spatial Working Memory Interference, *Learning and Individual Differences,* vol. 19, pp. 577-589, 2009.
[4]. Heuer, R. *The Psychology of Intelligence Analysis*. Washington, DC: Central Intelligence Agency, 1999.
[5]. Huang, W. Eades, P. and Hong, S. Beyond time and error: A cognitive approach to the evaluation of graph drawings, *Proc. 2008 Conference on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2008.
[6]. Morse, E., Steves, M.P., and Scholtz, J. Metrics and methodologies for evaluating technologies for intelligence analysts. *In Proc. Conference on Intelligence Analysis*, 2005.
[7]. Munzer, T. A Nested Model for Visualization Design and Evaluation, *IEEE Transactions on Visualization and Computer Graphics,* vol. 15, pp. 921-928, 2009.
[8]. North, C. Toward measuring visualization insight. *IEEE Computer Graphics and Applications 26* (3), pp 6-9, 2006.
[9]. Plaisant, C., Fekete, J., and Grinstein, G. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE Transactions on Visualization and Computer Graphics* 14 (1), pp. 120-134, 2008.
[10]. J Scholtz, E. Morse, M.P. Steves. Evaluation metrics and methodologies for user-centered evaluation of intelligent systems. *Interacting with Computers* 18, pp 1186-1214, 2006.
[11]. Shah, P. and Miyake, A. Models of working memory: An introduction. In A. Miyake and P. Shah (eds.). *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, Cambridge University Press, Cambridge, UK, pp 1-27,1999.
[12]. W. B. Verwey and H. A. Veltman, "Detecting Short Periods of Elevated Workload: A Comparison of Nine Workload Assessment Techniques," Journal of Experimental Psychology, vol. 2, pp. 270-285, 1996
[13]. Gawron, V. (2008). Human performance, workload, and situational awareness handbook. Boca Raton, FL: Taylor & Francis.
[14]. Meshkati, N., Hancock, P.A., & Rahimi, M. (1989). Techniques of mental workload assessment. In: J. Wilson (Ed.).Evaluation of Human Work: Practical Ergonomics Methodology. London: Taylor and Francis.
[15]. Wierwille, W.W., and Eggemeier, F.T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35, 263-282.
[16]. Wickens, C.D., Hyman, F., Dellinger, J., Taylor, H., & Meador, M. (1986). The Sternberg memory search task as an index of pilot workload. *Ergonomics, 29,* 1371-1383.