Evaluating Visual Analytics: The 2007 Visual Analytics Science and Technology Symposium Contest

Catherine Plaisant
University of Maryland
Georges Grinstein
University of Massachusetts Lowell
Jean Scholtz, Mark Whiting
Pacific Northwest National Laboratory
Theresa O'Connell, Sharon Laskowski
National Institute of Standards and Technology
Lynn Chien, Annie Tat, William Wright
Oculus Info Inc.
Carsten Görg, Zhicheng Liu, Neel Parekh, Kanupriya Singhal, John Stasko
Georgia Institute of Technology

Abstract

The second Visual Analytics Science and Technology (VAST) contest's data consisted of a heterogeneous synthetic collection of news articles with additional supporting files and contained a scenario with embedded threats that provided ground truth. Using visual analytic tools, participants sought evidence of illegal and terrorist activities. We describe the contest and the evaluation methodology. We report results and lessons learned.

Keywords: evaluation, visual analytics, human information interaction, sense making, metrics, contest.

1. Background

The second VAST contest [VAST07] ran from March through July 2007, in conjunction with the Visual Analytics Science and Technology (VAST) 2007 Symposium [VAST]. Its objectives were to provide the research community realistic tasks, scenarios and data used in analytic work, to help visual analytics (VA) researchers evaluate their tools, and to improve and enrich interactive visualization evaluation methods and metrics. Competitions of this sort have been useful in other domains, e.g., the Text REtrieval Conference (TREC) [TREC], the Knowledge Discovery and Data Mining Cup [KDD], the Critical Assessment of Microarray Data Analysis [CAMDA], and the IEEE InfoVis contests [Pla07].

This paper was co-authored by the contest committee members and the authors of the winning entries. In this paper, the pronoun "we" refers to the contest committee, except in section 6 where the "we' refers to the contest winners (6.1, Oculus; 6.2, Georgia Tech).

We discuss the contest, the evaluation methodology, the results, the two winning entries and key lessons learned.

2. Dataset and tasks

Participants received a dataset developed by the National Visualization and Analysis Center (NVAC™) Threat Stream Generator project team at Pacific Northwest National Laboratory (PNNL). This dataset contained:

- 1500 news stories simulating an online news archive
- Two blog excerpts with entries extracted over time segments
- Ten pictures (in jpg format)
- Three small databases (in XLS and CVS format)
- Reference information (in Word and PDF formats).

The synthetic dataset included data from analyst-generated scenarios, providing ground truth. The 2006 and 2007 datasets consisted mostly of text, but in 2007 we included new data types (e.g., blogs, hand-drawn images, and images with annotations), several major subplots or scenarios instead of one, and information gaps which required teams to identify and deal with incomplete data. Information about the plots was spread across differently formatted documents requiring participants to make links or associations for evidence in support of hypotheses.

Participants received background information about the task plus some directions. The dataset was provided under a closed-world assumption: contestants did not need to go outside the dataset to generate hypotheses and collect supporting evidence. No domain expertise was required.

Participants chose between using raw data or preprocessed data with extracted entities such as names, places, times, or money. They could use tools and visualizations they developed or off-the-shelf tools. Their task was to find the major plots embedded in the dataset, identifying the individuals involved in suspicious activities, the time frame for these activities and provide a list of important events. Participants had to write a debrief describing the situation, identify the associated key documents used in their hypothesis generation, and suggest recommendations for further investigations based on

theories developed in their analysis. A form was provided to report that information and the process used. Screen shots and a video were required to highlight insights provided by the tools and to facilitate entry evaluation. After the contest, we collected anonymous responses to a survey on participants' experiences in the contest.

3. Timing and incentives

The dataset was made available in early March 2007 and the July 13th deadline gave participants approximately four months to prepare their entries.

To encourage participation, we provided incentives. Top entries received awards and presented their work at a VAST Symposium panel. There were awards for best entries and appreciation certificates for all other entries. The top entries were invited to an interactive session at the symposium where professional analysts worked with them on a new task. Finally, the winners coauthored this journal paper with the contest committee. All accepted entries are at the Information Visualization Benchmark Repository [INFOVIS] with links from the VAST contest page and a NIST webpage [NIST].

Because a great deal of work goes into submitting to the contest and those entries become a community resource, we wanted to provide broad recognition for contributions. In 2007, we invited all participating teams to present posters at VAST and publish a two-page summary in the proceedings.

4. Judging

While most scoring criteria remain subjective, ground truth made it possible to calculate quantitative scores for accuracy of responses about the embedded threats, specifically, major players and activities, and where and when the activities occurred. The accuracy scoring was done first, then judging took place with six experts in VA, human-computer interaction (HCI) and visualization and seven professional analysts.

The accuracy scoring was done by hand, a first step toward automation which we anticipate will be available for future contests. The dataset included elements (i.e. people, events, location) that are part of the plots and some that are not. We defined the variables T_P, T_N, F_P and F_N for each type of element as follows:

- T_P: Number of correctly identified plot elements (i.e. True Positive)
- T_N: Number of elements reported as not part of the plot (e.g. True Negative such as innocent bystanders identified)
- F_P: Number of elements incorrectly reported as part of the plot (False Positive)
- F_N: Number of unidentified plot elements (False Negative)

The scores were then calculated using the following:

- Who score = $T_P + T_N - \frac{1}{2} (F_P + F_N) + 0.25 *$ Association, where Association is defined as the number of correctly identified person/organization pairs
- What/ When score = $T_P + T_N - 1/2\, F_N$
- Where score = $T_P + T_N - \frac{1}{2} F_N - \frac{1}{4} *$ Not Specific, where Not Specific is defined as correct but too broad (such as the country instead of the city)
- Debrief score = Correct key events – missing events. Key events were defined as events associated with persons required for correct debriefing, not just the events reported correctly.

A few days before meeting, the 13 judges received the ground truth: the contest scenario, a timeline of events by plot and subplot, pointers from the subplots to items in the dataset that captured the ground truth, and a social network diagram. They also received the accuracy scores and rating sheets. Judges attended a parallel meeting held on the U.S. East and West coasts. They rated each visualization and assessed the utility of each of its components. University entries were judged separately from corporate entries. During the meeting, all judges finished their reviews and discussed individual entries using the following criteria

- Accuracy (correctness of answers)
- Quality of the debrief , evidence provided
- Overall system utility
- Visualization quality
- Process description

Then each judging team summarized their conclusions and suggested winners. Activities were coordinated with conference calls. Over the next few days, the contest committee made the final award decisions and prepared evaluation summaries for each entry.

5. VAST 2007 Contest Results

We received seven entries, three from academia. The two winners were:.

Winner , Corporate Category: Oculus Info Inc.
Winner, University Category: Georgia Tech.

Both entries had high accuracy in answering the who, what, when and where questions (over 85%). Ratings of the systems' utility and visualization quality were high.

Judges gave high ratings to Oculus Sandbox's support for the analysis of competing hypotheses (ACH), its emphasis tool and its auto template feature. Oculus provided two different time/space visualizations of the data (done by two different analysts) to show the tool's flexibility.

Judges gave high ratings to Georgia Tech's coordinated views and the automated graph layout feature. The process description was clear, illustrating the system's useful features and those needing improvement.

We also gave a *Best Debriefing Award* to the University of British Columbia and Simon Fraser University [Cha07]. This team, using off-the-shelf tools and a great deal of manual labor, had the best accuracy. They identified all the major threads and players, and prepared an exemplary

debrief, albeit the manual process does not scale, an important contest measure.

Selection of the two winning teams and a third team from Applied Technology Systems [Sch07] for the interactive session was based on the quality of their entries and on system robustness. Software had to be capable of processing a new dataset in less than 30 minutes and to reliably handle two hours of intensive use.

6. Summary of the winning entries

6.1 Winning corporate entry: nSpace and GeoTime – Oculus Info Inc.

nSpace and GeoTime are systems that are currently in research and development at Oculus with the support of the Intelligence Advanced Research Projects Activity (IARPA) [Wri06]. These were developed in collaboration with analysts and are being used and evaluated by analysts on a day-to-day basis. nSpace is a general analysis tool for unstructured massive datasets. GeoTime focuses on structured data with geo-temporal registration.

nSpace combines TRIST (The Rapid Information Scanning Tool) and the Sandbox, while integrating advanced computational linguistic functions using a Web Services interface and protocol [Pro07]. For the VAST contest, the HNC/Fair Isaac linguistics system was used. nSpace aims to support every step of the process of analysis. TRIST is a massive data triaging tool with capabilities such as planned query execution, automatic information extraction, and customizable multi-linked dimensions, which help provide rapid scanning, result characterization and correlation. Information, including full documents, text fragments, and entities, gained from TRIST, can be dragged and dropped into the Sandbox for evidence marshalling and further analysis. The Sandbox is a flexible and expressive thinking environment for analysts to visualize their cognitive mechanisms. It supports both ad-hoc and more formal analytical sense-making through capabilities such as "put-this-there" cognition, automatic organization of evidence, assertions and evidence assembly, and ACH. Analysts alternate between TRIST and the Sandbox to continually develop and refine their analyses.

GeoTime focuses on interactions between entity movements, events and relationships over time within a geospatial (or any conceptual) context to amplify the concurrent cognition of time and space. Entity behaviors and relationships, along with their patterns in both space and time are easily identified. Entities and events are charted in a single interactive three-dimensional view [Kap04]. The ground plane is the geographic space represented by the X and Y axes; the vertical T-axis represents time. Events can be animated in time.

We used an iterative, multi-threaded workflow to analyze the VAST dataset. Analysis tasks are described below in a sequential form, but, in practice, analysts can jump back and forth from one analysis activity to another.

6.1.1 Brainstorming and querying for information

Before analysis, we used the Sandbox to plan out the process. We gathered and organized contest instructions; generated questions and keywords for querying; and annotated notes, thoughts, and prior knowledge. The entire dataset of 1,600 documents was indexed and loaded into nSpace, then reviewed using TRIST. A series of exploratory queries were refined as understanding grew. Visualization of changes highlighted query refinement results.

6.1.2 Scanning data for characterization and correlation

Information retrieved was viewed within TRIST's multi-dimensional framework. The date-published dimension provided a time range of the information objects; topics were automatically extracted in another dimension. The automatic cluster dimension was a great place to start. Automatic entity extraction of people, places and organizations provided additional dimensions. In each dimension, rows of icons or charts representing information objects showed the distribution of information across the categories in each dimension. User-tailored dimensions were added as key topics and entities emerged.

Relationships between key issues, topics, events, players, organizations and locations were tracked by viewing and linking the dimensions side by side (Figure 1).
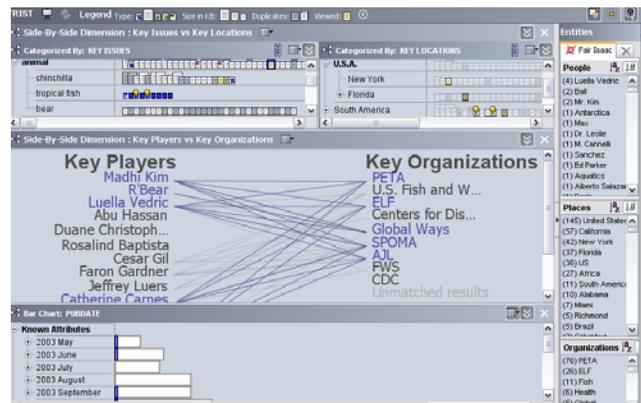
Figure 1. **Triaging key topics and relationships in TRIST**: **Clicking on the key issue "tropical fish" highlighted all key players and organizations associated with the issue, and the relationships between the players and organizations.**

6.1.3 Examining geo-temporal structured data

The Animal Import database, with more than 2,000 transactions, was transferred into GeoTime to further investigate connections between suspicious players and organizations. Using GeoTime's charting and filtering tools the analyst quickly reviewed exporters' behaviors and identify suspicious patterns. GeoTime's link analysis tools allowed the quick review of history and connectivity of suspicious exporters.

6.1.4 Reading documents and exploring relevant issues

After identifying pertinent players, organizations, issues and locations, analysts began to read relevant documents using the document workspace in nSpace. Entities and search terms were automatically highlighted in documents to facilitate quick identification of key content. As important events were discovered and transferred to the Sandbox, we created an evidence marshalling timeline view there. To understand patterns and behaviors, events were then transferred into GeoTime and plotted in GeoTime's animated three-dimensional view (Figure 2).
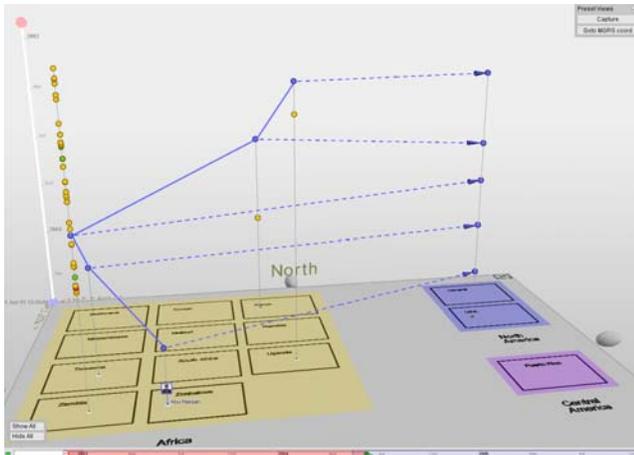
Figure 2. Tracking entity transactions in GeoTime: Countries in the Animal import database were mapped conceptually to better focus on relevant events.

6.1.5. Assembling data to synthesize information and interpret findings

Throughout the analysis, discovered insights and supporting evidence were saved and annotated in the Sandbox. Sandbox tools, such as links, entities, and groups, helped us organize data. We created layouts that corresponded with our mental models, e.g., various social networks and hierarchies (Figure 3a). The Sandbox supported the development and assessment of meaningful hypotheses, captured as assertions. Evidence for assertions was marshaled through evidence gates by dragging and dropping supporting or refuting evidence from the left and right side of an assertion (Figure 3b). Conflicting evidence and competing assertions were analyzed using the ACH tool, which helped clarify and diagnose the strength of evidence and hypotheses (Figure 4).

6.1.6 Collaborating to enhance analysis and verify conclusions

Sometimes several analysts worked on the dataset simultaneously needing to share their work. Various elements of content in nSpace were imported from one workspace into another. User-defined dimensions, such as Key Players and Key Locations, and templates were exported, shared and then modified to help uncover additional unexpected patterns. We were able to corroborate answers quickly because findings were presented easily using nSpace. When in presentation mode, nSpace menu items became hidden and both TRIST and Sandbox showed findings in a single display. Tools such as the Powerful Finger temporarily enlarged items when the Sandbox was at low zoom. Bookmarks quickly set and re-accessed desired views.
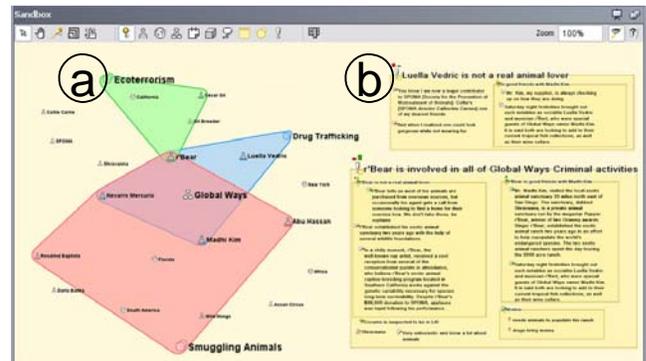

Figure 3. Information synthesis in the Sandbox: (a) A link diagram between key entities and overlapping color-coded sets helped identify plot lines, (b) Assertions with weighed evidence helped identify villains.

Figure 4 Analysis of Competing Hypothesis: The diagnosticity of individual pieces of evidence helps verify which hypothesis is the most probable.

6.1.7. Reporting

Final reports were generated by dragging and dropping objects from TRIST and the Sandbox directly into MS Word or PowerPoint, preserving formatting. Sources and references for documents and fragments were automatically added.

6.1.8. Lessons learned

nSpace supports analytical work-in-progress. It can be tailored to adapt to both the task at hand and analysts' cognitive mechanisms. Every step of an analyst's workflow is assisted with automated and user-driven analytical visualizations. nSpace suited the contest data types. The new ACH tool added utility.

GeoTime enabled the analysis of information connectedness over time and geography within a single, highly interactive 3D view. Expected relationships were observed and investigated from various viewpoints and unexpected patterns were discovered without having to wade through multiple spread-sheets, tables, maps, and other cross-referenced data that are often simultaneously needed during analysis.

Still, we uncovered problems by working through this exercise. We were able to overcome some; others have been incorporated into future development plans. Records in CSV files, when imported into nSpace, were processed as a single field rather than separate records with multiple fields. This is a feature that may be implemented in the near future. We observed that while nSpace is a rich tool with many capabilities, for a novice user, a simpler version would be less overwhelming.

The connection between nSpace and GeoTime included some manual steps and required two analysts to work independently. Future goals include linking the two applications for more efficient collaboration.

In conclusion, nSpace and GeoTime proved to be powerful systems that enabled novice analysts to perform full analysis of real-world problems with proficiency. The tools proved to work well both independently and collectively. The VAST 2007 contest provided an excellent opportunity to test and refine advanced visual analytic capabilities.

6.2 Winning academic entry: Jigsaw - Georgia Tech

Jigsaw [Sta07] is a visual analytic system that provides multiple coordinated views to show connections between entities extracted from a document collection. It was developed during 2006 and 2007 at the Information Interfaces Lab at Georgia Tech. Usability and simplicity are key design features to make Jigsaw's operations intuitive and easy to use.

Jigsaw presents information about documents and their entities using different types of views. A large amount of screen space is beneficial, so we used a computer with four monitors. Jigsaw's views do not show the entire dataset at once but use an incremental query-based approach to show a subset of the dataset. At start-up, the views are empty. Analysts can populate them either by querying for entities or by expanding visible entities. This approach allows Jigsaw to operate on datasets where simply showing all the entities would result in a crowded and cluttered display. Five views can be instantiated multiple times:

- The Text View displays documents, highlights the entities within them, and shows how often a document has been displayed.
- The List View draws links and uses colors to show connections between entities organized in lists. Users can specify the number of lists to be displayed and the type of entity shown in each list. They can choose among sorting options.

- The Graph View displays connections between entities and documents in a node-link diagram. It supports step-by-step network exploration by expanding and collapsing nodes to show or hide their connected entities or documents.
- The Scatter Plot View highlights pair-wise relationships between any two entity types, showing co-occurrence in documents.
- The Calendar View provides an overview of the documents and the entities within them with respect to the publication date of the documents.

Views are coordinated using an event mechanism: interactions with one view (selecting, adding, removing, or expanding entities) are transformed into events that are then broadcast to all other views. Thus, the views stay consistent and provide different perspectives on the same data. Jigsaw also has an option to freeze the view by turning event listening off in case a specific view shows an interesting data subset that should not be changed.

Figure 5 shows four views on the contest dataset. The List View shows the people and organizations to which r'Bear is connected. Orange coloring and lines between entities in neighboring lists indicate entity connections. The Text View shows a set of four documents that mention Luella Vedric with entities highlighted for easier scanning. The Scatterplot View displays person-by-place connections with the central diamonds indicating reports that include the pairs of entities. The Calendar View gives an overview of documents published in May to October 2003 and shows the entities of one selected document on the side. (Jigsaw's Graph View is not shown on Figure 5, but is illustrated in Figure 6.)

Jigsaw does not have capabilities for finding themes or concepts in a document collection. Instead, it acts as a visual index, helping to show which documents are connected to each other and which are relevant to a line of investigation. Consequently, we began working on the problem by dividing the news report collection into four pieces (for the four people on our team doing the investigation). We each skimmed the 350+ reports in our own unique subset to become familiar with general themes in those documents. We took notes about people, organizations or events to potentially study further.

We wrote a translator to change text reports and pre-identified entities from the contest dataset into the xml form that Jigsaw reads. We then ran Jigsaw and explored a number of potential leads identified by our initial skim of the reports. First, we looked for connections across entities, essentially the same people, organizations or incidents being discussed in multiple reports.

Surprisingly, there was relatively little in the way of connections across entities in the documents. After a few hours of exploration, we had no definite leads, just many possibilities. So we returned to the text reports and some team members read subsets of the reports they had not examined before. At that point, we began to identify potentially "interesting" activities and themes to examine

further. It became clear that the time we spent earlier exploring the documents in Jigsaw was not wasted. It helped us become more familiar with many different activities occurring in the reports. We were curious, however, why some connections did not show up in Jigsaw initially. Returning to the system, we learned why. Some key entities were only identified in some of the documents in which they appeared or they were not identified at all. To remedy this, we created software to scan all the text documents and identify missed pre-identified entities. This resulted in adding more than 6,000 new entity-to-document matches. Thus, the entity-connection-network became much denser. This also added noise by multiplying unimportant or wrongly extracted entities. Therefore, we manually removed the false positive entities (wrongly classified or extracted) that occurred with high frequency for each entity type, and we identified a few entities that had been missed.
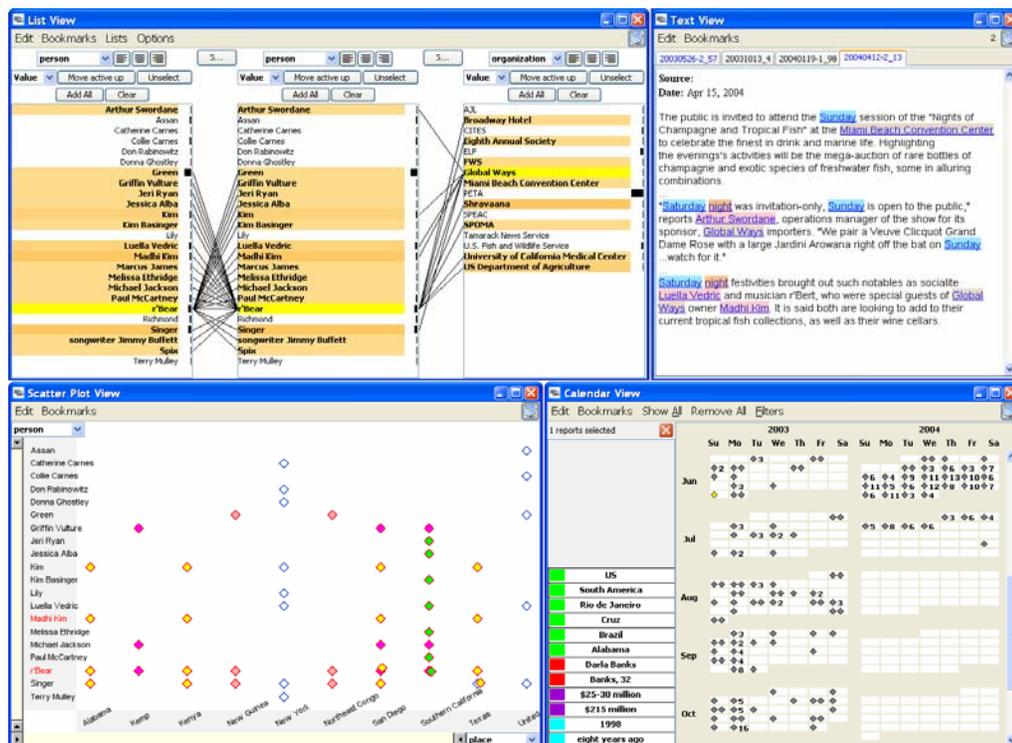

Figure 5: The List View, Text View, Scatterplot View, and Calendar View show data from the contest.
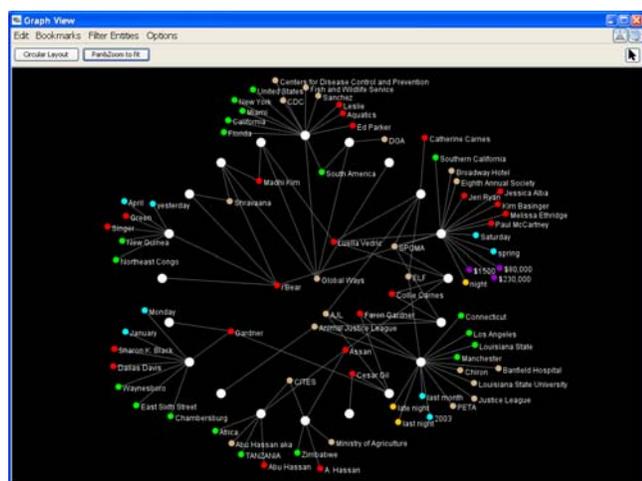

Figure 6. Circular Layout in the Graph View. All entities connecting to more than one document are drawn in the middle making it easier to focus on them.

This process provided us with a consistent connection network that was mostly devoid of false positives. Since less than one quarter of the entities across the entire collection appeared in more than one report, we added an option in Jigsaw that allows the user to filter out all entities that appear in only one report. This allowed us to focus on highly-connected entities at the beginning of the investigation and to add entities when more specific questions arose later during the analysis. When we resumed exploring the documents using Jigsaw, it was much easier to track plot threads and explore relationships between actors and events given this refined entity information.

Having multiple instances of one view helped us during analysis. We could follow different threads in parallel during our investigation. When two threads seemed to merge or seemed to be related, we could create a new view containing the entities of both threads and investigate further with this combined and extended set of entities. Having multiple instances of one view was useful in another context: we used multiple instances of the Text View to group related documents (e.g., multiple open Text Views each showed the documents about a key person in the plot).

The Graph View helped immensely in finding highly connected entities. This view provides a special layout operation that repositions all the visible reports equidistant around a large circle in the view (Figure 6). Entities connecting to only one report are drawn near that report, but outside the circle. Entities connecting to more than one report are drawn inside the circle with the entities having the most report connections going closer to the center showing that they may be related in important ways and likely should be examined more closely.

Another useful approach was to increase the context of an entity collection step-by-step. We began with a small entity collection in the Graph View, expanded these entities to get the documents in which they occur, and then expanded those documents to identify their entities. Next, we explored this new entity and document set in more detail in one of the other views. By reading the documents, we decided which new entities would be relevant. We removed the ones that were not of interest and then we continued to expand the context.

Our investigation disclosed a drawback: Jigsaw lacks functionality to change extracted entities on the fly. Since the system uses the co-occurrence of entities to build the connection network, it is crucial that entities are properly identified. Missing or unidentified entities result in a knowledge gap: connections that are not there cannot be visualized. To overcome this shortcoming, we plan to extend Jigsaw's functionality to allow the user to add, remove, or change extracted entities dynamically.

Our interactions with an analyst in the interactive session further showed how Jigsaw's utility is limited when facing a new, unexplored document collection without identified directions or themes. Jigsaw works better in helping to explore potential leads and uncover chains of connected entities and events given some initial exploration. We also had to show the analyst certain strategies that we have learned to most effectively use the different views.

The analyst felt that the Graph View and its circular layout operation were particularly helpful. He liked how entities were highlighted in the Text View to allow quick skimming of the documents. He also felt that a new evidence marshalling view was beneficial. It allowed "interesting" entities to be dropped in, comments registered about each, and hypotheses to be developed.

Overall, our participation in the contest was extremely beneficial in terms of improving and updating the Jigsaw system. Having such a large, high-fidelity dataset available was invaluable in numerous ways and it particularly allowed us to observe the utility of the different views in an actual investigation scenario. We learned more about the system while working on the contest than we did at any other time. This knowledge motivated us to fix usability problems, create new operations and views, and consider future avenues for growth and expansion.

7. Interactive session

The Oculus, Georgia Tech and ATS teams participated in the interactive session using a smaller synthetic dataset. Each team worked with one analyst who provided feedback about the tools. Thirty minutes were allocated to train analysts on their system. We recommended that one of the developers use the system while the analyst lead the analysis. After training, analyst/developer teams had two hours to make as much progress as possible on the task. One observer sat with each team, taking notes on the process used and comments made. When an interesting event happened, the observer photographed the screen(s). These notes were provided to the teams. An additional analyst and other observers (including dataset and scenario creators) moved between teams to observe and compare their work and progress.

It appeared difficult for some analysts to get started. While they seemed to understand the systems' capabilities, it seemed difficult to understand how to best employ them. As the developers had considerably more experience using the systems and also some experience working on the type of tasks presented in our datasets, they gave suggestions to the analysts on how to get started. Analysts often tried to use the system themselves, with some success, but developers mostly "drove" the UI so that everyone could focus on the task instead of focusing on learning the UI. Still, we observed that each team quickly became a unified set of analysts as its members started to get engaged in the task. Sometimes, different team members performed different tasks in parallel to speed up the process.

Analysts' observations helped us plan refinements to the evaluation metrics.

8. Lessons learned

The 2007 contest saw improvements over 2006. In this second contest, participants had the advantage of having access to the 2006 task and solution for practice before release of the 2007 contest data, and could look at past submissions. This clearly made a difference as the analyses were much improved over those received in 2006. We also provided better guidelines and criteria for our evaluations of the quality of the visualizations.

We identified areas for improvement in future contests. Providing a preprocessed version of the dataset allowed participation by two teams that did not have access to text analysis capabilities so we will continue and refine the data preprocessing in the future. We had intended to score entries separately depending on whether the entries used the preprocessed data or not, but, with only seven submissions, this was not necessary.

In 2007, for the first time, we calculated quantitative accuracy scores based on the ground truth and used systematic scoring guidelines and ratings both for the utility of the systems' components and the quality of the visualizations. Thus we could provide participants with more detailed comments on their entries.

Several entries focused on describing their tools' functionality and features instead of the overall process. What was more important to us was a description of the usage of the tool in the context of the analytical task. This is a common problem we encountered in other contests [Pla07]. More guidelines and explicit examples of good and bad process descriptions need to be made available.

The committee HCI and visualization experts developed the utility and visualization quality ratings during the year and tested them with entries from the previous year. However, these metrics were created for UI evaluation, an area in which most analysts are not trained. Some of the professional analysts preferred providing overall comments on the quality of the debriefings, the overall utility of the systems and the scalability of the process used. The committee members thus focused more on scoring the visualizations' quality and used the analysts' valuable comments to rate the systems' utility in the context of use. Because the ratings were not used consistently, the teams received only comments, but we intend to refine our rating criteria and procedures to be able to report more objective and consistently derived quantitative scores.

Submissions reflected the diversity of possible problem-solving approaches. One student team used only off-the-shelf tools. In the process of working on the problem they learned a great deal about the analysis process and are now better prepared to become effective VA developers.

Participants stated that the number one reason for entering was for systems evaluation, namely to test their tools and find potential problems with their systems. Teams found the effort worthwhile and recommended participation. One team's comment typified the reasons, "It gave us a great metric on how our tools are used and a good excuse to document how they are used. It gave us a feel for how long an analysis takes using our utilities and what its strengths/weaknesses are (and where we can improve)."

Participants said that the time spent in producing their entry was definitely worth it. Several noted that producing the video was time consuming. Videos are indispensable for the judges to understand how the tools work so this problem will remain. Others noted that producing a debrief reflecting all the subtleties took time. We hope that the availability of more examples of successful debriefs will help participants. Teams noted that the data had inconsistencies (resembling real world data).

In the survey, three teams anonymously provided details about time and effort spent. Two reported in terms of team hours: one spent about a week; another approximately a month. One team spent 56 person hours. One team read every document; another read 25% of them; a third did not count documents read. Teams reported time-consuming problems including tying together non-text, unstructured data; hitting dead-ends; and resolving data inconsistencies. While we do not know if the effort put forth by these teams is typical across all entries, we are satisfied that their effort solving the problem is reasonable given the contest's four month time span.

Teams who participated in the interactive session reported learning much about analysis while working with a professional analyst. In some cases, it was the first and only time they ever had this opportunity.

Analysts learned about new technologies. Working on a problem together as a team with the developers seemed more effective to understand the potential of a tool than seeing any number of demonstrations, even though the exercise only took a couple of hours. This method may be useful for workplace training. Analysts' experiences interacting with these tools emphasized the importance of designing for usability. High usability reduces training time and makes visualization tools more attractive to software purchasers.

9. Panel discussion at the symposium – input from the community

The contest panel at the VAST Symposium consisted of one or two representatives of each of the three interactive session teams, an analyst from the Department of Homeland Security who had also participated in the interactive session, and the contest committee. We summarized the contest and the interactive session, with each team's representative describing their system, then took questions. There were questions about the specificity of the dataset to homeland security. The answer highlighted the broad applicability of the systems that did participate and pointed out that the type of task was also applicable to business intelligence, patent and intellectual property monitoring and exploration, or publication analysis in medicine.

Other questions addressed the perception that only teams that already have developed extensive tools for the

type of task proposed could possibly participate. We commented that one team had used only off-the-shelf tools to solve the task, and that a combination of existing and new tools could be used. We also noted that we encourage participants to find partners with complementary skills and tools and offer to act as matchmakers if participants ask for help. Some comments from the audience made clear that VA researchers are working on tools that address different tasks than the type proposed in the contest (i.e. a "Who done it" scenario) or with data types not included in the dataset (e.g. video). While it is our goal to expand the contest's scope progressively, we discussed the challenge to keep the contest simple enough to first increase participation and refine our evaluation methodology. Other comments addressed the "risk" in not succeeding or inquired about next year's dataset and task.

10. The path forward

We have been encouraged to see that about 75 individuals or organizations have downloaded the 2007 datasets so far. Future datasets may be larger with more data types, uncertainty and deception. We may increase scenario complexity. The challenge is to keep the problem accessible to enough people. For this reason, we will continue to provide various versions of the data with multiple levels of preprocessing.

For 2008, we would like to explore providing coherent subsets of the data that can be processed, visualized and analyzed independently from the rest of the data. For example, a social network of name entities can be extracted and made available for analysis. This analysis would not yield a complete task solution, but would allow teams working on network VA to participate at some level. Revealing elements of the ground truth plots through this focused analysis will make it possible to generate accuracy scores for the social analysis component. Analysis of imagery components, videos or table data can also become independent components of the contest. This reorganization of the contest data would both increase participation and provide a more competitive base of tools and components.

We plan to refine and publish our evaluation criteria prior to the 2008 contest. We will analyze our results to determine correlations between the qualitative and quantitative measures and metrics. Our goal is to use the quantitative scores along with the qualitative comments to arrive more rigorously at objective overall scores for the VAST 2008 contest.

Competitions include an inherent risk for participants who may fear not doing well. Several directions are possible to address this problem. We hope to provide a Web utility for participants to upload their answers and automatically receive an accuracy rating, allowing them to gauge how they are doing. This requires again that we reorganize the dataset into components that can be computationally evaluated independently. Another possibility is to move from the current competition mode where only winners present their work to a workshop where all participants who worked through the problem present, discuss promising solutions, and work on sharing or combining their tools.

Based on our experience so far, we encourage researchers and potential contest participants to:
- Participate in contests (all who participate report many benefits)
- Find partners if you cannot complete the task alone or ask assistance to find partners
- Use off the shelf tools for components you did not develop but need

We encourage VA tool users to:
- Help prepare sanitized and interesting benchmark datasets and tasks
- Offer your help to generate ground truth for benchmark datasets
- Promote evaluation activities
- Evaluate tools and train analysts while working on realistic analysis exercises

We believe that we have developed useful resources for researchers and are beginning to understand how to better evaluate VA tools. Competitions encourage the community to work on difficult problems, improve their tools and build baselines for others to build upon. We need to accept that although we would like to automate as much of the judging process as possible, human evaluation will still be required in the future. We will continue to evolve a collection of datasets, scenarios and evaluation methodologies that reflect the richness of the many VA tasks and applications.

References

[Cha07] William Chao, Daniel Ha, Kevin Ho, Linda
    Kaastra, Minjung Kim, Andrew Wade, The Bricolage,
    Proceedings of VAST 2007, IEEE Press, pp. 239-240

[Kap04] Kapler, Thomas, and William Wright, GeoTime
    Information Visualization, IEEE Symposium on
    Information Visualization 2004. 25 - 32

[Pla07] Plaisant, C., Fekete, J. D., Grinstein, G., Promoting
    Insight Based Evaluation of Visualizations: From
    Contest to Benchmark Repository, Transactions on
    Visualization and Computer Graphics (to appear)

[Pro07] Proulx, P., Chien, L., Harper, R., Schroh, D.,
    Kapler, T., Jonker, D., Wright, W., nSpace and
    GeoTime: A VAST 2006 Case Study," *IEEE Computer
    Graphics and Applications*, vol. 27, no. 5, pp. 46-
    56, Sept/Oct, 2007

[Sch07] Schwendiman, L., McLean, J., Larson, J., Data
    Analysis Using NdCore and REGGAE, Proceedings of
    VAST 2007, IEEE Press, pp. 245-246.

[Sta07] J. Stasko, C. Görg, Z. Liu, and K. Singhal. Jigsaw:
    Supporting Investigative Analysis through Interactive
    Visualization. In IEEE Symposium on Visual Analytics
    Science and Technology, pp. 131-138, 2007.

[Wri06] Wright, W., Schroh, D., Proulx, P., Skaburskis. A.,
    and Cort, B., The Sandbox for Analysis – Concepts and
    Methods, ACM CHI 2006. pp. 801 - 810

Web links:

[CAMDA] http://www.camda.duke.edu/
[INFOVIS] http://www.cs.umd.edu/hcil/InforvisRepository
[KDD] http://www.kdd.org/
[NIST] http://www.vac.nist/gov
[TREC] http://trec.nist.gov/
[VAST] http://conferences.computer.org/vast
[VAST07] http://www.cs.umd.edu/ hcil/VASTcontest07

GeoTime, nSpace TRIST and nSpace Sandbox are
    trademarks of Oculus Info Inc.