Information Visualization
http://ivi.sagepub.com/

Reflections on the evolution of the Jigsaw visual analytics system

Carsten Görg, Zhicheng Liu and John Stasko

Information Visualization 2014 13: 336 originally published online 23 July 2013
DOI: 10.1177/1473871613495674

The online version of this article can be found at:
http://ivi.sagepub.com/content/13/4/336

Published by:
SAGE
http://www.sagepublications.com

Info Vis

Article

Information Visualization
2014, Vol. 13(4) 336–345

DOI: 10.1177/1473871613495674
ivi.sagepub.com
SAGE

Reflections on the evolution of the Jigsaw visual analytics system

Carsten Görg[1], Zhicheng Liu[2] and John Stasko[3]

Abstract

Analyzing and understanding collections of textual documents is an important task for professional analysts and a common everyday scenario for nonprofessionals. We have developed the Jigsaw visual analytics system to support these types of sensemaking activities. Jigsaw's development benefited significantly from the existence of the VAST Contest/Challenge that provided (1) diverse document collections to use as examples, (2) controlled exercises with a set of analytic tasks and solutions for judging results, and (3) visibility and publicity to help communicate our ideas to others. This article describes our participation in a series of VAST Contest/Challenge efforts and how this participation helped influence Jigsaw's design and development. We describe how the system's capabilities have evolved over time, and we identify the particular lessons that we learned by participating in the challenges.

Keywords

VAST Contest, visual analytics, investigative analysis, intelligence analysis, information visualization, sensemaking, multiple views

Introduction

Suppose that you are given a big box full of the pieces from many different jigsaw puzzles and you are asked to put the pieces together from one or two of the most "interesting" puzzles and describe what you see. Oh, by the way, not all the pieces of those "interesting" puzzles are in the box. Investigative analysts, particularly those in fields such as law enforcement or intelligence, frequently confront this kind of challenge in their work. They are given large collections of seemingly unconnected documents and are tasked with identifying a plot or threat that is hinted at, but not clearly communicated, by a small subset of the documents in the collection.

We have developed a visual analytics system called Jigsaw[1,2] to help investigators faced with such challenges. Jigsaw provides a suite of visualizations that depict different perspectives on the documents and the entities (people, places, organizations, etc.) within these documents. Each visualization (called a "view" in Jigsaw) communicates a different aspect of the documents and how the different entities relate to each other. Jigsaw allows an analyst to search for a particular entity and then the system visually communicates the context of that entity, such as the documents in which it appears and the other entities to which it is connected. Alternately, Jigsaw provides different overviews of the document collection so that an analyst can gain some initial evidence about where to begin exploring in more depth. Jigsaw does not automatically find suspicious threads throughout the collection or tell an analyst what to examine first. Instead, it acts

[1]Computational Bioscience Program, School of Medicine, University of Colorado, Aurora, CO, USA
[2]Department of Computer Science, Stanford University, Stanford, CA, USA
[3]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

Corresponding author:
Carsten Görg, Mail Stop 8303, 12801 E 17th Ave, Aurora, CO 80045, USA.
Email: Carsten.Goerg@ucdenver.edu

more as a visual index, helping to show which documents are connected to each other and which are relevant to a particular line of investigation being pursued.

History and contest/challenge participation

In 2004 and 2005, John Stasko participated in a series of meetings that helped to develop an initial definition and research agenda for the field of visual analytics, documented in the book *Illuminating the path*.[3] At those meetings, Professor Frank Hughes of the Joint Military Intelligence College conducted two analysis exercises to provide attendees with a good example of the type of work investigators often conduct. Each exercise involved a set of short, synthetic intelligence reports. These reports were one to a few paragraphs in length and described events that would be of interest to law enforcement and intelligence officials. The events included specific details about particular people, places, organizations, and dates. Attendees were tasked with assimilating these reports, making sense of their contents, and most crucially "connecting the dots" to synthesize a larger crime or threat in the planning stages. No one document itself was enough to understand the plot. A variety of information had to be integrated from across the documents to construct a more complete narrative. These exercises were done using pencil and paper, which are close to representing the state of tools used by many analysts in the field at that point.

Working on these exercises was very challenging for the workshop attendees, even when the exercises consisted of relatively few documents. In particular, it was difficult to keep all the relevant entities clear in one's mind, to remember the context at which they were discussed, and to connect them to other activities that were noted. It occurred to us that visual analytics might provide capabilities to assist with such investigative, sensemaking, and knowledge synthesis tasks. This realization was the genesis of Jigsaw, and it motivated us to create a system that would help investigators and analysts who are confronted with large document collections and need to rapidly understand their contents.

We began the design of Jigsaw in 2006 and shortly thereafter built the initial visualizations within the system. The VAST Contest started in 2006, but our system was not mature enough to be used at that time. In early 2007, the second annual VAST Contest was released. It consisted of approximately 1500 news reports (short text documents) each of a few paragraphs. We decided that this collection would be a good test bed for our system.

We started working on the problem by dividing the news report collection into four piles (for the four people on our team doing the investigation). Each of us skimmed the 350+ reports in our own unique subset just to become familiar with general themes discussed in those documents. We also jotted down notes about people, organizations, or events to potentially study further.

Next, we came together to examine the entire news report collection. Using Jigsaw, we explored a number of the potential leads that each person identified in the initial skim of the reports. At first, we looked for connections across entities, essentially the same people, organizations, or incidents being discussed in multiple reports. After about 6 hours of exploration, we really had no definite leads and were left with many possibilities. Therefore, we returned to the text reports, and some team members read subsets of the reports they had not examined before. At that point, we identified some potential interesting activities and themes to examine further. What also became clear was 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. Closer, more deliberate examinations and readings of the documents uncovered more promising leads and we found additional connections across some actors and organizations in the dataset. Ultimately, we discovered a sinister plot in which animals smuggled into the country were infected with a serious disease that could be transmitted to humans. The contest judges viewed our entry as being extremely accurate about the potential threat, and we were declared the top entry within the academic division of the contest that year. Details of our Contest entry and the analytical process are discussed in Görg et al.[4,5]

Some years later, we entered the VAST Challenge (as it was now known) in 2010. In particular, Mini Challenge 1 in 2010 provided a document collection and an objective much like in 2007—identify a latent threat across the collection and describe the particular details involved in that threat. Unlike the larger collection in the VAST Contest of 2007 where one had to find the "needle in the haystack," here, many different documents contributed to a complex, multifaceted storyline. The plot involved arms dealers from different countries who all convened at a particular location.

Since the number of documents in the dataset was relatively small (just over a hundred compared to more than a thousand in the 2007 Contest), we were able to quickly familiarize ourselves with most of the documents using the views in Jigsaw. We soon realized that unlike in the 2007 Contest data where only a small subset of the documents was relevant to the final solution, most documents in this Mini Challenge seemed to contribute to a larger story.

We used some of Jigsaw's new functionality—computational text analyses and evidence

marshaling—to scaffold our investigation. We started our exploration by examining the high-frequency entities and their connections in the List View and the Graph View. This enabled us to directly focus our attention on important people and places in the dataset. Showing document clusters grouped by topics in the Document Cluster View helped us to keep track of the different threads of the stories embedded in the dataset; in addition, this view indicated which documents we had already read and explored. We created multiple pages in the Tablet (our new approach for note-taking). The pages organized our findings and thinking processes in terms of different perspectives and themes, including social networks, timelines, specific topics such as weapon and fund transfers, and geographically connected people and events. We iteratively modified and refined our hypotheses and findings represented in the Tablet as we read the documents in greater depth and discovered connections between interesting entities. In the end, we uncovered a social network illustrating associations among key players in the arms dealing, patterns of people meeting arms dealer Nicolai Kuryakin in Dubai in the period between April 17 and April 20 in 2009, and patterns of bank fund transfers. Our Challenge entry won an award for "Good Support for Data Ingest." Details of our entry and the analytical process we used for the investigation are discussed in Liu et al.[6]

The 2011 VAST Challenge and its Mini Challenge 3 provided a much larger document collection to explore. It contained 4744 text documents, each in the form of a news report. Jigsaw's organizational and filtering capabilities helped narrow the collection and made it possible for us to browse many of the documents rapidly. As we explored and read more documents, we began to notice that the majority of the documents in the collection were modified versions of actual news articles from the 1990s with key entity names changed. Ultimately, we believed that these documents were not related to the embedded challenge plot. Other interesting and potentially relevant documents, however, were typically shorter and seemed to center around recent activities at a fictitious city called Vastopolis. We uncovered organizations that were planning to make a bioterrorist attack on the city. For this Mini Challenge, Jigsaw was most useful for rapid triage on the documents, helping to determine their potential relevance to the plot. It provided multiple analytical perspectives on the documents' text. Our Challenge entry won an award for "Good Use of the Analytic Process." Details of our entry and the analytical process are discussed in Braunstein et al.[7]

In the following sections, we describe how our participation in the VAST Contest/Challenge influenced the evolution of the Jigsaw system (and our design decisions in particular), and we describe the lessons we learned. We have covered related work of visualization and visual analytics approaches for textual data in two previous journal articles[1,2] and therefore do not provide an explicit section on related work in this article. We do discuss work of other researchers who used Jigsaw to work on their own data in section "Adoption and dissemination."

Lessons learned and Jigsaw evolution

When we started to work on the VAST '07 Contest, we had just finished the implementation of the first prototype of the Jigsaw system. Grounded in our expertise as visualization researchers, the system heavily relied on the interactive visual representation of connections between entities identified across textual documents, and it did not provide any kind of automated text analyses, such as document clustering or summarization. It neither supported the automated identification of named entities, such as people, places, and organizations, in the documents, so we solely relied on the provided dataset, which included identified entities. Details about the state of the system at that point are described in a previous article.[8] In this section, we discuss how our experience from participating in the VAST Contest/Challenges influenced our design decisions and the development of the Jigsaw system.

Reading the documents still matters

One important lesson we learned from working on the contest datasets is that the interactive visualization of connections between entities and documents alone cannot replace the reading of reports. Repeatedly and carefully reading reports is crucial to incrementally expand knowledge about the dataset and to understand details in the underlying plot. The initial version of Jigsaw was helpful in this respect by identifying a small subset of reports that are relevant to an idea being explored and that can be examined closely. However, besides a basic Text View, no other view was tailored towards the visualization of textual data.

To address this shortcoming, we integrated automated text analysis, such as text summarization and clustering, into later versions of the system and made them available throughout the views. These analyses can facilitate the reading of documents. We also improved the Text View (renamed as the Document View) itself since it is such an important component of the system. The views in Figure 1(a) and (b) show the four documents mentioning *Luella Vedric*. The initial Text View (Figure 1(a)) only displayed a document with highlighted entities. The documents in the currently loaded document set were represented as tabs,

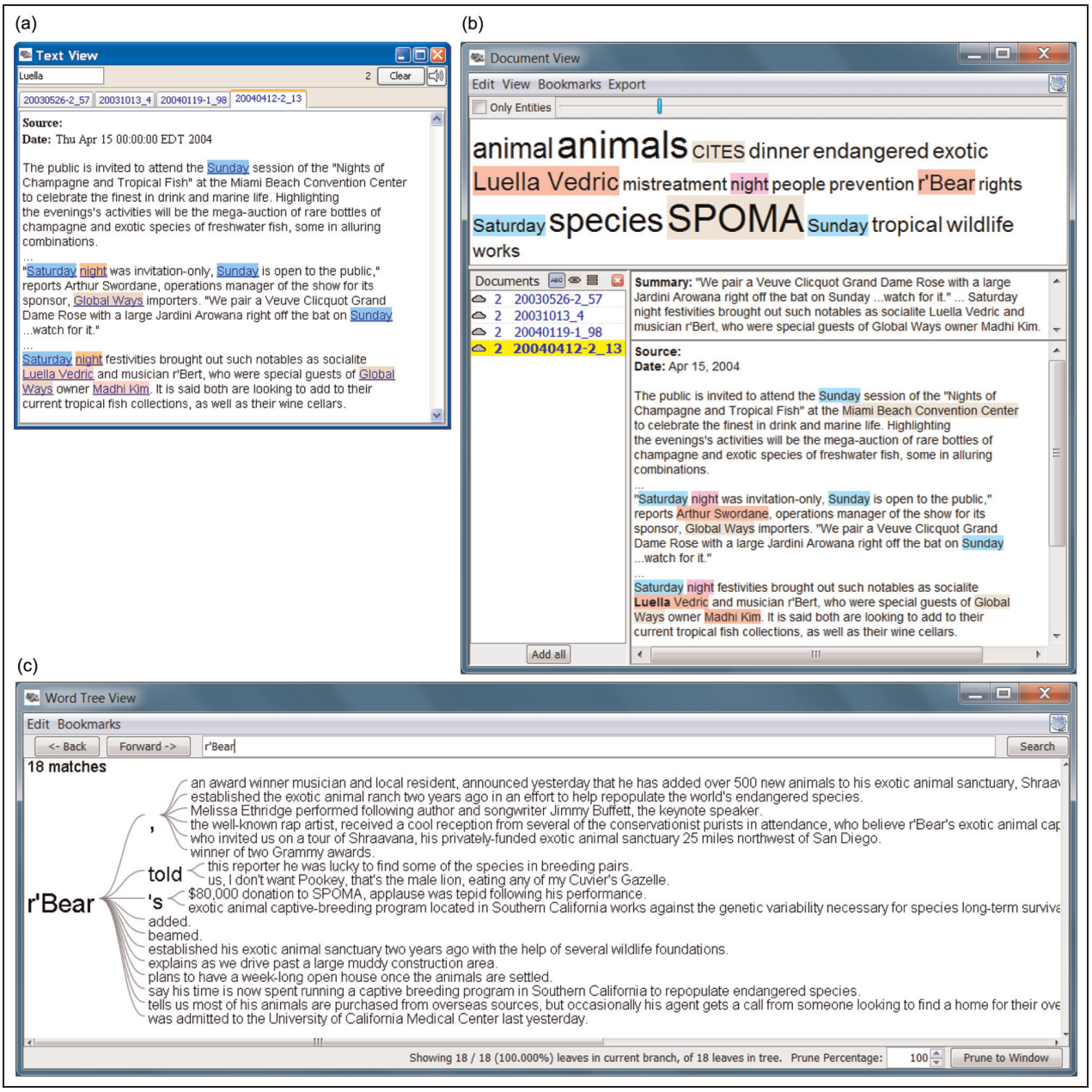

Figure 1. (a and b) The evolution of the Document View. (a) The initial Text View and (b) the current Document View with tag cloud and one-sentence summary of the displayed document. Both views show the same set of documents mentioning *Luella Vedric*. (c) The Word Tree View for *r'Bear*, summarizing the 18 sentences that mention him across 1500 documents.

limiting the number of documents in the view to a few dozen. The current version of the Document View (Figure 1(b)) provides more functionality. It stores the set of loaded documents in a scrollable list (left panel) and thus can handle thousands of documents. It displays a tag cloud (top) for all the documents in the current set to summarize their content.

In this example, we see that the person *r'Bear* in the Blue Iguanodon dataset (2009) and the organization *SPOMA* are key entities in the documents mentioning *Luella Vedric*. The selected document in the set of documents (yellow background) is displayed with highlighted entities (right). Affiliated entities that are connected to a document but are not mentioned in it (e.g. document metadata) are displayed below the document text (not shown in this view because of space constraints); a one-sentence summary of the document is displayed above the document to facilitate the quick scanning of documents. We modified the Document View to count the number of times a

document had been viewed and to allow each view to be named. We frequently found our investigations to have many Document Views present, each with a small number of reports, and naming the view allowed us to recall what the focus of the view was.

To support reading across documents, we implemented Wattenberg's and Viégas' Word Tree approach.[9] The Word Tree View shows all occurrences of a word or phrase across all documents in the context of the words that follow it, each of which can be explored further by a click. The Word Tree View in Figure 1(c) shows occurrences of the person *r'Bear* and the most common phrases that follow that word in sentences within the documents of the Blue Iguanodon dataset (2009). The view illustrates that besides being a musician, *r'Bear* is also involved with the *SPOMA* organization and funds an exotic animal sanctuary, important details in the plot.

Flexible data import is challenging, but vital

Importing and processing data from various sources and in different formats is a crucial feature of any visual analytics system that evolves beyond a lab prototype. Even though we focused our research efforts on the visualization aspects and the integration of computational analyses, we also spent a considerable amount of time and effort on features to ingest and process data. For our participation in the VAST '10 Challenge, we integrated a number of packages to automatically identify entities in text documents, including GATE,[10] LingPipe (http://alias-i.com/lingpipe), the OpenCalais web service (http://www.opencalais.com), and the Illinois Named Entity Tagger.[11] Additionally, we implemented a rule-based approach where we define regular expressions that match dates, phone numbers, zip codes, as well as email, web, and Internet Protocol (IP) addresses. Finally, we added a dictionary-based approach that allows analysts to provide dictionaries for domain-specific entity types that are identified in the documents using basic string matching. All these approaches can be applied to plain text documents, PDF documents, Word documents, and HTML documents. We also implemented a reader for CSV and Excel files that can extract a column with textual data and link it with attributes in other columns.

The improved data import functionality turned out to be very useful beyond our Challenge submission. It is a powerful feature for Jigsaw users outside of academia who have to work with all kinds of text files on a daily basis. Jigsaw users in the law enforcement domain found the reader for Excel files especially useful, and journalists working with Jigsaw often imported their documents from PDF files. These external users discovered a number of bugs and issues in our import functionality and helped us further improve Jigsaw. However, it also became clear to us that we were not able to address all types of exceptions that exist in various file formats and that the data import functionality of a research prototype—which is an important part of a system but not a research contribution—will never be as complete as the data import functionality of a commercial product.

Finally, we defined an XML-based data file format (.jig file) that describes the attributes of documents and the entities within them. In addition to importing files in that format, Jigsaw can also generate these files for documents that were originally imported from other file formats, for example, Word files. This supports the easy sharing of datafiles among Jigsaw users. We have made a number of datasets available in the Jigsaw file format (http://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html).

Entity identification is imperfect and needs help

Although algorithms and libraries for entity identification have improved significantly, they are still far from perfect. In this context, we found during our investigation of the VAST Contest dataset that the missing functionality of being able to change the identified entities on the fly was a significant drawback. Since Jigsaw uses the co-occurrence of entities to build a connection network, it is crucial that the entities are properly identified. Missing or unidentified entities result in a knowledge gap: connections that are not there cannot be visualized. Thus, we added a feature to address this issue. Through direct manipulation in its Document View, Jigsaw now supports manually adding entities that were missed by the entity identification process, changing the type of, or altogether deleting wrongly identified entities. Additionally, we addressed the aliasing or duplication problem: the same logical entity may be identified by different strings in different documents. Jigsaw now provides an operation that allows analysts to merge different entities (strings) under one alias. After assigning a primary identifier to the merged entities, that identifier represents all the initially different entities in Jigsaw's visualizations. Jigsaw uses italics to indicate entities with aliases. The alphabetic sort function in the List View can be helpful to find similar entities.

Assist analysts to start an investigation

At the beginning of an investigation, the amount of data to consider is often overwhelming, and it is difficult for analysts to find a starting point. This is especially true for open-ended, strategic analysis scenarios

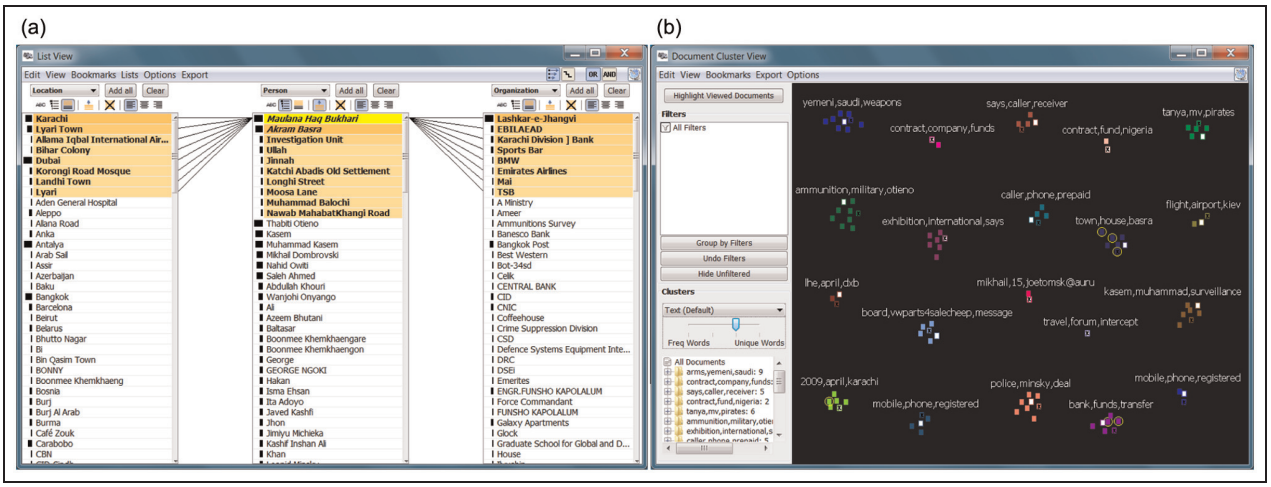

Figure 2. (a) List View showing locations, persons, and organizations connected to *Maulana Haq Bukhari* and (b) Document Cluster View showing different clusters of related documents (small rectangles in different colors). Documents mentioning *Bukhari* are selected (surrounded by a yellow circle).

in which analysts are not tasked with examining specific people, places, and organizations. Instead, the analyst must initially learn about the different topics and contents within the data and decide what to investigate first.

Jigsaw initially did not have capabilities for finding themes or concepts in a document collection, and it was challenging for analysts to get started with an investigation. When we began working on the Contest dataset, we split the documents among ourselves and read all of them to find initial leads. We noted the need for a more global view of all the reports, one that could show which documents have been examined and that would allow the documents to be partitioned into groups.

To better assist analysts in browsing and understanding text documents in a more structured manner, we coupled the interactive visualizations in Jigsaw with automated computational analysis capabilities such as analyses of document summarization, document similarity, and document clustering. We integrated three different types of document summaries. We summarize single documents with a one-sentence summary by extracting the most important sentence from the document. We show the one-sentence summary in the Document View and also provide it as a tooltip wherever a document is shown by its icon. To summarize a collection of documents, we use word clouds (if there is enough space available, for example, in the Document View) or keyword summaries (if there is not a lot of space available, for example, in the Document Cluster View). The document summaries help analysts to quickly decide whether to read (a set of) documents in detail. Document similarity can be based on either the document text or the entities identified in or associated with the documents. Similarity helps analyst to understand whether a particular document is an outlier in a collection or part of a bigger theme (if there are similar documents). Document clustering can also be based on either the document text or the entities of the documents. It provides an overview of the document collection and helps analysts to explore the documents more systematically. (Additional details of the computational analyses are described in Görg et al.[2]) We integrated the computational analyses across a number of views in the system, as described below.

The Document Cluster View visualizes document clustering results and indicates which documents already have been read. One of Jigsaw's key capabilities, cross-filtering across views, becomes now even more powerful since it can highlight connections across entity-centric visualization, such as the List View, and document-centric visualization, such as the new Document Cluster View. Figure 2 shows an example in the context of the Challenge 2010 dataset. The List View (Figure 2(a)) shows locations, persons, and organizations connected to *Maulana Haq Bukhari*. *Karachi* and *Lyari Town* are the most connected locations, *Akram Basra* is the most connected person, and *Lashkar-e-Jhangvi* is the most connected organization. The Document Cluster View (Figure 2(b)) shows different clusters of related documents (small rectangles in different colors). Documents mentioning *Bukhari* are selected (surrounded by a yellow circle). The view illustrates that *Bukhari* is strongly connected to documents in the "town, house, basra" and the "bank, funds, transfer" clusters.

Additionally, we implemented a view that is tailored toward the representation of text analysis results. The Document Grid View can present, analyze, and compare a variety of document metrics, such as document similarity or sentiment. The view organizes the

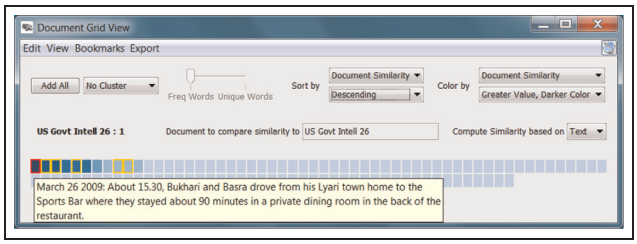

Figure 3. Document Grid View with the document (small rectangle) order and shading set to correspond to the documents' similarity to the selected report on *Bukhari*.

documents in a grid and provides an overview of all the documents' similarity to a selected document via the order and color of the documents in the grid representation. The Document Grid View in Figure 3 shows all documents of the Challenge 2010 dataset grouped and colored by similarity to a report on *Bukhari*. The tooltip displays the one-sentence summary of that report. The highlighted documents (yellow rectangle) also mention *Bukhari*. The computational analyses, in particular the document clustering, proved to be very useful in guiding the process of reading and making sense of the documents.

Do not neglect evidence marshaling

Taking notes of hypotheses and findings, tying them back to evidence, and keeping track of an investigation are important parts of analysis, especially for investigations that are carried out over a longer period of time. Jigsaw's support for these activities also evolved over time. When we participated in the VAST '07 Contest, the system did not provide any support for note-taking and we only relied on our manual notes on paper. After realizing this shortcoming, we developed the ShoeBox window to support evidence marshaling and note taking. We chose a structured approach for the view so that analysts could create hypotheses and then connect supporting as well as contradicting evidence to the hypotheses that were linked back to documents as provenance. The view provided a number of advanced features, such as "hypothesis-slides" that could be laid over one another similar to different graphic layers in image editing software such as Photoshop. This feature allowed analysts to compare hypotheses and ask "what if" questions, investigating different scenarios. However, the ShoeBox window was just too complex and did not allow analysts to take free-style notes in the way they would do on paper. Therefore, it was seldom used, and we abandoned it after some iterations.

As a replacement, we developed the Tablet as our new evidence marshaling tool. The Tablet adopts a minimalistic design, intending to offer greatest flexibility for visual thinking and sensemaking. Entities in Jigsaw's views can be directly added to the Tablet via popup menu commands. The added entities retain their original color coding according to their types. Analysts can also create their own items representing customized entities or events. Any two items or entities can be linked and the links can be labeled. Additional information about the items can be represented as post-it-notes (on a yellow background). Analysts can also create timelines and link entities or items to specific points on the timeline. All the visual items in the Tablet can be freely moved around and repositioned. With the Tablet, we took a free-style approach, mimicking analysts' note-taking behavior on paper. The basic constructs allow analysts to organize significant events into timelines, log and connect related people and organizations, and gradually build up hypotheses. The Tablet also allows the user to integrate bookmarks of views, as provenance or evidence, and the views can be re-instantiated at a later point to follow up on the original analysis. We used the Tablet in the VAST '10 Challenge submission and had much better results than we had with the ShoeBox. Figure 4 shows the initial ShoeBox (Figure 4(a)) and the new Tablet (Figure 4(b)).

Observations from our contest participation

Feedback from contest participation

Our participation in the VAST Contest/Challenges was extremely beneficial for us. It helped us to both improve and evolve the Jigsaw system, and it provided a hands-on learning experience through which we gained a better understanding of the analytical process in these types of investigations. This better understanding of the analytical process then directly guided some of our design decisions.

The availability of such large, high-fidelity datasets was invaluable in numerous ways, and it particularly allowed us to observe the utility of the different views in an actual investigation scenario. The feedback about our system we garnered from our own experience working on the contests was a useful complement to the feedback we received from other Jigsaw users. Participating in the contests allowed us to put ourselves into the shoes of an investigator and experience firsthand the virtues and shortcomings of our system. This knowledge motivated us to fix usability problems, create new operations and views, and consider future avenues for growth and expansion.

In particular, the live Contest session with professional analysts to which the winners of the VAST '07 Contest were invited was highly useful.[12] Working together with an analyst, seeing the analyst's perspective on an investigation, the kind of questions he asks,

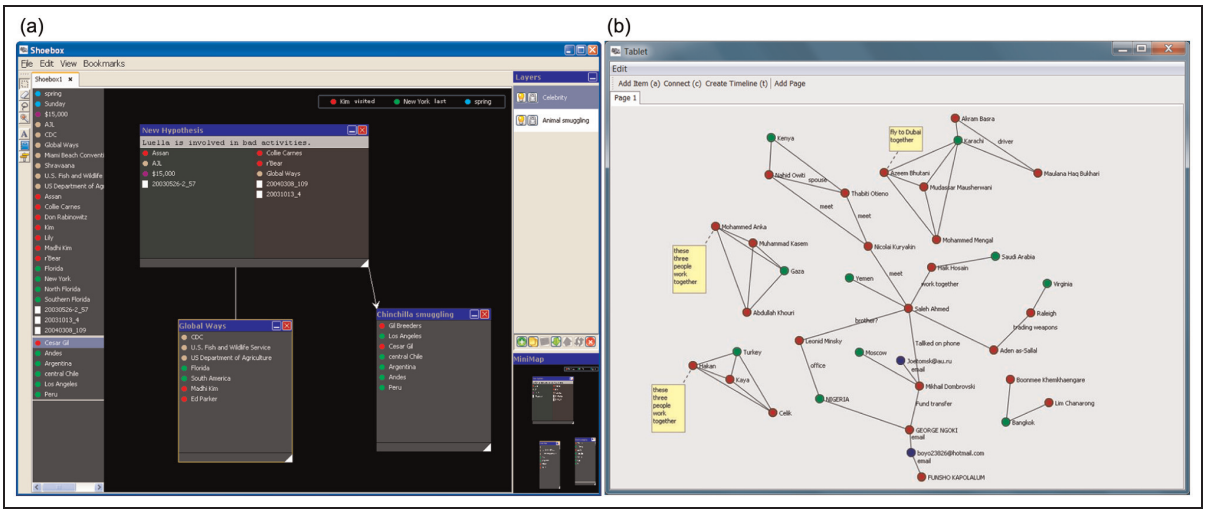

Figure 4. (a) The initial ShoeBox and (b) the new Tablet.

the kind of evidence he looks for, the way he builds trust in the data, and the way he interprets visual representations were very illuminating. It became clear which parts of the system aligned with the workflow of a professional analyst and which parts needed to be changed and improved.

The written feedback we received from our contest submissions was not as useful, however. Naturally, feedback that is based on a summary report and a video that only shows the "shiny" side of a system but not its shortcomings cannot have the same quality as feedback received from an interactive session with an analyst. Without observing the actual analytical process and the tool usage during the investigation, it is difficult for reviewers to provide accurate feedback that can drive the further development of a system.

Using contest datasets for evaluation

Our analysis activities in the VAST Contest/Challenges using the large examples with embedded ground truth exposed a number of shortcomings in the Jigsaw system, and thus, the activities functioned very much in a formative evaluation sense. The contest datasets have been used in class projects[13] and lend themselves for evaluation studies since answers can be easily graded because of the embedded ground truth. However, the contest datasets are too large and too complex to be used in lab experiments for formal evaluations of visual analytics systems. Such studies must be completed in a few hours, and the scope of the contest datasets does not fit this purpose. To evaluate the Jigsaw system in a formal experiment, we developed our own small dataset with embedded ground truth. The contest dataset and scoring scheme served as a model. Details of our system evaluation are described in Kang et al.[14] We do appreciate the effort it takes to prepare datasets with embedded ground truth, and we applaud the VAST Contest organizers' effort over the last 7 years.

Adoption and dissemination

Over the years, Jigsaw became better known as a visual analytics system through our participation in the Contest/Challenges. The publicity it received through our challenge participation and in particular the title of a winning entry of the VAST '07 Contest certainly helped to raise its visibility in the community. Analysts from a variety of areas contacted us and wanted to try the system on their own data.

We believe that two additional aspects of the system helped foster its adoption by other people: (1) an emphasis on usability and (2) simplified data import. We focused on two principles for achieving usability. First, Jigsaw uses only two primary interactive operations: a single click selects an entity or document and a double click expands the context of an entity or document (brings in additional entities or documents that are connected to it). Of course, there are a number of other advanced operations, but they are not required for a novice user to get started and they are not in the way of the simple interactions. Second, we consistently implemented brushing and linking across all the views to encourage the user to use the views together. The 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 of the system stay consistent and provide different perspectives on the same data.

User support is crucial for the adoption of any system, and we employed a two-pronged approach. To

introduce novel users to the system, we created a number of training videos that guide the user step-by-step through the views and the analytical process. We also provide a number of example datasets that could readily be used. Additionally, we offer to help users to import their own data efficiently. The current version of Jigsaw does support the import of a number of file formats, including text, PDF, Word, Excel, and HTML. However, many different data formats exist, and we have found data import to still be a major hurdle.

Through its high visibility, Jigsaw gained popularity within the visual analytics research community. Researchers from various organizations used Jigsaw in conjunction with other tools to work on VAST Challenges. In the 2011 VAST Mini Challenge 3, for example, teams from the University of Konstanz and the City University London used Jigsaw in their visual analytical processes. The Konstanz team first identified candidate documents that were potentially relevant to the plot using keyword-based filtering and supervised machine learning classification. The researchers then used Jigsaw to explore the entities and documents within these candidate selections. Three views were particularly useful: the List View, the Document Cluster View, and the Document View. Similarly, the City University London team first used their own tool to perform keyword-based document filtering and to discover interesting organizations. With this initial set of leads, they used Jigsaw's Graph View to explore connected entities and the Document View to read documents. The Tablet View was useful for documenting how various entities were discovered in the analytical process.

These exemplary usage cases are consistent with our experience of using Jigsaw. Without a predefined theme or a set of keywords, it is usually difficult to identify a subset of documents related to the ultimate solution narrative. To our knowledge, few effective natural language processing tools can accomplish this task competently. The Document Cluster View in Jigsaw, backed by clustering algorithms, was moderately helpful on the challenge datasets. Jigsaw's strengths lie in its highly coordinated views, and both teams that used the system took advantage of this feature.

Outside of the VAST Contests, researchers have and are using Jigsaw for different purposes including targeting analysis and hypothesis generation based on police/intelligence case reports, comparing aviation documents,[15] understanding source code files for software analysis and engineering,[16] and genomics research based on PubMed articles.[17] Other domains or types of documents that have been analyzed with Jigsaw include investigative reporting, fraud, consumer reviews, academic publications, business intelligence, webpages, and blogs. Another article[18] reviews six individuals, including those from law enforcement and intelligence, who have used Jigsaw for periods from 2 to 14 months.

Synthetic datasets

The VAST Contest/Challenges are presented in a significantly different way than other visualization contests, such as the former InfoVis Contest: instead of using real-world datasets with open-ended questions, the organizers of the VAST Contest decided to create synthetic datasets with an embedded ground truth. Using synthetic datasets has a number of advantages: it is easier to ensure that a dataset has the right scope since the organizers can decide how large and complex the dataset should be, and the ground truth allows the organizers to better judge the submitted entries on their accuracy. Additionally, synthetic datasets often motivate the participants since they know that a solution does exist, and they also know that the analysis is feasible for students and researchers, not requiring the knowledge and background of a professional analyst. However, we also noted a few shortcomings of the text-based synthetic datasets used in the VAST Contests. These datasets are a mix of "real" text documents and "contrived" text documents. Real events and people are not likely to be involved in the embedded ground truth, and therefore, we excluded (sometimes consciously, sometimes maybe subconsciously) documents that mentioned real events or real people. The synthetic datasets are very useful for promoting and demonstrating a visual analytics system. Having a dataset that is clearly not a toy example but also not too large, and knowing an interesting story within it, lends itself for good system demonstrations.

Scalability

Scalability is an important aspect of any visual analytics system. Jigsaw evolved over the years from using an in-memory data model, capable of handling a few thousand documents and tens of thousands of entities, to an architecture using a database framework, capable of handling tens of thousands of documents and millions of entities. This evolution was driven by the demand of real-world clients and their applications and not by our participation in the VAST Contests. The synthetic datasets used in the contests were not large enough to motivate the move from an in-memory model to a database model. This decision of the VAST Contest organizers is understandable given the multiple purposes that the contest data were serving. Keeping the datasets at more modest sizes also encouraged more teams to work on the problems and submit entries.

Conclusion

A primary task in the VAST Contest/Challenges is to "connect the dots" or "put the pieces together." This task aligns very closely with the primary purpose of Jigsaw, and we found the existence of the contest datasets very beneficial for the development of the system. In this article, we have described how Jigsaw evolved from a very visualization-centric system to a balanced visual analytics system that provides and integrates both computational text analyses and interactive visualizations of entities and documents. Our experiences from the VAST Contest and Challenges influenced our design decisions and guided us throughout this process. Additionally, we discussed our observations from participating in the Contest and Challenges, including feedback from our participation, lessons on evaluation, adaption, dissemination, and properties of synthetic datasets.

Acknowledgements

Many other students have contributed to Jigsaw's development and our participation in the VAST Challenges, including Meekal Bajaj, Elizabeth Braunstein, Jaegul Choo, Alex Humesky, Jaeyeon Kihm, Vasilios Pantazopoulos, Neel Parekh, Kanupriya Singhal, Gennadiy Stepanov, and Sarah Williams. We also would like to thank the organizers of the VAST Contests and Challenges for their continued effort in creating the datasets and judging the submissions.

Funding

This research is based upon the work supported in part by the National Science Foundation via Awards IIS-0414667, IIS-0915788, and CCF-0808863; by the National Visualization and Analytics Center (NVAC), a US Department of Homeland Security Program; and by the US Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

References

1. Stasko J, Görg C and Liu Z. Jigsaw: supporting investigative analysis through interactive visualization. *Inform Visual* 2008; 7(2): 118–132.
2. Görg C, Liu Z, Kihm J, et al. Combining computational analyses and interactive visualization for document exploration and sensemaking in Jigsaw. *IEEE T Vis Comput Gr*, in press.
3. Thomas JJ and Cook KA. *Illuminating the path*. Washington, DC, USA: IEEE Computer Society, 2005.
4. Görg C, Liu Z, Parekh N, et al. Jigsaw meets Blue Iguanodon—the VAST 2007 contest. In: *IEEE VAST*, Sacramento, CA, October 2007, pp. 235–236. Washington, DC, USA: IEEE Computer Society.
5. Görg C, Liu Z, Parekh N, et al. Visual analytics with Jigsaw. In: *IEEE VAST*, Sacramento, CA, October 2007, pp. 201–202. Washington, DC, USA: IEEE Computer Society.
6. Liu Z, Görg C, Kihm J, et al. Data ingestion and evidence marshalling in Jigsaw. In: *IEEE VAST*, Salt Lake City, UT, October 2010, pp. 271–272. Washington, DC, USA:IEEE Computer Society.
7. Braunstein E, Görg C, Liu Z, et al. Jigsaw to save Vastopolis—VAST 2011 Mini Challenge 3 Award: "Good Use of the Analytic Process." In: *IEEE VAST*, Providence, RI, October 2011, pp. 323–324. Washington, DC, USA: IEEE Computer Society.
8. Stasko J, Görg C, Liu Z, et al. Jigsaw: supporting investigative analysis through interactive visualization. In: *IEEE symposium on visual analytics science and technology 2007 (VAST 2007)*, Sacramento, CA, October 2007, pp. 131–138. Washington, DC, USA: IEEE Computer Society.
9. Wattenberg M and Viégas FB. The word tree, an interactive visual concordance. *IEEE T Vis Comput Gr* 2008; 14(6): 1221–1228.
10. Cunningham H, Maynard D, Bontcheva K, et al. *Text processing with GATE* (Version 6). Gateway Press CA, 2011.
11. Ratinov L and Roth D. Design challenges and misconceptions in named entity recognition. In: *CoNLL*, Boulder, CO, June 2009, pp. 147–155. Stroudsburg, PA, USA: Association for Computational Linguistics
12. Plaisant C, Grinstein G, Scholtz J, et al. Evaluating Visual Analytics: The 2007 Visual Analytics Science and Technology Symposium Contest. *IEEE Computer Graphics & Applications* 2008; 28(2): 12–21.
13. Whiting MA, North C, Endert A, et al. VAST contest dataset use in education. In: *IEEE symposium on visual analytics science and technology 2009 (VAST 2009)*, Atlantic City, NJ, October 2009, pp. 115–122. Washington, DC, USA: IEEE Computer Society.
14. Kang Y-A, Görg C and Stasko J. How can visual analytics assist investigative analysis? Design implications from an evaluation. *IEEE T Vis Comput Gr* 2011; 17(5): 570–583.
15. Pinon OJ, Mavris DN and Garcia E. Harmonizing European and American aviation modernization efforts through visual analytics. *J Aircraft* 2011; 48: 1482–1494.
16. Ruan H, Anslow C, Marshall S, et al. Exploring the inventor's paradox: applying Jigsaw to software visualization. In: *ACM SOFTVIS*, Salt Lake City, UT, October 2010, pp. 83–92. New York, NY, USA: ACM.
17. Görg C, Tipney H, Verspoor K, et al. Visualization and language processing for supporting analysis across the biomedical literature. In: *Knowledge-based and intelligent information and engineering systems* Setchi R, Jordanov I, Howlett RJ and Jain LC (eds). LNCS Berlin Heidelberg: Springer, 2010.
18. Kang Y-A and Stasko J. Examining the use of a visual analytics system for sensemaking tasks: case studies with domain experts. *IEEE T Vis Comput Gr* 2012; 18(12): 2869–2878.