Editor: Mike Potel
www.wildcrest.com

Evaluating Visual Analytics at the 2007 VAST Symposium Contest

Catherine Plaisant
University of Maryland

Georges Grinstein
University of Massachusetts Lowell

Jean Scholtz and Mark Whiting
Pacific Northwest National Laboratory

Theresa O'Connell and Sharon Laskowski
National Institute of Standards and Technology

Lynn Chien, Annie Tat, and William Wright
Oculus Info

Carsten Görg, Zhicheng Liu, Neel Parekh, Kanupriya Singhal, and John Stasko
Georgia Institute of Technology

The second Visual Analytics Science and Technology (VAST) contest ran from March through July 2007, in conjunction with the VAST 2007 Symposium. 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, such as the Text Retrieval Conference (TREC), the Knowledge Discovery and Data Mining Cup (KDD), the Critical Assessment of Microarray Data Analysis (CAMDA), and the IEEE Information Visualization (InfoVis) contests.[1]

In this article, we report on the contest's data set and tasks, the judging criteria, the winning tools, and the overall lessons learned in the competition. Because the contest committee members and the contest winners collaborated on this article, we would like to note that for most of the article, the pronoun "we" refers to the contest committee. The exception is in the "Summary of the winning entries" section, where the use of "we" refers to the contest winners.

Data set and tasks

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

- 1,500 news stories simulating an online news archive,
- two blog excerpts with entries extracted over time segments,
- 10 pictures (in JPEG format),
- three small databases (in XLS and CVS format), and
- reference information (in Word and PDF formats).

The synthetic data set included data from analyst-generated scenarios, providing ground truth. The 2006 and 2007 data sets consisted mostly of text, but in 2007 we included new data types (such as blogs, hand-drawn images, and images with annotations), several major subplots or scenarios instead of one, and information gaps that 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 data set was provided under a closed-world assumption: Contestants did not need to go outside the data set 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 data set, identifying the individuals involved in suspicious activities, the time frame for these activities, and provide a list of important events.

They 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.

Timing and incentives

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

 Published by the IEEE Computer Society 0272-1716/08/$25.00 © 2008 IEEE

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 article with the contest committee. All accepted entries are at the InfoVis Benchmark Repository with links from the VAST contest page and NIST web page (we mention the specific websites in the Web Links sidebar).

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 symposium proceedings.

Web Links

The following are helpful resources for learning more about benchmarks and overall results from various visual analytics conferences and contests:

- IEEE Visual Analytics Science and Technology 2007 Contest (VAST07)—www.cs.umd.edu/hcil/VASTcontest07
- IEEE Symposium on VAST—http://conferences.computer.org/vast
- IEEE VAST 2008 Challenge—www.cs.umd.edu/hcil/VATchallenge09
- National Institute of Standards and Technology (NIST)—www.vac.nist/gov
- Information Visualization (InfoVis) Benchmarks —www.cs.umd.edu/hcil/InfovisRepository
- Text Retrieval Conference (TREC)—http://trec.nist.gov/
- Critical Assessment of Microarray Data Analysis Conference (Camda)—www.camda.duke.edu/
- ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)—www.kdd.org/

Judging

While most scoring criteria remain subjective, ground truth made it possible to calculate quantitative scores for the accuracy of responses about the embedded threats—specifically, major players and activities, and where and when the activities occurred. We scored each project's accuracy first, then judged the projects using six experts in VA, human–computer interaction (HCI), and visualization as well as seven professional analysts.

Although we scored the accuracy portion by hand, we anticipate automating this process for future contests. The data set included elements (for example, people, events, and locations) that are part of the plots and some that aren't. We defined the variables T_P, T_N, F_P, and F_N for each type of element as follows:

- a true positive, or T_P, is the number of correctly identified plot elements,
- a true negative, or T_N, is the number of elements reported that aren't part of the plot (such as innocent bystanders identified),
- a false positive, or F_P, is the number of elements incorrectly reported as part of the plot, and
- a false negative, or F_N, is the number of unidentified plot elements.

We then calculated the scores using the following formulas:

- the who score = $T_P + T_N - 1/2(F_P + F_N) + 0.25 *$ association, where association means the number of correctly identified person/organization pairs,
- the what/when score = $T_P + T_N - 1/2\ F_N$,
- the where score = $T_P + T_N - 1/2\ F_N - 1/4 *$ not specific, where not specific is correct but too broad (such as the country instead of the city), and
- the debrief score = correct key events − missing events (key events include 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 data set 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 US East and West coasts. They rated each visualization and assessed the utility of all the components. University entries were judged separately from corporate entries. During the meeting, the judges finished their reviews and discussed individual entries using the following criteria:

- accuracy (correctness of answers),
- quality of the debrief and evidence provided,
- overall system utility,
- visualization quality, and
- process description.

Then each judging team summarized their conclusions and suggested winners. We coordinated activities with conference calls. Over the next few days, we made the final award decisions and prepared evaluation summaries.

VAST 2007 contest results

We received seven entries, three from academia. The two winners were Oculus Info (for the corporate category) and Georgia Tech (for the university category). Both entries were very accurate in answering the who, what, when, and where questions (scoring more than 85 percent). Ratings of the systems' overall utility and visualization quality were also 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.

nSpace and GeoTime are deployed systems with ongoing research and development at Oculus.

We also gave a Best Debriefing Award to the University of British Columbia and Simon Fraser University.[2] This team, using off-the-shelf tools and a great deal of manual labor, had the best accuracy. Although their manual process didn't scale, which was an important contest measure, they successfully identified all the major threads and players, and prepared an exemplary debrief.

We selected the two winning teams and a third team from Applied Technology Systems[3] for the interactive session, based on the quality of their entries and on system robustness. The software had to be capable of processing a new data set in less than 30 minutes and to reliably handle two hours of intensive use.

Winning corporate entry: nSpace and GeoTime, Oculus Info

nSpace and GeoTime are deployed systems with ongoing research and development at Oculus with the support of the Intelligence Advanced Research Projects Activity (IARPA).[4] 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 data sets. GeoTime focuses on structured data with geotemporal registration.

nSpace combines Trist (The Rapid Information Scanning Tool) and the Sandbox (a flexible and expressive thinking environment for analysts to visualize their cognitive mechanisms), while integrating advanced computational linguistic functions using a Web services interface and protocol.[5] (We should mention here that GeoTime, nSpace Trist, and nSpace Sandbox are trademarks of Oculus Info.) For the VAST contest, Oculus Info used the HNC/Fair Isaac linguistics system.

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 multilinked dimensions that help provide rapid scanning, result characterization, and correlation. Users can drag and drop information—including full documents, text fragments, and entities gained from Trist—into the Sandbox for evidence marshalling and further analysis. The Sandbox 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. The system easily identifies entity behaviors and relationships, along with their patterns in both space and time. It then charts the entities and events in a single interactive 3D view.[6] The ground plane is the geographic space represented by the X and Y axes; the vertical T-axis represents time. The system is also capable of animating events in real time.

We used an iterative, multithreaded workflow to analyze the VAST data set. We describe the analysis tasks in a sequential form—but in practice, analysts can jump back and forth from one analysis activity to another.

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 data set of 1,600 documents was indexed and loaded into nSpace, then reviewed using Trist. A series of exploratory queries were refined as our understanding grew. The system visualized the changes highlighted in query refinement results.

Scanning data for characterization and correlation. We viewed the information retrieved within Trist's

multidimensional framework. The date-published dimension provided a time range of the information objects, while the system automatically extracted topics 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 (see Figure 1).

Examining geotemporal 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 identified suspicious patterns. GeoTime's link analysis tools allowed the quick review of history and connectivity of suspicious exporters.

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 identifying key content quickly. As important events were discovered and transferred to the Sandbox, we created an evidence-marshalling timeline view there. To understand patterns and behaviors, we then transferred events into GeoTime and plotted in GeoTime's animated 3D view (see Figure 2).

Assembling data to synthesize information and interpret findings. Throughout the analysis, we saved and annotated discovered insights and supporting evidence in the Sandbox. Sandbox tools—such as links, entities, and groups—helped us organize data. We created layouts that corresponded with our mental models, such as various social networks and hierarchies (see Figure 3a). The Sandbox supported the development and assessment of meaningful hypotheses that were captured as assertions. The system marshaled evidence for assertions through evidence gates by dragging and dropping supporting or refuting evidence from the left and right side of an assertion (see Figure 3b). We analyzed conflicting evidence and competing assertions using the ACH tool, which helped

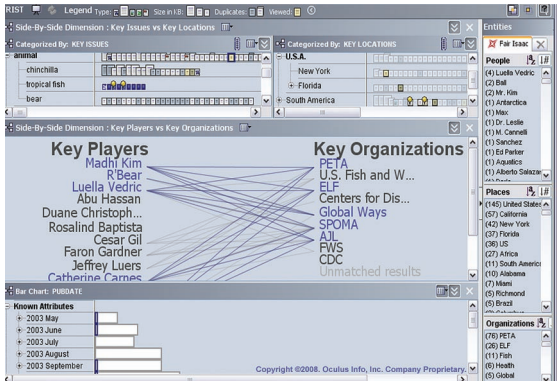

Figure 1. Triaging key topics and relationships in Trist: Clicking on the key issue "tropical fish" highlighted all of the key players and organizations associated with the issue, as well as the relationships between the players and organizations.

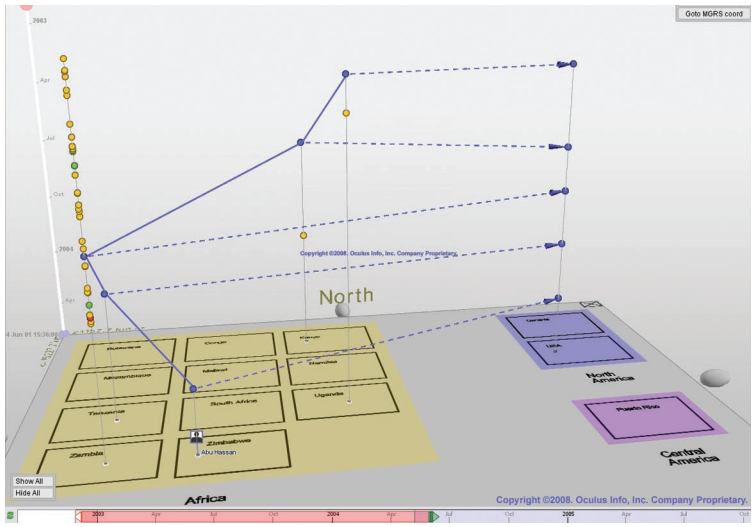

Figure 2. Tracking entity transactions in GeoTime: We mapped countries in the animal import database conceptually to better focus on relevant events.

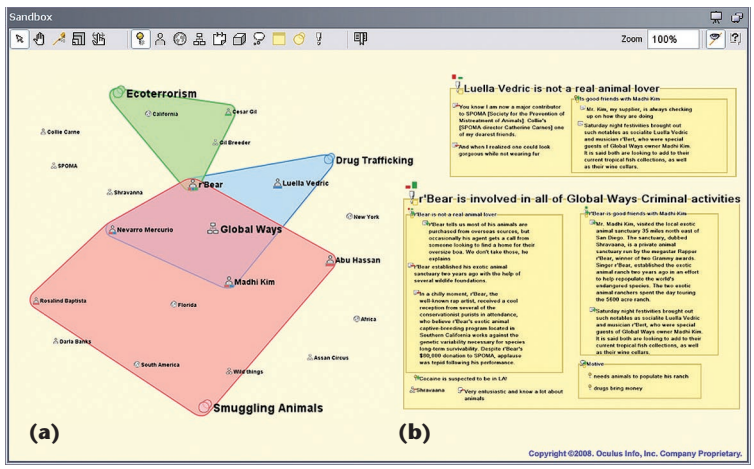

Figure 3. Information synthesis in the Sandbox. (a) A link diagram between key entities and overlapping color-coded sets help identify plot lines. (b) Assertions with weighed evidence help identify villains.

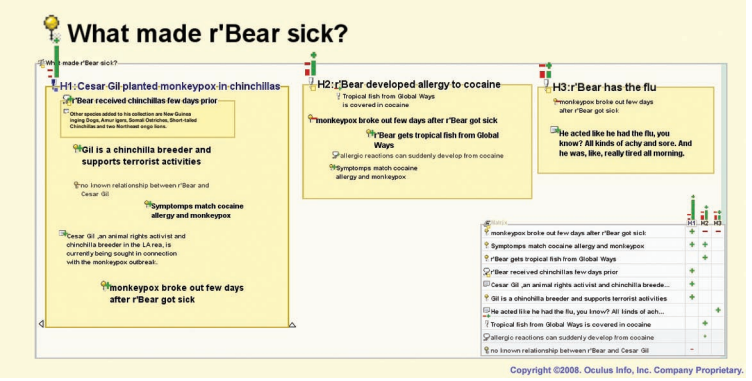

Figure 4. Analysis of competing hypotheses. Diagnosing individual pieces of evidence helps verify which hypothesis is the most probable.

clarify and diagnose the strength of evidence and hypotheses (see Figure 4).

Collaborating to enhance analysis and verify conclusions. Sometimes several analysts worked on the data set simultaneously and needed to share their work. Various elements of content in nSpace were imported from one workspace into another. User-defined dimensions—such as key players, key locations, and templates—were exported, shared, and modified to help uncover additional unexpected patterns. We were able to corroborate answers quickly because nSpace easily visualized and represented our findings. 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.

Reporting. We generated final reports by dragging and dropping objects from Trist and the Sandbox directly into Microsoft Word or PowerPoint, preserving formatting. The system automatically added sources and references for documents and fragments.

Lessons learned. nSpace supports analytical work-in-progress. We can tailor it 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. We were able to observe and investigate expected relationships from various viewpoints, and this helped us discover unexpected patterns without having to wade through multiple spreadsheets, 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 Comma Separated Value (CSV) files, when imported into nSpace, were processed as a single field rather than separate records with multiple fields. This is a feature that the developers hope to implement 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.

Overall, 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.

Winning academic entry: Jigsaw, Georgia Tech

Jigsaw[7] 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 data set at once but use an incremental query-based approach to show a subset of the data set. 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 data sets 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 a step-by-step network exploration by expanding and collapsing nodes to show or hide their connected entities or documents.

- The Scatter Plot View highlights pairwise 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 documents' publication date.

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 shouldn't be changed.

Figure 5 shows four views on the contest data set. 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 from May to October 2003 and shows the entities of one selected document on the side. Jigsaw's graph view is illustrated in Figure 6.

Jigsaw doesn't include 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 more than 350 reports in our own unique subset to become familiar with the 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 preidentified entities from the contest data set 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 wasn't wasted. It helped us become more familiar with many different activities occurring in the reports. We were curious, however, why some connections didn't 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 weren't 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

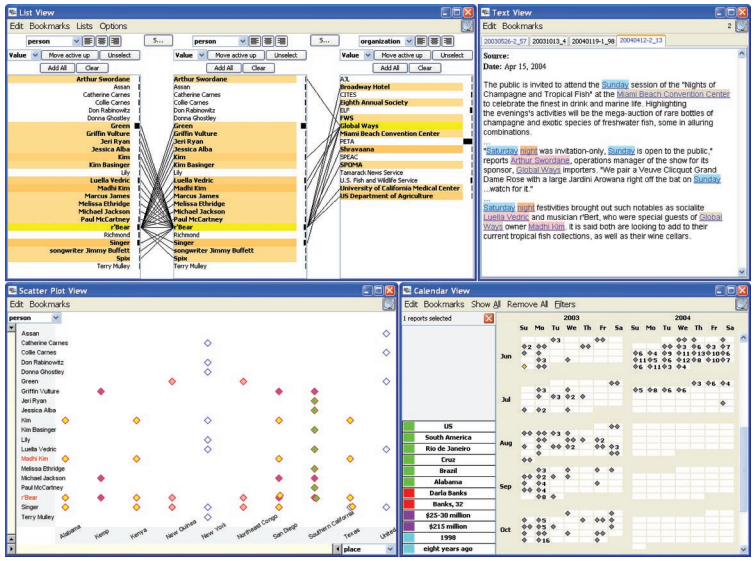

Figure 5. The list view, text view, scatterplot view, and calendar view show data from the contest.

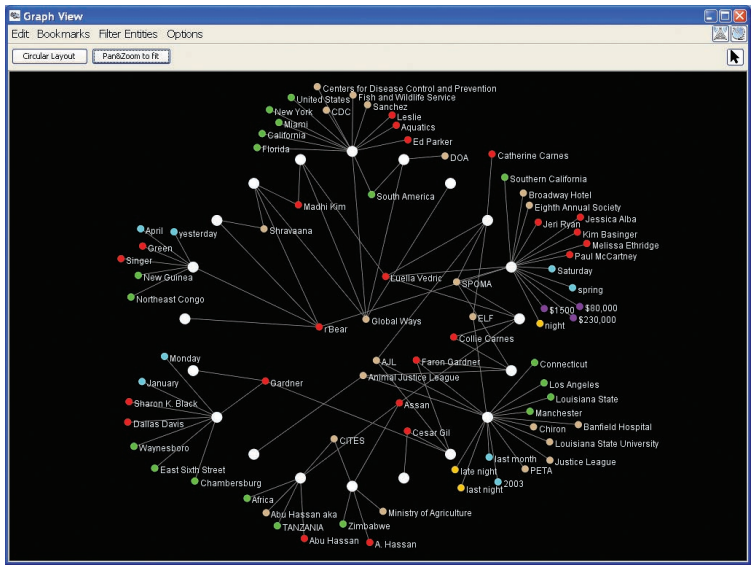

Figure 6. Circular layout in the graph view. All entities connecting to more than one document are drawn in the middle, making it easier to focus on them.

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.

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.

We learned more about the system while working on the contest than we did at any other time. This motivated us to consider future avenues for growth and expansion.

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 (for example, multiple open text views each showed 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 (see 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 might 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 weren't of interest and then we continued to expand the context.

Our investigation disclosed a drawback: Jigsaw lacks the functionality to change extracted entities on the fly. Because the system uses the co-occurrence of entities to build the connection network, it's crucial to properly identify entities. Missing or unidentified entities result in a knowledge gap, because connections that aren't 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 the current limitations of Jigsaw's utility 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 let us drop in interesting entities, register comments about each, and helped us develop hypotheses.

Overall, our participation in the contest was extremely beneficial in terms of improving and updating the Jigsaw system. Having such a large, high-fidelity data set 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.

Interactive session

The Oculus Info, Georgia Tech, and ATS teams participated in the interactive session using a smaller synthetic data set. 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 led 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 data set and scenario creators) moved between the 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. Because the developers had more experience using the systems and working on the type of tasks presented in our data sets, 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 user interface (UI) so that everyone could focus on the task instead of focusing on learning the interface.

Still, we observed that each team quickly became a unified set of analysts as its members engaged in the task. Sometimes, different team members performed different tasks in parallel to speed up the process. Overall, analysts' observations helped us plan refinements to the evaluation metrics.

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 solutions for practice before the 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 data set allowed participation by two teams that didn't have access to text analysis capabilities, so we'll continue to refine the data preprocessing in the future. We had intended to score entries separately depending on whether the entries used the preprocessed data, but with only seven submissions, this was unnecessary.

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 tool's use in the context of the analytical task. This is a common problem we encountered in other contests.[1] 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 aren't 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 weren't used consistently, the teams received only comments, but in the future we intend to refine our rating criteria and procedures to be able to report more objective, consistently derived quantitative scores.

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.

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 anonymous survey 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 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 the 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 percent of them; a third didn't count documents read. Teams reported time-consuming problems—including tying together nontext, unstructured data; hitting dead ends; and resolving data inconsistencies. While we don't know if the effort put forth by these teams is typical across all entries, we're satisfied that their efforts in solving the problem are 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 in understanding a tool's potential than seeing any number of demonstrations, even though the exercise only took a couple of hours. This method might be useful for workplace training. Analysts' experiences in 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.

Panel discussion and 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 US 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. We then answered questions about the specificity of the data set 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 whether 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 offered to act as matchmakers if participants ask for help. Some comments from the audience made it clear that VA researchers are working on tools that address different tasks than the type proposed in the contest (that being a mostly "who did it" scenario) or with data types not included in the data set (such as video). While it's 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 data set and task.

The path forward

We have been encouraged to see that about 75 individuals or organizations have downloaded the 2007 data sets so far. Future data sets might be larger with more data types, uncertainty, and deception. We might also increase the scenario's complexity. The challenge is to keep the problem accessible to enough people. For this reason, we'll 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 participants can process, visualize, and analyze independently from the rest of the data. For example, perhaps a social network of name entities could be extracted and made available for analysis. This analysis wouldn't yield a complete task solution, but it would allow teams working on a network VA to participate at some level. Revealing elements of the ground truth plots through this focused analysis will also 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.

In general, this competition includes an inherent risk for participants who worry that their system

might not do 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're doing. This requires again that we reorganize the data set into components that we can evaluate computationally and independently. Another possibility is to move from the current competition mode—where only the 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 (all who participate report many benefits),
- find partners if you cannot complete the task alone or ask for assistance to find partners, and
- use off-the-shelf tools for components you didn't develop but need.

We encourage VA tool users to

- help prepare sanitized and interesting benchmark data sets and tasks,
- offer your help to generate ground truths for benchmark data sets,
- promote evaluation activities, and
- evaluate tools and train analysts while working on realistic analysis exercises.

We believe that by organizing these contests, we're creating 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 develop baselines for others to build or improve 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 necessary in the future. We'll continue to evolve a collection of data sets, scenarios, and evaluation methodologies that reflect the richness of the many VA tasks and applications.

Acknowledgments

The contest was supported in part by the Intelligence Research Advanced Projects Agency (IARPA), the National Science Foundation Collaborative Research IIS-0713198, and the National Visualization and Analytics Research Center (NVAC) located at the Pacific Northwest National Laboratory (PNNL). PNNL is managed for the US Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RL01830. We thank Jereme Haack, Carrie Varley, and Cindy Henderson of the PNNL Threat Steam Generator project team and Ryan Beaven, Chris Deveau, Curran Kelleher, and Christine Mathai—the University of Massachusetts Lowell students who preprocessed the data and performed evaluation computation. We also thank John Grantham at the National Institute of Standards and Technology for technical support in managing the submission FTP website.

Jigsaw research was supported by the National Science Foundation via Award IIS-0414667 and NVAC, a US 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 postdoctoral program of the German Academic Exchange Service (DAAD).

Oculus Info's research was supported in part and monitored by the National Geospatial-Intelligence Agency (NGA) under contract number HM1582-05-C-0022. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision, unless so designated by other official documentation. The authors wish to thank IARPA and all IARPA and NGA staff for their support and encouragement.

References

1. C. Plaisant, J.D. Fekete, and G. Grinstein, "Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository," *Trans. Visualization and Computer Graphics* (to appear).
2. W. Chao et al., "The Bricolage," *Proc. VAST 2007*, IEEE Press, pp. 239-240.
3. L. Schwendiman, J. McLean, and J. Larson, "Data Analysis Using NdCore and Reggae," *Proc. VAST 2007*, IEEE CS Press, pp. 245-246.
4. W. Wright et al., "The Sandbox for Analysis—Concepts and Methods," *Proc. ACM Computer-Human Interaction*, 2006. pp. 801-810.
5. P. Proulx et al., "nSpace and GeoTime: A VAST 2006 Case Study," *IEEE Computer Graphics and Applications*, vol. 27, no. 5, pp. 46-56.
6. T. Kapler and W. Wright, "GeoTime Information Visualization," *Proc. IEEE Symp. Information Visualization*, 2004, pp. 25-32.
7. J. Stasko et al. "Jigsaw: Supporting Investigative Analysis through Interactive Visualization," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 131-138.

Readers may contact Catherine Plaisant at plaisant@cs.umd.edu.

Contact Applications department editor Mike Potel at potel@wildcrest.com.