Developing Qualitative Metrics for Visual Analytic Environments

Jean Scholtz
Pacific Northwest National Laboratory
Richland, WA
+1 (509) 375-2121
Jean.scholtz@pnl.gov

ABSTRACT

In this paper, we examine reviews for the entries to the 2009 Visual Analytics Science and Technology (VAST) Symposium Challenge. By analyzing these reviews we gained a better understanding of what is important to our reviewers, both visualization researchers and professional analysts. This is a bottom-up approach to the development of heuristics to use in the evaluation of visual analytic environments. The meta-analysis and the results are presented in this paper.

Categories and Subject Descriptors

H 5.2 [**Information Interfaces and Presentation**]: User Interfaces – *evaluation / methodology*

General Terms

Design, Experimentation, Human Factors.

Keywords

VAST Challenge; heuristic evaluation; qualitative metrics; visual analytic environment.

1. INTRODUCTION

Quantitative measures have been used in computer science for evaluating software for many years. Benchmark datasets have been used to produce performance measures such as processing time and memory use. As more graphical user interfaces were developed, usability metrics of effectiveness (percentage of tasks completed), efficiency (time to complete tasks), and user satisfaction became important. User-centered software engineering [6] emphasized integrating users in the software development process and collecting qualitative feedback during prototyping and implementation. Human-computer interaction professionals have developed a number of methods to facilitate the design of usable and useful software tools. In particular, task analysis [4] has been used for documenting user tasks and processes. User testing [3] and heuristic reviews [7] are common methods for evaluating the usability of user interfaces.

Researchers in the visual analytics community realize the value of qualitative metrics and have proposed various solutions in recent years. Bertini and Santucci [2] proposed quality metrics for effectiveness and feature preservation of visualization. In this work, they note a number of practical issues that quality measures for visualization need to address. In particular, metrics should be generalized to different types of visualizations and baselines need to be established to allow comparisons.

A number of visualization researchers have also looked at heuristic evaluations for visualizations [1,8,9,11].While the consensus is that these are useful, there is considerable work that must be done before an agreed upon set of heuristics exists. There are many levels of heuristics that can be applied ranging from low level perceptual aspects of visualizations to higher level interactions with visualizations to perform tasks. In applying heuristics to visual analytic systems, we also need to evaluate the analytic process supported by the system.

In this work our task is to develop metrics for evaluating the utility of visual analytic environments for professional analysts. As such the evaluation must cover a number of issues including the visualizations, how the visualizations facilitate analysis, user interactions with the visualizations, and the support the environment provides for the analytic process, including the results of the analysis. A constraint is that these metrics will be applied in a contest setting in which the evaluators will not have direct access to the visual analytic systems. In this paper we address several questions.

1. What materials should be provided to evaluators to ensure that they can adequately assess the visual analytic systems?
2. What aspects of visual analytic systems are important to reviewers?
3. Is there an advantage to selecting evaluators from different domains of expertise (visualization researchers and professional analysts)?

BELIV'10, April 10-11, *2010*,Atlanta, GA, USA.

2. BACKGROUND

For the past four years the Visual Analytics Science and Technology Symposium (VAST) has included a contest or challenge. The objective of these VAST competitions is to provide researchers in visual analytics with tasks and data representative of real world problems and to provide researchers with feedback on the utility of their software tools. We have always provided datasets with embedded ground truth. This has been very popular as our participants know when they have correctly assessed the situation. While we realize that this is not realistic in actual analysis, providing fictional scenarios with known answers is used for training intelligence analysts.

The competition has increased in popularity from entries numbering in the single digits in 2006 (6 entries) and 2007 (7 entries) to a double digit number of entries in 2008 (73) and 2009 (49). This increase is attributable to a restructuring of the tasks and data using the concept of mini challenges. In the earlier competitions we provided heterogeneous data and it was necessary for teams entering the competition to analyze all the data. A mini challenge uses homogeneous data and has specific questions that teams need to answer. To enter the grand challenge the teams need to combine their analysis of all the mini challenge data.

Another change involved moving from designating "winners" to giving awards. This allows us to recognize teams that demonstrate particularly interesting or innovative work in the visualizations, the analysis process, the analytic product or some aspect of the tool design. This provides incentives for student teams and commercial teams and removes the stigma of "not being a winner."

In order to make awards, we need some basis for assessing the visual analytic tools. Currently we use both quantitative and qualitative measures. The data for the competitions is provided by the Threat Stream Generator Project at the Pacific Northwest National Laboratory [10]. This team provides tasks and synthetic data in which ground truth is embedded. This means that we can assess whether or not teams find the ground truth. We recognize that real analysts rarely know that they have "found" the correct answer. However, real analysts are given case studies with known answers as a training tool [5].Therefore, these competitions can be viewed as training for the researchers in analysis. This then leads to relatively easy quantitative metrics. We can look at the "answers" which could be a list of people, incidents, transactions, social networks, etc. The concept of having to find a known item or items in the data makes the competitions very engaging and provides a way to determine how effective the visual analytic tools are in the analytic process. This also gives the evaluators a basis of comparison – both for the quantitative results and for comparing the analytic process and the usefulness of the visualizations in that process.

This brings us to the qualitative measures. As we are working on evaluation within the domain of visual analytics tools, the key is that analysts will use the visualizations, explore them and interpret them to arrive at some conclusions. The analyst is an integral part of the process and therefore, needs to be an integral part of the evaluation.

There are many different types of analysts, depending primarily on the material they are experienced in analyzing. For example, financial analysts, signals analysts, image analysts and open source analysts deal with different types of input materials. The materials we have used at this point in our datasets are most similar to those used by open source analysts and do not require expertise in a specific area such as financial transactions or genetics. We recruit analysts in our review process who are representative of open source analysts and either are currently employed as an analyst or have been employed as an analyst. We recruit the analysts we work with either through recommendations from others or through professional organizations. The analysts who participate differ each year, based on their work schedules and their interest in the datasets and challenge questions for that year.

In the following sections, we discuss the methodology of the meta- analysis and how the results apply to the questions posed in the introduction.

3. A META- ANALYSIS OF THE 2009 VAST CHALLENGE REVIEWS

3.1 The Review Process

For the first two competitions (2006 and 2007) we had two evaluation meetings for each contest. The meetings occurred on the same day but one was held on the west coast and another on the east coast. Results from the two meetings were consolidated in conference calls. Attendees at the meetings were the VAST contest committee members who are primarily visualization experts. A number of analysts joined each meeting to provide assessments as well. Although we gave the analysts access to the submissions approximately a week ahead of the meetings and provided them with comment forms, not all analysts had the time to fill out these sheets prior to the meeting. However, the discussions during the meetings provided much information about the analysts' views of the different software tools. We took notes during the meetings and then combined these notes into feedback that we returned to the teams. This was very labor intensive and all teams did not receive feedback on all aspects of their submissions due to limited discussion time.

We were not expecting the huge number of submissions for the VAST 2008 Challenge that we received after re-structuring the tasks and data following the 2007 competition. This forced us to quickly devise a way to make the review process more efficient while still providing reasonable feedback to the teams. For 2008, each committee member took charge of one mini challenge and went through those entries to determine which entries deserved a more in-depth review. We held only one evaluation meeting attended by committee members and analysts. The committee members gave a presentation on the mini challenge they had reviewed, presenting the entries that they felt were deserving of more discussion. Each presentation was followed by discussion from the committee. Comments from this discussion were used to augment feedback. We then selected entries that the committee felt should be given awards and the type of award was determined.

The problem with this method was that only one committee member looked at each mini challenge entry and as some mini challenges had a large number of entries, the committee members needed to devote much effort to their reviews in a relatively short amount of calendar time. While the committee could weigh in during the discussions in the meeting, they were likely seeing the entry for the first time. We felt that this process could result in overlooking some deserving entries.

We agreed that a different process would be needed for 2009 based on the assumption that a large number of entries would be submitted. We decided to use a peer review system similar to paper reviews. We requested reviewers from visualization researchers (not on the committee) and analysts. We sent out e-mails to academic and industry visualization researchers. Both faculty and doctoral students were recruited. A submission system was developed along with an accompanying review web site by students at U. of Massachusetts Lowell under the direction of Dr. Georges Grinstein. We assigned three reviewers to each entry; two visualization specialists and one analyst. One committee member was then assigned to be a meta-reviewer for the various mini challenges. As in 2008, the committee member who was the meta- reviewer for each mini challenge gave a presentation at our two-day evaluation meeting followed by discussion. However, this time we had two or three reviews for each entry that were used as a basis along with the meeting discussions for deciding upon awards.

The grand challenges were not assigned for peer reviews but were kept for discussion in the evaluation meeting. Although we had a large number of entries overall, we still only a single digit number of grand challenges. As these entries required more work and had a lengthy intelligence assessment that required evaluation, the committee members agreed to review these entries rather than solicit outside reviewers. We felt that it was fairer to review all of these and compare and contrast them. The second part of the meeting focused on discussions about the grand challenge entries.

We designed review forms for the peer review system based on both review sheets we had distributed in 2006 and 2007 and based on comments we had heard in all the previous evaluation meetings. For the meta-analysis, we analyzed the reviews from the 43 mini challenge entries submitted. One entry was omitted as this entry was more of a meta-analysis of the Challenge than an actual entry. Each entry was originally assigned to three evaluators (reviewers); one was an analyst and two were visualization researchers. We received a full set of reviews (three) for 32 of the 42 entries. Unfortunately, some of the reviewers who had agreed to review for us did not complete their reviews and our schedule as such that we did not have time to recruit additional reviewers. The following sections address the three questions posed in the introduction.

3.2 The Materials Used for Review

What materials should be provided to evaluators to ensure that they can adequately assess the visual analytic systems? Currently the evaluations are done based on materials submitted by the teams; we do not have access to the systems. The materials requested for the competitions include:

- files of the answers in a specific format to allow automatic checking
- a description of the process used to arrive at a given answer (specified in each mini challenge) along with screen shots
- a video showing this process (so the interactions with the visualizations can be demonstrated)
- a description of the results found

In addition, we provided reviewers with the ground truth solution and the entry's accuracy scores.

Starting in 2007, competition participants were provided examples from the previous years to view. We have seen better quality explanations of the process and better analytic products since then. We emphasize to participants that good clarity in their descriptions and videos is essential. If the evaluators are not able to understand what the teams have done, the resulting assessments may not be favorable.

In addition to asking our evaluators questions about the content of the material, we also asked them to rate the clarity of the material. We wanted to convey to the Challenge participants that comments from the evaluators should be interpreted depending on how well evaluators understood the submissions. Therefore, we first looked at the clarity scores to determine where significant differences in perceived quality existed.

Of the 42 entries, 22 of them received a clarity rating of 4 or above (1 was low and 7 was high) from all reviewers. Results included:

- There were 17 entries for the Flitter Mini Challenge (a social network challenge) and 12 of these entries had clarity ratings of 4 or above. Of these 17 entries, only two entries had clarity ratings that differed by more than 2 points.
- There were 22 Traffic Mini Challenge entries (analyzing network traffic and logs of employees entering areas of the building). Of these 22 entries, 10 received clarity ratings of 4 or above from all reviewers and nine entries had disagreements about the clarity of the material.
- Of four video entries, four received clarity ratings of 4 or above from all reviewers but there were differences in agreement about clarity of the materials on two entries.

Interestingly, there was no overlap between entries whose clarity ratings were 4 or above and those where there was a disagreement in clarity. Clearly clarity of the explanation affected the scores of the evaluators. This is not a surprising result but a point that needs to be emphasized to submitters.

We asked teams to submit videos so that reviewers could get a sense of the analytic process used and see the interactions. There were several problems with the videos. First of all, not all teams submitted videos – some teams substituted slides with audio. In these submissions it was difficult to understand the interactions that were available to use in exploring the visualizations. Also, we did not dictate a particular format for the video. Therefore, some reviewers had problems getting the videos to play or getting the audio portion to play. In some cases, the audio was almost impossible to hear. It would be helpful for teams to send their videos to others before submitting it to ensure that they can play and understand the video.

Another issue was the process description. In some cases reviewers commented that much of the video was devoted to describing all aspects of the visual analytic tool, but only a small portion of the tool was actually used in solving the problem. As we accepted a video no longer than four minutes, any time taken to describe features not used in the solution takes away from time that could better be spent helping reviewers understand the step-by-step process. This should include describing the interactions used with the visualizations, noting what was automated and what was manually accomplished. The process should focus on the process from the analyst's point of view, not the technical description.

We also provided reviewers with access to the solutions and the accuracy scores of the entries. It is clear that the accuracy scores affected the reviews. Comments included both positive and negative impacts. Reviewers commented that the entry did not appear efficient but that that correct answer was obtained and consequently rated the entry somewhat higher than their comments justified. The opposite was also true. Reviewers liked an entry but commented that it was not ready for prime time as the team did not obtain the right answer. Finding the right answer was not always a problem of the software but was more frequently an incorrect assumption on the part of the team doing the analysis.

Several reviewers commented that it would be very helpful for them to understand how much work was needed to ingest the data into the system. While the process of analyzing the data might be efficient, it might be offset by the work needed to be done to enter the data originally.

Entries included visual analytic systems and systems customized for these challenges using toolkits. As we asked reviewers to evaluate the efficiency of the analytic process, those reviewing entries that made use of toolkits were confused. They wondered if the time needed to customize the toolkit should be considered as part of the efficiency. While it was relatively easy to judge the efficiency of the process of the analysis, we did not ask for toolkit developers to provide us with the time needed to customize the toolkit.

In conclusion, if done clearly the materials are mostly sufficient for reviewers to do an adequate job of understanding the software environment, with the exception of providing an understanding of how much effort is needed to get the data into the system initially. While we pointed out problems with the submissions above we should also say that the quality of the submitted materials has improved significantly since we started the VAST Challenge in 2006. As we post the submissions on an archival website participants in the next challenges are able to see past entries and note if those received awards. One issue we need to pursue is whether to provide access to the quantitative metrics for team submissions, as this does impact reviewers' perceptions. A second issue is whether entries based on custom-built applications from toolkits include the development effort required to customize the toolkit as part of the analytic process efficiency.

3.3 An Analysis of the Reviews

What aspects of visual analytic systems are important to reviewers? We did not provide heuristics for the reviewers, as an agreed upon set of heuristics for visual analytic environments does not currently exist. Instead we provided the following question categories for ratings and comments.

Reviewers were asked to view the materials submitted by the teams (text descriptions and video) and they were given the accuracy scores for the mini challenge questions. The reviewers were then asked to answer the following questions using a 1-7 scale where 1 is low and 7 is high and to provide comments:

1. Rate your overall satisfaction with the submission.
2. How clearly does the submission explain how visual analytic tools were used to analyze the data and scenario?
3. Rate the usefulness, efficiency and intuitiveness of the analytical process.
4. Rate the usefulness, efficiency and intuitiveness of the visualizations.
5. Rate the usefulness, efficiency and intuitiveness of the interactions.
6. Rate the novelty in this submission (data processing, visualization, interaction, hypotheses generation or evaluation, overall process, etc.).

Reviewers (as well as participants) were given the following brief explanations of what constituted usefulness, efficiency and intuitiveness on the Challenge web site (http://hcil.cs.umd.edu/localphp/hcil/vast/images/uploads/Instructions%20for%20Reviewers.doc):

Usefulness:

- Was the process useful in arriving at an answer or understanding of the situation?
- Were the visualizations useful in understanding the situation?
- Were the interactions useful in providing different views for different purposes?

Efficiency:

- Was the process used efficient or were there many repetitive steps?
- Was it easy to obtain the necessary visualizations or did the user have to go through multiple steps?
- Were there multiple interactions that were used to obtain different views?

Intuitiveness:

- Was it easy to understand the next step to take in the process?
- Was it easy to understand the visualization?
- Was it easy to understand what an interaction did? Was it easy to understand which interaction technique to apply to produce a desired result?

We analyzed discrepancies in the ratings to identify any issues with clarity of the questions. We analyzed the reviews for comments to determine what aspects of the visual analytic environments our reviewers felt were important. The comments, classified under the appropriate category, are listed in Table 1.

These comments can certainly be revised as heuristics that researchers and subsequent challenge reviewers can use as a guide to evaluating visual analytic tools. While none of these are surprising, these comments do help to show what aspects of each category are most important to researchers and analysts.

Analysts were very concerned with the complexity of visualizations. In fact, in one submission an analyst pointed out that the correct suspect was identified in several visualizations but the team actually missed it because of the sheer density of the display.

Reviewers assigned rating scores to the usefulness, efficiency and intuitiveness of each category (process, visualizations and interactions). However, they were only asked to comment about the whole category. While we did not analyze the individual scores given for the three different aspects of the categories, these scores did not seem to vary much for individual reviewers. Feedback from the Challenge participants was that the comments were much more useful to them than the scores. However, asking the reviewers to think about usefulness, efficiency and intuitiveness was indeed helpful judging from the times they mentioned in their comments something affecting these aspects either positively or negatively.

As explained earlier, the review sheet asked for ratings and comments in various categories. While table 1 shows the comments in the correct category, the reviewers did not always place these comments in the correct category. In particular, comments relating to visualizations and interactions were often placed in the analytic process category.

The comments in table 1 were taken directly from the reviews and are negative. This does not imply that the positive versions of these were not commented on in the actual reviews. It is difficult to provide the positive comments in this paper without describing details of the software environments. We intend to rewrite the comments shown in table 1 as heuristics in the near future and we will supplement them with positive examples.

3.4 Analysis of Reviews based on Reviewer Expertise

Is there an advantage to selecting evaluators from different domains of expertise (visualization researchers and professional analysts)?

One approach to answering this question is to look at entries where the average scores given by reviewers differed substantially (greater than two points). Table 2 shows nine entries for which this is true. In the labels, T refers to an entry to the traffic mini challenge, F to the flitter mini challenge and V to the video mini challenge. Please note that although the columns are labeled Analyst, Vis #1 and Vis #2, they do not imply that the same people reviewed the nine entries.

Table 1. Comments from the reviews

Category	**Important aspect**
Analytic process	Highly manual processes
	Repetitive steps in process
	Large amount of data that analysts have to visually inspect
	Automation that might cause an analyst not to see an important piece of data
	Need for analyst to remember previously seen information
	Too many steps
	For automatic identification of patterns and/or behaviors, users need an explanation of what the software is programmed to identify
	Analysts need to document their rationale for assumptions in their reports
	Document the selection of a particular visualization if several are available
	Show filters and transformations applied
	Participants need to explain how the visualizations helped the analysis process
Visualizations	Complexity of visualizations
	Misleading color coding, inconsistent use of color
	Lack of labels; non intuitive labels
	Non intuitive symbols
	Using tooltips instead of labels causes analysts to mouse over too many items
	No coordination or linking between visualizations
	The use of thickness of lines to represent strength of associations is difficult to differentiate
	Difficult to compare visualizations if can only view them serially
	Is the visualization useful for analysis or is it a reporting tool?
	Use of different scales in visualizations is confusing
	Need to relate anomalies seen in visualizations to analysis
Interactions	Too much scrolling
	Interactions embedded in menus
	Too many levels of menus and options to check
	Need to be able to filter complex visualizations
	Need to be able to drill down to actual data

For the eight entries where we have a full set of reviews, the ratings are lower for the analyst in two instances and for one of the visualization researchers in five entries. In the eighth entry the analyst and one visualization researcher are lower than the other visualization reviewer.

We looked at comments where the analyst's rating was lower than those of the visualization researchers. In the Traffic 1 entry the analyst commented that the visualizations were not tied together. While the clarity rating was somewhat low on this entry, one visualization researcher commented that he had heard about this tool at a conference. This may have helped improve his understanding of the tool. In the other case where the analyst's rating was lower, the analyst felt that the tool was more useful for developers than for analysts and there was no explanation of how the team arrived at the answers.

We analyzed the comments where a visualization researcher had lower ratings than the other visualization researcher and the analyst. We were particularly interested to see if the comments that were negative were of a more technical nature about the visualizations. In fact, we found the opposite. The following comments were given in entries where the visualization researcher provided lower ratings:

- The tool does not seem flexible enough to investigate other scenarios
- Should a way to quickly browse through video snippets be considered a visualization?
- One visualization is provided. More are needed to look at other data.
- The analytic process was not well described
- Suspicious events were highlighted in the visualization but it was unclear how they were found
- The visualization is distracting and not useful for analysis
- The analytic process is described in terms of using this tool to detect certain kinds of events without justifying if this type of event is related to the mini challenge question
- The visualization was too compressed. It was difficult to see groupings.
- Tool required an iterative process and it was difficult to remember what had been done.

It was interesting that the visualization researchers commented on a number of analytic issues as well as visualization issues. In reviewing the comments from all reviews we found that analysts tended to comment more from an analyst perspective but visualization researchers were also thinking more about the analytic process and how the tools supported it.

We asked the question about having both perspectives because it is difficult to get analysts as reviewers. The visualization reviewers tend to be academics and reviewing is part of the job. This is not true for analysts, so reviewing for them is definitely an outside activity. While we value their participation and will continue to recruit analysts as reviewers, we are pleased that the visual analytics community is now able to incorporate the user perspective in their comments.

Table 2. Entries where reviewers differed by more than 2 points on ratings

Label	Analy. clarity	Analy. ave.	Vis #1 clarity	Vis #1 ave	Vis #2 clarity	Vis #2 ave.
T1	3	2	5	4	4	4
T2			6	6	6	3
T3	5	4	3	2	6	5
T4	5	5	2	5	1	2
T5	6	6	6	2	7	7
F1	7	7	3	3	6	5
V1	5	3	4	3	4	6
V2	2	3	5	5	7	6
V3	7	7	5	3	6	6

4. CONCLUSIONS

Andrews [1] noted that there are several weaknesses with user testing of visual analytic environments:

- buggy implementation of new interfaces
- testing the wrong users
- familiarity of test users with traditional interfaces
- testing the wrong tasks

Using the VAST Challenge as a way to provide feedback on visual analytic software circumvents a number of these problems. First of all, the tasks and datasets are generated by a team that consists of analysts and visualization experts. The "users" in the case of the VAST Challenge are the developers themselves. No analysts are subjected to buggy software. While one might argue that the wrong users are still being tested, we argue that it is extremely beneficial for developers to have to use their software to understand whether or not it is useful in carrying out the task. Furthermore, having to document their process should help developers reflect on the utility and efficiency. The users, in this case the developers, are eager to use a nontraditional interface, and the reviewers are in fact asked to comment on the novelty of the submission.

Zuk et al. [11] noted that agreed upon heuristics for evaluating visual analytic software do not yet exist. Zuk et al. [11] also noted that higher level heuristics require a holistic evaluation of entire systems. While we do not claim that the tasks used in the VAST Challenges exercise all of most systems, we do note that they definitely require a reasonable analytic process. Asking our reviewers to evaluate the part of the system used in that process is certainly an advantage over just a usability test. By asking the reviewers to look at the different categories, processes, visualizations and interactions, we are able to find aspects under each category that are important. This provides a good starting place for heuristics that can later be validated in the laboratory or in field studies. These heuristics are also a good starting place for researchers to use in assessing their own work. We intend to provide heuristics based on the comments in table 1 to researchers and reviewers in the near future. Analysis of future reviews will be conducted to contribute to this knowledge.

In addressing our questions of whether the materials requested were sufficient for review, what aspects of the visual analytic environments were important to our reviewers and whether it is necessary to ensure that we have reviewers knowledgeable in visualization and analysis, the following summarizes our findings:

- Assuming the materials are expressed clearly they are sufficient for the reviewers to evaluate the analytic process and interactions used in analyzing the given data and situation. However, we will reconsider providing the quantitative evaluation scores as that seems to bias reviewers. We also need to decide if an "efficiency" measure for the analytic process should include the amount of time to ingest data or the amount of time to customize a toolkit.

- An analysis of the reviews has provided us with a starting point of issues that are important to our reviewers. These points will be turned into heuristics for developers for visual analytic software.

- While we are most appreciative of the analysts who help with our review process and we will continue to ask for their help, we are pleased that many in the visualization community are now more aware of the role of the user and can comment on aspects of analytic process and user interaction.

As we are developing qualitative metrics in a bottom-up fashion (that is, based on the reviews we receive for various challenges), it is necessary to ensure that we have provided reviewers with the necessary materials for their review and that we have solicited reviews that address both the technical quality of the visualizations and the utility that the visualizations provide the analysts in their work. It is essential that we continue to do a meta-analysis of the reviews after each VAST Challenge to ensure that this is the case. We can then continue to augment our qualitative metrics based on reviewer input.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Visualization and Analytics Center™ (NVAC™) located at the Pacific Northwest National Laboratory in Richland, WA. The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RL01830. This work was also funded by the National Science Foundation (IIS-0712770).

6. REFERENCES

[1] Andrews, K. 2006. Evaluating Information Visualisation. In the Proceedings of BELIV '06 (Venice, Italy), ACM. New York, NY.

[2] Bertini, E. and Santucci, G. 2006. Visual Quality Metrics, In the Proceedings of BELIV '06 (Venice, Italy), ACM. New York, NY.

[3] Dumas, J. and Redish, J. 1999/1993.A Practical Guide to Usability Testing. Revised Edition. Intellect. Bristol, UK.

[4] Hackos, J. and Redish, J. 1998. User and Task Analysis for Interface Design. John Wiley and Sons, New York, NY.

[5] Hughes, F. J. and Schum, D. Evidence Marshalling and Argument Construction. Washington, DC: Joint Military Intelligence College, Student Course Notes.

[6] Mayhew, D. 1999. The Usability engineering Lifecycle. Morgan Kaufmann Publishers.

[7] J Nielsen, R Mack. 1994. Usability Inspection Methods. John Wiley and Sons, New York, NY.

[8] Tory, M. and Möller, T. "Evaluating Visualizations: Do Expert Reviews Work?" IEEE Computer Graphics and Applications. September/October 2005. pp. 9 – 11.

[9] Trainer, E., Quirk, S., de Souza, C.R.B. and Redmiles, D.F., 2008. Analyzing a Socio-Technical Visualization Tool Using Usability Inspection Methods. In the Proceedings of the IEEE Symposium on Visual Languages and Human Centric Computing, IEEE Computer Society, Washington, DC.

[10] M. Whiting, M., Haack, J., and Varley, C. 2008. Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. In the Proceedings of BELIV '08. ACM, New York, NY, 1-9.

[11] Zuk, T., Schlesier, L., Neumann, P., Hancock, M., Carpendale, S. Heuristics for Information Visualization Evaluation, In the Proceedings of BELIV '06 (Venice, Italy), ACM. New York, NY.