Issues and Methodologies for Evaluating the Jigsaw Visual Analytic System

Carsten Görg* Sarah Williams† John Stasko‡

School of Interactive Computing & GVU Center
Georgia Institute of Technology

ABSTRACT

This article presents an evaluation plan for Jigsaw– an interactive visualization system to support investigative analysis. Jigsaw provides multiple views of a document collection and the individual entities within those documents, with a particular focus on exposing connections between entities. We describe how we plan to evaluate Jigsaw on different levels: general usability first, then focusing on the analytical process, and finally conducting comparative studies.

Keywords: Visual analytics, investigative analysis, intelligence analysis, information visualization, evaluation, metrics

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 INTRODUCTION

The research area of visual analytics recently became more and more established and a number of visual analytic systems have been developed over the last years. Now, the time is ripe to start evaluating and comparing these system to gain insights for improvement and further development.

Such evaluation turns out to be a challenging problem. Card and Pirolli [1] present the sensemaking process that is build by interweaving the foraging loop and the sensemaking loop. Different visual analytic systems cover different parts of those loops and provide support that is more or less automated. This makes it hard to differentiate insight gained using a system from insight gained from human intelligence. Systems that cover different phases of the sense making process are difficult to compare to each other. Other factors make the evaluation process even harder: the systems themselves are big and complex, the input dataset can be huge, analysts use different approaches to solve a problem, and the analytical task can vary from a specific question to an open questions.

Regardless of all those challenges, it is very important to start working on a base foundation for evaluating visual analytic systems. Scholtz [5] introduced hypotheses and measures for evaluation aspects and emphasizes that evaluating usability is an important first step but by far not sufficient to evaluate visual analytic environments. Areas like collaboration, interaction, creativity, and utility also need to be addressed.

In this paper we present a three-fold plan to evaluate our Jigsaw system – our plan covers some of the areas mentioned above. The next section briefly introduces the system. In Section 3, we present our evaluation process along with study logistics, metrics, and sample questions. Finally, we conclude the paper by reflecting on open questions of our plan.

*e-mail: goerg@cc.gatech.edu
†e-mail: sarahw@cc.gatech.edu
‡e-mail: stasko@cc.gatech.edu

2 JIGSAW SYSTEM

Jigsaw provides multiple views that show connections between entities (like people, places, organizations, dates, *etc.*) across documents. A connection between two entities is defined as a co-occurrence in at least one document.

The Text View allows analysts to validate connections, provides their context, and gives access to information that is not extracted as an entity. The List and Graph Views display connections between entities and allow analysts to explore the connection network. The Scatter Plot View highlights pairwise relationships between any two entity types. The Time Line and Calender Views organize entities and reports by date to ease the search for time patterns.

To allow Jigsaw to handle large datasets, the views do not show the entire dataset at once but use an incremental query-based approach to show a subset of the dataset. The query system allows analysts to search for entities and also provides a text search within the documents.

For a detailed description of the system, we refer the reader to an article [7] about Jigsaw in the VAST '07 proceedings and to a video on the project website [3] that shows interaction with the system.

We used Jigsaw to work on the VAST '07 Contest data set. This was the first in-the-field use of our system and we considered participation in the contest also as a first evaluation. We found some shortcomings and made a number of changes to each view in our system as we were working on the contest. Those insights and details about the analytic process we used are described in [2].

3 EVALUATION PROCESS

In a concept map portraying overall metrics that could be used to evaluate an interactive dialog information system, Scholtz defines three high level areas for evaluation: Usability (effectiveness, efficiency and user satisfaction), Performance (savings on intelligence analyst time, and accuracy) and Utility (trust, information/effort, Product quality) [6].

We have designed an evaluation process for Jigsaw that addresses metrics in all three of these areas:

1. Usability: a heuristic evaluation and usability study with simple tasks and lay users
2. Utility: a usability study involving real analysts and a mix of simple and complex tasks
3. Performance: a comparative evaluation between users and analysts with or without Jigsaw.

The first stage of our evaluation addresses usability and is motivated by the fact that Jigsaw users must be able to complete simple tasks with each view before they can perform more complex analytic activities involving multiple views. A heuristic evaluation followed by a think aloud study will be used to assess how the system supports these tasks.

	Jigsaw	No Jigsaw
Analysts	Best results expected	Better than students?
Students	Not as good as analysts without Jigsaw?	Worst results expected

Table 1: Setup for a Comparative Study.

The second stage addresses the utility of Jigsaw among our ultimate target users, the intelligence analysts themselves. During this stage we hope to use a range of straightforward and complex tasks with these users and gain valuable insight into feature improvements needed to positively impact their work.

Finally, after iterating over improvements in the first two stages, we will conduct a comparative study to understand the effect of the analysts and of the system on analytic products. This stage will involve a larger problem given to various groups of analysts and non analysts, only some of whom will have Jigsaw. In addition to collecting their final 'answers', qualitative data about their experience with or without Jigsaw will be collected.

Table 1 shows a setup for the comparative study. We expect the best results from analysts using Jigsaw and the worst results from students without Jigsaw, but we are not sure what to expect in between. Should students with Jigsaw perform better than analysts without Jigsaw?

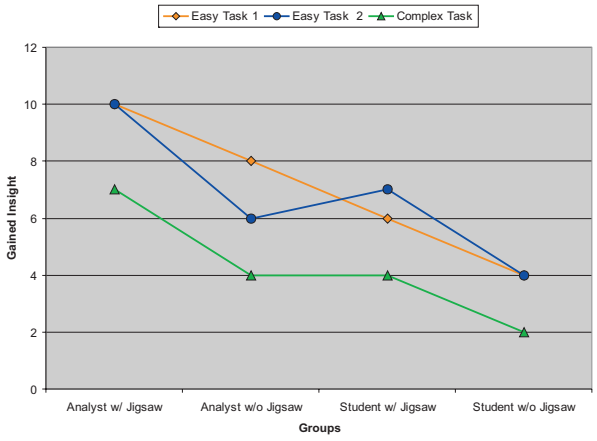

Figure 1: Different Possible Results for a Comparative Study.

Figure 1 presents more possible outcomes of a comparative study that depend on the given task. The performance results for analysts without Jigsaw and students with Jigsaw could swap for different kinds of tasks or they could be the same.

3.1 Study Logistics

Our goals in evaluating Jigsaw are to assess the basic usability of each view which includes how the views facilitate the tasks for which they are uniquely suited and how they handle tasks for which the user may think they are intended. Moreover, we would like to get a sense of how easy our system is to learn.

To address the usability of each view, we have composed a mixture of questions that can be solved with only one view or by using multiple views in concert. In addition, there are tasks for which different views can be used but one may be more efficient than another. Our tasks also range from identifying direct connections to composing a network of connections given bits of information about that network. The study will begin with basic connection identification questions and then move on to more complex questions. To maintain balance, some easier questions will also be interspersed among the later questions in the study.

To address learnability, we will lead the participant through a specific training period before the study begins. This period will consist of an explanation of the views, a demonstration of how they can be used, and a small set of practice questions for the user to complete. The length of this training period will depend on the user. The study will not begin until they have mastered the practice questions.

In addition, we have designed the study so that we can compare the user's initial behavior with the system with their choices at the end of the study. Specifically, the study will be divided into three phases:

1. Participants will be allowed to decide which view to use for each task.

2. For each task, we will tell the user which view to use. The mandatory view will be determined based on whether or not it is the best view for that specific task.

3. The final phase will once again allow the user to decide how to approach a task.

Naturally, there will be the same types of tasks in each of the three phases. This structure allows us to compare how much the users have learned about which view is best for a particular task.

We plan to use the VAST '07 contest as a dataset for our first evaluation stage, since it is publicly available and already tested to be sound during the contest. This also allows to reuse our process to evaluate usability and to compare Jigsaw to other systems.

Since the availability of significant screen space is very beneficial for using Jigsaw with multiple views, we will use a computer with four monitors in our lab to provide a good system environment in the first and second evaluation stage. The third stage will run over a longer time period and the users will be able to chose where to work.

As in any usability study, defining failure is also a challenge. Johnston's work [4] made a distinction between intelligence error and failure: the former being factual inaccuracies resulting from poor or missing data and the latter being systemic organizational surprise resulting from incorrect, missing, discarded or inadequate hypotheses. These definitions focus on user misjudgment and data deficiencies. In addition to these, we have a system centric failure definition, *i.e.* failure of the system to intuitively support a task that the user needs to complete. These could result from bad design decisions or deficiencies in system performance.

During the first stage of our evaluation, we will only ask questions for which there is a clear answer, *i.e.* no data necessary for the answer will be missing. Therefore, even though the VAST '07 dataset contains several dead ends and ambiguous information, data centric errors should not be an issue. User failure will refer to the inability to find these answers or giving an incorrect answer and system failure will be the defined as it is above.

During the final two stages when tasks become more complex, and questions can be determined by the participants themselves as they try to test a hypothesis, user failure will refer to inadequate hypotheses while system and data failures will remain the same.

Recognizing data failures will be somewhat straightforward. However, distinguishing between user and system failures will be a challenge. In some cases both system and user failures can work together to produce the wrong outcome.

3.2 Metrics for Results

We will be collecting the following data during the usability stage of our evaluation.

1. Whether or not the user is able to answer the given questions about the dataset.

2. How the participant found those answers to the questions.

3. The pitfalls they encountered along the way.

To assess system usability, we will look at the above measures in addition to task completion time. To assess learnability we will examine time to finish training and differences in how tasks are approached at the beginning of the study and at the end.

Another powerful way to gather feedback about the user session is through Jigsaw's interaction history: the user's mouse clicks and view usage in timestamped text format. Interaction history gives us the unique opportunity to triangulate what our participants say about their experiences with Jigsaw and how they actually interacted with it. Moreover, with a record of each participant's interaction with the system, we can easily compare approaches to each task across participants without having to rely on subjective assessments of each user's actions. Moreover, we can learn whether or not analysts approach tasks differently from lay users. This empirical data could prove to be a valuable way to identify current pain points in Jigsaw usage and inspire new feature ideas.

Besides covering the usability of each view, our evaluation must address the query interface that is the gateway to defining what is displayed in the views. We will be able to get feedback on this feature during the training session and in the first and third phases of our study (where users are not constrained by which view they use). We will be looking at whether the interface accurately maps to users' mental models of the scope of their search.

3.3 Questions about the VAST '07 Contest Dataset

In this section we present eight different categories with sample questions that we plan to use for the first evaluation stage. We also comment on how we expect to answer the question using Jigsaw.

1. Simple questions, one step connections between different entity types

   - What position does Faron Gardner have at the Animal Justice League?
     [Query for Gardner and AJL, find connecting report in Graph or List View and read it in Text View]

2. Simple questions, one step connections between same entity types

   - Have John Fraley and Erik Wenum worked together before?
     [As above but easier with Graph View. In the List View either multiple entities have to be selected in one list, or two people lists are needed.]

3. Questions about number of connections

   - Is the Sea Shepherd Conservation Society more connected to Greenpeace than to the U.S. Forest Service or to the Animal Liberation front?
     [Query for entities and count connecting reports in either List or Graph View.]
   - Which state is Dennis Kucinch most associated with?
     [Query for Kucinch and use monochromatic color scale in List View that shows how strong entities are connected. Using the Graph View would be more difficult: showing all connected places and counting the connecting reports.]
   - Which two entities are most associated with the Physicians Committee for Responsible Medicine?
     [As above but all entity types have to be inspected. Using the List View is getting even more efficient than using the Graph View.]

4. Questions about frequency

   - What were the five most frequently mentioned organizations across all the data that you have?
     [Use List View, add all entities to an organization list and sort by frequency.]
   - You have heard something about a man with initial 'R.A.' Does anyone in this dataset have those initials? If so, how many people?
     [Use List View, add all entities to a person list, sort alphabetically and inspect all entries starting with 'R'.]
   - You have received some news about a terrorist called Sarah Williams. You wonder if any other family members are involved. Is there anybody with the last name Williams in your dataset?
     [Query for Williams and inspect query result. Using the List View like above would be more difficult since it's not possible to sort by last name, so the whole list of people would need to be inspected.]

5. Questions about validating connections

   - How does Mary Alice Smothers know Susan Fellows?
     [Query for entities, find the connecting report in the List or Graph View and read it in the List View.]

6. Questions about indirect connections

   - What city did Mary Meekins and Paul Laurent live in?
     [Query for entities and explore connections either in the List or Graph View.]

7. Questions about dates

   - What did Ron Lawrence do on 11/27/2003?
     [Query for Lawrence and read report. It's also possible to query for the date but that's more tricky because there's no standard format for dates within reports.]
   - What 10th anniversary event was happening on 02/24/2004?
     [More difficult than the question above since query has to be for the date.]
   - When did the killer whale Luna nearly collide with a landing float plane?
     [Query for Luna, read the report and compare with publication date of the report in the Text View – the reports says 'Luna nearly collided two days ago'.]

8. Questions about report and entity density

   - On what day were the most reports published?
     [Load all reports in the Calender View and inspect the days.]

It has been more difficult than we expected to find suitable questions in the VAST '07 contest data set. Even though the data set contains about 1500 documents, it is challenging to find small networks with indirect connections, because on the one hand most entities have only few connections but on the other hand few entities (like 'U.S.') have very many connections and the search space would explode if they are included in a search.

3.4 Additional Remarks

In discussing metrics for evaluating human information systems [6], Scholtz encourages a shift from focusing on usability and performance along to one that includes measures of utility and impact. It is important to note that all of our evaluation plans so far take place outside of analysts' actual environment, which makes it difficult to assess impact comprehensively.

We can address this blind spot in two ways. First, we could introduce a fourth stage of evaluation focusing on analysts in their native environment or in a closely simulated environment. Second, and more simply, we suggest that it is also helpful to evaluate impact based on a system's ability to address pain points that already exist in the intelligence culture as outlined in Johnston's Analytic Culture in the US Intelligence Community [4]. For example, one observation Johnston made was that analysts felt that most of their day was spent simply reading and writing rather than analyzing. It is conceivable that with a tool like `Jigsaw`, even the simple act of daily reporting could become more of an analytical exercise.

On the management side, one can imagine daily briefings and reports from analysts containing insights that can be recognized more quickly. For instance, a screen shot from `Jigsaw`'s graph view showing repeated connections between a suspicious character and an organization across multiple reports whose dates are in a particular range could be more effective than a snippet of text.

There is an additional category of task that we have not mentioned, *i.e.* those tasks for which there is no answer in the dataset. Realistically speaking, there are times when insufficient or conflicting data hinder an analyst's ability to give a strong answer to a particular question. Should this use case be covered in all stages of our evaluation or will this cause unnecessary frustration on the part of our participants? One approach could be to use these types of tasks in the second and third stages of evaluation where tasks will be more complex and dead end results may be more expected.

4 Reflections and Conclusion

It is important that as researchers, we remember that our goal in visual analytic evaluation is two fold: to improve the system we are evaluating and to improve the life of the analyst. It may be easy to get caught up in the numbers we are collecting but ultimately, there must be positive impact for the analysts.

In composing this evaluation plan for `Jigsaw`, we recognized many current and future challenges. First, `Jigsaw` is a large system with multiple views that facilitate varied interaction. How can we ensure full system coverage in our usability study? Since many screens are involved in our lab setup, how many screens should the participant be allowed to work with? Should we determine that amount or should the users, depending on what they are comfortable using? What effect would this variable have on the outcome of the study? Would too many monitors overwhelm the user and hinder work or would the opposite occur?

Additionally, would it be useful in these early stages of evaluation – where usability is our intended focus – to have a control group of users who work with the same dataset but only have Google's search functionality and Microsoft Word as tools for working on the same tasks?

Finally, there seems to be a fine line between evaluating the `Jigsaw` user and evaluating `Jigsaw` itself. We must be careful to present our results from a system failure rather than a user failure point of view.

While we do not yet know whether the results of this study will be generalizable to other visual analytics systems, our approach could be applied to other evaluation in two ways. First, we have composed a set of tasks of varying difficulty on the VAST '07 contest dataset. These questions can most likely be reused in another study regardless of the system being tested. Second, our three step evaluation process can be reused, starting from benchmark, atomic tasks in a small scale usability study and ending at larger, more complex analytic study among one or more groups of participants. It may be tempting to start evaluating a system using intricate problem solving tasks. However, evaluators need to ensure that basic usability is in place before jumping to this step.

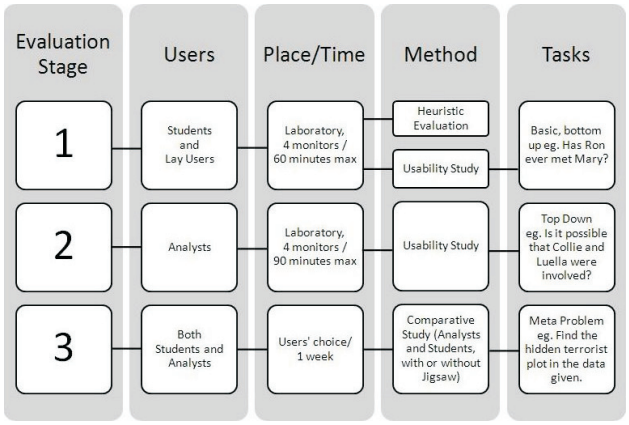

Figure 2: Evaluation Plan.

In this paper, we presented a plan to evaluate our visual analytical system `Jigsaw`. Figure 2 summarizes and compares the three stages of our plan.

Acknowledgements

This research is supported in part by the National Science Foundation via Award IIS-0414667 and the National Visualization and Analytics Center (NVAC^{TM}), a U.S. Department of Homeland Security Program, under the auspices of the Southeast Regional Visualization and Analytics Center. Carsten Görg was supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD).

References

[1] S. Card and P. Pirolli. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Through Cognitive Task Analysis. In *Proceedings of 1st International Conference on Intelligence Analysis*, 2005.

[2] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko. Jigsaw meets Blue Iguanodon - The VAST 2007 Contest. In *IEEE Symposium on Visual Analytics Science and Technology*, 2007.

[3] Jigsaw project. `http://www.gvu.gatech.edu/ii/jigsaw/`.

[4] R. Johnston. *Analytic Culture in the U.S. Intelligence Community – An Ethnographic Study*. Central Intelligence Agency, 2005.

[5] J. Scholtz. Beyond Usability: Evaluation Aspects of Visual Analytics Environments. In *IEEE Symposium on Visual Analytics Science and Technology*, 2006.

[6] J. Scholtz. Metrics for evaluating human information interaction systems. *Interact. Comput.*, 18(4):507–527, 2006.

[7] J. Stasko, C. Görg, Z. Liu, and K. Singhal. Supporting Investigative Analysis through Interactive Visualization. In *IEEE Symposium on Visual Analytics Science and Technology*, 2007.