Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw

Carsten Görg, *Member*, *IEEE*, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, *Student Member*, *IEEE*, Haesun Park, *Member*, *IEEE*, and John Stasko, *Senior Member*, *IEEE*

Abstract—Investigators across many disciplines and organizations must sift through large collections of text documents to understand and piece together information. Whether they are fighting crime, curing diseases, deciding what car to buy, or researching a new field, inevitably investigators will encounter text documents. Taking a visual analytics approach, we integrate multiple text analysis algorithms with a suite of interactive visualizations to provide a flexible and powerful environment that allows analysts to explore collections of documents while sensemaking. Our particular focus is on the process of integrating automated analyses with interactive visualizations in a smooth and fluid manner. We illustrate this integration through two example scenarios: An academic researcher examining InfoVis and VAST conference papers and a consumer exploring car reviews while pondering a purchase decision. Finally, we provide lessons learned toward the design and implementation of visual analytics systems for document exploration and understanding.

Index Terms—Visual analytics, information visualization, sensemaking, exploratory search, information seeking, document analysis

1 Introduction

Everyday, analysts and investigators confront large collections of data as they make decisions, solve problems, or simply seek to understand a situation better. Frequently, the data collections include text documents or documents with key text components. While numerical or structured data are more amenable to statistical and computational analysis, text data are conversely often messy and noisy, requiring a very sequential, slow processing (reading documents one-at-a-time, in order).

Investigators working with such document collections gather bits of information as they explore the data, hoping to form new insights about the issues at hand. Large, unstructured document collections make this task more difficult; the investigator may not know where to begin, what is important, or how concepts/events are related. The following situations are examples of these kinds of tasks:

- An academic researcher moves into a new area and seeks to understand the key ideas, topics, and trends of the area, as well as the set of top researchers, their interests, and collaborations.
- A consumer wants to buy a new car but encounters a large variety of possible models to choose from, each of which has 10 to 20 "professional" reviews and a web forum with hundreds of postings.
- A family learns that their child may have a rare disease and scours the web for documents and information about the condition, easily encountering many articles.
- A police investigator has a collection of hundreds of case reports, evidence reports, and interview transcripts and seeks to "put the pieces together" to identify the culprits behind a crime.

Such processes, sometimes called Sensemaking [39], [50], [54], Information Seeking Support [44], or Exploratory Search [43], [66], go beyond the initial retrieval of data or the simple return of the "right" document. Instead, they involve analysts browsing, exploring, investigating, discovering, and learning about the topics, themes, concepts, and entities within the documents, as well as understanding connections and relationships among the entities.

One approach to this problem is the computational analysis of document text, including text mining [3], [22]. However, as many researchers have noted [37], [58], simply performing computational analysis of the documents may not be sufficient for adequate understanding of a document collection—the investigator inevitably will think of some question or perspective about the documents that is either not addressed by the computational analysis or not represented accurately enough to draw a conclusion.

- *C. Görg is with the Computational Bioscience Program, University of Colorado, Mail Stop 8303, 12801 E 17th Ave, Aurora, CO 80045. E-mail: Carsten.Goerg@ucdenver.edu.*
- *Z. Liu is with the Department of Computer Science, Stanford University, 379 Gates Hall, Stanford, CA 94305. E-mail: zcliu@cs.stanford.edu.*
- *J. Kihm is with Cornell CIS, 301 College Ave., Ithaca, NY 14850. E-mail: jk2443@cornell.edu.*
- *J. Choo and H. Park are with the School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332. E-mail: {joyfull, hpark}@cc.gatech.edu.*
- *J. Stasko is with the School of Interactive Computing, Georgia Institute of Technology, 85 5th St., NW, Technology Square Research Building, Atlanta, GA 30332. E-mail: stasko@cc.gatech.edu.*

Manuscript received 8 May 2012; revised 3 Oct. 2012; accepted 5 Dec. 2012; published online 21 Dec. 2012.
Recommended for acceptance by F. van Ham.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2012-05-0082.
Digital Object Identifier no. 10.1109/TVCG.2012.324.

1077-2626/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

Another approach leverages information visualization to show information about document contents [40], [47], [59]. However, interactive visualization itself may not be sufficient for sensemaking either—as the size of the document collection grows, interactively exploring the individual characteristics of each document may simply take too much time.

Our approach to the problem combines these two analytics methods: 1) automated computational analysis of the text documents and 2) interactive visualization of the documents and of the analysis results. Such a combination is described as a *visual analytics* approach [36], [58], and it leverages the strengths of both the human and the computer. Humans excel at the interactive dialog and discourse of exploration and discovery. They develop new questions and hypotheses as more and more information is uncovered. They reason about the importance of new facts that are discovered. The computer excels at complex analyses to calculate metrics, correlations, connections, and statistics about the document collection. It can rapidly analyze large collections of documents in ways that would be prohibitively time consuming for people to do.

Relatively few systems to date have deeply and smoothly incorporated both automated computational analysis and interactive visualization while providing a tight coupling between the two. Systems (as discussed in the related work) usually focus on one of the two approaches and provide a few elements from the other. For instance, computational analysis tools sometimes provide rudimentary visualizations to depict analysis results. Alternatively, interactive visualization systems may provide a few simple analysis techniques such as filtering or statistical analysis of the data.

Elaborating on this notion, Keim et al. [36] state:

> Visual analytics is more than just visualization. It can rather be seen as an integral approach to decision making, combining visualization, human factors, and data analysis. The challenge is to identify the best automated algorithm for the analysis task at hand, identify its limits which cannot be further automated, and then develop a tightly integrated solution, which adequately integrates the best automated analysis algorithms with appropriate visualization and interaction techniques.

In this paper, we explore this coupling through Jigsaw [55], a system for helping analysts explore document collections. Jigsaw is a relatively mature system, and has garnered trial use in the field by analysts in law enforcement, investigative reporting, fraud detection, and academic research, among other areas. An initial user study of the system showed its potential in helping investigators work with documents and in supporting different analysis strategies [34].

Earlier versions of Jigsaw emphasized multiple, coordinated visualizations but provided relatively little computational analysis of documents' text. The system primarily visualized connections between entities across documents to help investigators follow trails of information. More recently, we have added a variety of automated text analyses to the system including analysis of document similarity, document sentiment, document clusters by content, and document summarization through a few words or sentences. These new analyses aid investigators in determining the documents to examine first, the documents to focus on or discard, and the documents that may be related to different investigative angles.

Our focus is not on developing novel innovative algorithms for computational text analysis. Instead, we explore ways to smoothly integrate existing computational analyses into an interactive visual interface in a seamless manner that will provide a natural and fluid user experience. Furthermore, new computational analysis algorithms frequently are developed for well-defined tasks or problems with carefully constructed inputs and data. Real-world visual analytics systems, conversely, encounter messy, noisy data and must support open-ended analytical reasoning and sensemaking. Thus, our research also examines how computational analysis techniques can be used throughout visual exploration on challenging real-world data.

The contributions of this research include: 1) methods for fluidly integrating computational text analysis and visualization, 2) illustration of the utility of such an approach through two example usage scenarios, and 3) lessons learned toward the design and construction of visual analytics systems for document exploration and understanding. Additionally, we provide implementation advice and experience on the integration of text analysis algorithms as a broader benefit for other researchers.

2 Related Work

Computationally aided analysis and visualization of text and documents to assist human investigators with sensemaking has been a topic of intense research interest recently. Furthermore, different subdisciplines of computer science each bring their own focus to the problem. Thus, a comprehensive examination of related work likely would take a complete paper itself. Here, we highlight some of the existing research most strongly related to our work to provide the reader with greater context and familiarity of the varied approaches others have taken.

Systems in this area typically focus on some aspect of a document or document collection to present. Broadly, they visualize

1. metadata about the documents;
2. the document source text (words);
3. computed features and attributes of the documents including entities; and/or
4. general concepts, themes, and models across the documents.

Visualization techniques have been developed for single documents or large collections of documents, though the techniques for individual documents often can be generalized to collections.

Systems with a specific focus on helping people understand various attributes of an academic paper collection are a good example of presenting *metadata about a set of documents*. PaperLens [40] employs a variety of bar chart, list, graph, and text-based visualizations to show author, topic, and citation data of past CHI papers. The system uses a clustering analysis to help group papers by topic as well. Selecting an author, paper, or concept in one visual representation loads related items into the other visualizations. A follow-on system, NetLens [33], focuses on visualizing content-actor data within document collections such as scientific publications and authors. NetLens uses bar charts, histograms, and lists to represent the data and help analysts understand statistics and trends from conference papers and their citations.

A number of innovative visualization techniques have been developed to represent the *words and source text of documents*. The SeeSoft system [21] represents a line of a text document by a row of pixels on the screen, with the length of the text line (number of characters) mapped to the length of the row of pixels. The goal of the technique is to visually depict documents that are larger than what can normally be shown on one screen. Other well-known source text visualization techniques such as TextArc [47], Word Clouds [61], Word Trees [64], and Phrase Nets [59] actually still show text, unlike SeeSoft. They also show frequency and relationships of particular words or terms within documents.

Many systems, in fact, inhabit a conceptual space that transitions from visualizing document source to visualizing *computed metrics or features of a document or documents*. For example, Viégas et al. [60] analyze collections of e-mail messages using a variant of the term-frequency inverse document-frequency (TF-IDF) algorithm that focuses on each sender. The system's visualization is temporally based and shows lists of keywords from the e-mails to characterize the main topics of messages during each month and over entire years.

Other techniques such as Arc Diagrams [63], DocuBurst [12], and Parallel Tag Clouds [13] compute metrics about a set of documents and visualize the computed metrics in unique ways. The PaperVis system [10] combines a relevance-determination algorithm with visualization to show relationships among academic articles. PaperVis performs citation and keyword analysis and presents the results through bulls eye and radial space filling visualizations. The size of a node (document) and its distance to other documents denote its importance and relevance, respectively.

Keim and Oelke [38] perform numerous text analysis measures not seen in other document analysis systems including measures such as average word length, sentence length, number of syllables per word, and other measures such as Simpson's index and Hapax Legomena and Dislegomena (number of words occurring once and twice). The visualization of the analysis results for each of these measures uses a heatmap style display. Together with colleagues they subsequently added sentiment analysis to their measures [45] and added node-link network visualization to communicate relationships among the documents' sentiments [46].

One particular computed attribute sometimes visualized by systems is an entity within a document or documents. Identifying entities may be as simple as looking for particular strings or expressions within a document's text or it may involve complex computations to determine unique entities and their types. Different systems then choose to visualize the results of the computation in unique ways.

FeatureLens [19] uses text mining to identify words or expressions that recur across a set of documents. The system presents lists of the frequently occurring words and expressions, small overview rectangles representing each document with term positions identified by small marks, graphs of appearance count across documents, and textual views with terms highlighted. Primary users of the system may be literary scholars or journalists reviewing books or speeches. A follow-on system, POSVis [62], performs word-based part-of-speech analysis on documents and then displays the results using pixel-based overviews, word clouds, and network diagrams.

Entity Workspace [4] focuses on entity-based analysis and provides a "snap-together" visualization of entities and their attributes. Its analysis capabilities include spreading activation techniques to calculate degree-of-interest for the entities.

The IVEA system [57] uses entities of interest from a document collection to support faceted navigation and browsing across the collection. The system employs a matrix-style visualization with semantic zooming to represent the facets within documents.

Another set of systems move beyond the calculation of specific features, entities, or linguistic metrics of documents. These systems employ sophisticated text mining techniques to compute document *models and abstractions, often including concepts or themes across the documents*. Models and abstractions become especially useful as the size of the document collection grows.

The ThemeRiver technique [28] uses a river metaphor to represent temporal themes across a document collection. The river visualization extends from left-to-right to show the chronological progress of documents, and individual currents (colored bands) within the river represent different concepts. The vertical width of a current portrays its strength at a certain point in time.

Document topic modeling through latent Dirichlet allocation (LDA) [6] has become a popular technique for driving visualizations of document collections. TIARA [41] performs LDA analysis to identify themes throughout documents, and it portrays the results using a ThemeRiver-style visualization that has been augmented with word clouds. The system, thus, shows how topics grow and decline in focus over time. The system also supports user interaction to drill down and provide more detail on concept regions and to see the actual documents (e-mails) generating the concepts. TIARA can be used in many domains such as consumer reviews, e-mail, and news. TextFlow [14] extends TIARA, showing how topics emerge and merge over time, how keywords relate to topics, and critical events within topics.

Parallel Topics [20] also employ LDA to model topics across a document collection and uses a ThemeRiver style visualization to present the results, coupled with a Topic Cloud to show important terms within topics, and a parallel coordinates visualization to show how individual documents contribute to the different topics. Other systems use LDA but provide different visualizations of the identified topics including word and topic lists [9], word clouds and sentences [23], force-directed networks [27], or custom-designed 2D projections [11].

The FacetAtlas system [8] helps an analyst understand relationships between entities and facets within collections of documents sharing traits similar to academic articles. FacetAtlas uses a density map-style visualization with bundled edge connections between facets and entities along with rich interactive operations to present complex relationships between concepts in a document collection. Users can either search for specific concepts or interactively explore through the visualization interface.

The IN-SPIRE [26], [30] system takes a different approach to visualizing document themes. It utilizes powerful automated analysis, clustering, and projection capabilities, primarily operating at the document level. IN-SPIRE computes high-dimensional similarities of documents and then visualizes these relationships through galaxy or themescape style projected representations that show the documents grouped into multiple clusters.

Finally, some visual analytics systems focus not on unique visualizations of text and documents but on creating environments, where an analyst can analyze and reason about the documents. Often these systems use visual representations to help analysts explore the documents and develop hypotheses, and their target domain is frequently intelligence analysis. The systems' main goal typically is to give an investigator a faster, better understanding of a large collection of documents to help understand plots, themes, or stories across the collection.

nSPACE/TRIST/Sandbox [32], [67] provide sophisticated document analysis including entity identification and relations, trend analysis, clustering, and other automated operations. The systems present the documents through views of the documents' text or via groups of documents as small icons, but they augment this representation with sophisticated user interface flexibility for analysts to reason and develop stories about the data.

Commercial tools such as i2's Analyst Notebook [31] help intelligence, law enforcement, and fraud investigators work with document collections, among other types of data. Analyst's Notebook primary visualization is a node-link graph that shows connections between key entities in an investigation. Typically, however, the human investigator establishes these connections and constructs linkages.

As we will show in the following sections, *our contribution beyond this vast body of related work centers around the breadth of computational analysis techniques paired with a suite of rich interactive visualizations and integrating the two in a fluid, consistent manner*. Jigsaw provides multiple, varied perspectives to portray analysis results that allow the investigator to rapidly explore documents in a flexible manner. The particular emphasis on communicating entity connections across documents within concept-, temporal-, and sentiment-based perspectives also distinguishes it from existing systems.

3 Computational Text Analyses

An earlier version of Jigsaw, described in [55], focused on interactive visualization support rather than on computational modeling and analysis of documents' text. In an evaluation study [34], we found that the system was overall useful and supported a variety of strategies participants used to conduct their investigations on a document collection. However, we also found a number of situations in which the participants might have benefitted from additional information provided by computational text analysis, especially to get started with their investigation.

Some participants first read many of the documents to gain familiarity with the collection. Automated text summarization could have helped them to speed up the initial reading by reducing the amount of text to examine; document metrics, such as documents' date or length, could have provided order and structure to make the initial familiarization more efficient. Other participants focused early in their investigation on certain entities and tried to learn everything about them. Document similarity measures or features for recommending related documents could have supported this task by highlighting related information in other documents; showing documents clustered by content also could have helped them to step back and see the topics already examined or overlooked. Another group of study participants first randomly selected a few documents for acquiring evidence on which to start their investigation. Clustering documents by content could have been beneficial to help them to choose documents from different clusters for broader initial evidence.

We made similar observations on the potential benefit of computational analyses from our own use of Jigsaw, especially through our participation in the VAST Contest and Challenges [24], [42], [70] as well as from other researchers' use of the system [49], [53]. In addition, we noticed that sentiment analysis would be another useful computational technique because product reviews are a natural document set for an investigation.

Computational text analyses are not without their own set of issues and concerns, however. As Chuang et al. [11] note, text mining algorithms generate *models* of a document collection to be visualized, as opposed to source data about the documents. When models are presented to the analyst, *interpretation* and *trust* arise as important concerns. In Jigsaw, we use an extensive suite of interactive visualizations to provide multiple perspectives on analysis results, thus enabling the analyst to review and explore the derived models and determine which are most helpful.

We now describe the suite of computational analyses added to the system and, most importantly, we focus on how the analyses integrate with different visualizations. First, we explain each analysis measure and how Jigsaw presents its results. Subsequently, we provide two example usage scenarios that illustrate how an analyst explores a document collection with the system (Section 4), and we present the implementation details of the analysis algorithms (Section 5). Our main focus has been on developing techniques for smoothly combining the computational analyses with interactive visualizations. We have emphasized an integrated user experience throughout, one that provides information where and when it is most helpful and that ideally feels natural and coherent to the analyst using it for an investigation.

3.1 Document Summarization

Jigsaw provides three different techniques to summarize a document or a set of documents: One sentence summaries, word clouds, and keyword summaries. A one sentence summary—a determination of the most significant sentence—of a single document helps analysts first to decide whether to read the full text of the document and subsequently to recall the content of a document read earlier. Jigsaw presents a one sentence summary above the full text of each document in its Document View (Fig. 4). Additionally, the one sentence summary appears via tooltip, wherever a document is presented through icon or name. Word clouds, the second type of document summary, help analysts to quickly understand themes and concepts within sets of documents by presenting the most frequent words across the selected documents. Jigsaw presents word clouds of selected documents in its Document View and flexibly allows a fewer or greater number of words to be shown. The final type of summary, keyword summaries of document sets, labels sets of grouped documents in the Document Cluster View (Fig. 5), and Document Grid View (Fig. 11, left) to help an analyst know what the group is about. Keyword summaries are based on different metrics: Word frequency in each set, word uniqueness across sets, or a combination of both. Summaries based on word frequency help to understand the content of each set, word summaries based on uniqueness help to analyze differences among sets. Jigsaw allows the analyst to interactively change the metric chosen. Overall, document summarization helps an analyst to quickly decide whether a document (or set of documents) is relevant for a specific task or question at hand and whether it should be investigated further.

3.2 Document Similarity

The similarity of two documents is measured in two different ways in Jigsaw: Relative to the text within the documents or to the entities connected to the documents. The latter similarity measure is of particular interest for semistructured document collections such as publications in which metadata-related entities (e.g., authors, years, and conferences) are not mentioned in the actual document text. Document similarity measures help an analyst to determine if a document is unique (an outlier in the collection) or if there exist related documents that should be examined as well. We implemented a new view in Jigsaw (the Document Grid View) to present, analyze, and compare document similarity measures. The view organizes the 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. In all other views showing documents, an analyst can retrieve and display the five most similar documents to any document through a simple menu command.

3.3 Document Clustering

Clustering of similar and related documents also is based on either document text or on the entities connected to a document. Clusterings can be either computed fully automatically (using default values for the parameters of the clustering algorithm), or the analyst can specify the number of clusters and themes within clusters by selecting seed documents. Additionally, the analyst can interactively change clusters and define new clusters based on identified entities or keyword searches across the document collection. Document clustering partitions the documents into related groups to help an analyst explore the collection more systematically. Jigsaw presents clusterings in its Document Cluster View. The Document Grid View also provides an option to organize the documents by cluster when showing document metrics.

3.4 Document Sentiment Analysis and Other Metrics

Jigsaw computes a document's sentiment, subjectivity, and polarity, as well as other attributes such as a document's length and its number of connected entities. These metrics help an analyst seeking documents that are particularly high or low in key attributes. Jigsaw integrates and presents these metrics in its new Document Grid View. One metric can be used to determine the order of the documents within the grid, and a second metric (or the first metric again) can be mapped to the documents' color. The combined representation of any two of these metrics (by the documents' order and color) provides a flexible and powerful analytical view.

3.5 Identifying Entities in the Documents

The initial version of Jigsaw used a statistical entity identification approach from the GATE [15] package. We have added additional packages for automated entity identification, and Jigsaw now provides three different approaches for automatically identifying entities of interest in text documents: 1) statistical entity identification, 2) rule-based entity identification, and 3) dictionary-based entity identification. It uses statistical approaches from GATE, Lingpipe,[1] the OpenCalais webservice,[2] and the Illinois Named Entity Tagger [52] to identify a variety of entity types, including person, place, organization, date, and money. For the rule-based approach, we define regular expressions that match dates, phone numbers, zip codes, as well as e-mail, web, and IP addresses. The dictionary-based approach allows analysts to provide dictionaries for domain-specific entity types that are identified in the documents using basic string matching.

The automatic identification of entities is still error prone, especially in noisy, real-world data. Therefore, Jigsaw also provides functionality to correct errors in the set of identified entities. Within different visualizations, an analyst is able to add entities that were missed (false negatives), remove entities that were wrongly identified (false positives), change the type of entities, and define two or more entities as aliases.

3.6 Recommending Related Entities

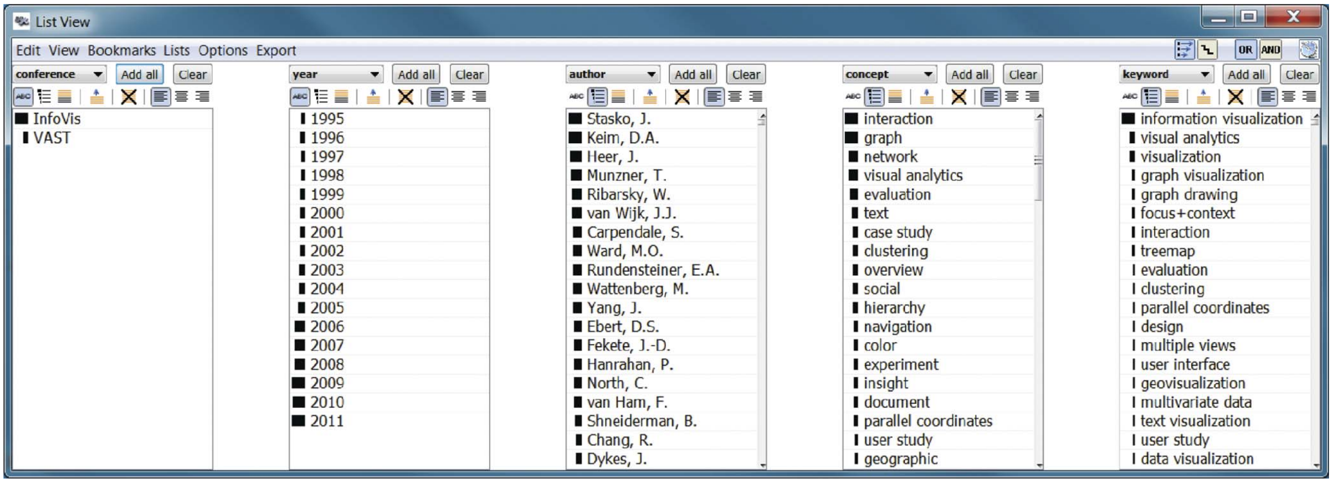
To find embedded connections among entities (that might be connected via a long chain of other entities and documents), Jigsaw recommends related entities for further examination. The recommended entities are computed by searching for connecting paths between two or more entities in the document-entity network. The chain(s) of connected entities and documents are presented in the Graph View.

4 Investigative Scenarios

To better understand how these computational analysis techniques operate within Jigsaw and aid an investigation,

[1] http://alias-i.com/lingpipe.

[2] http://www.opencalais.com.



conference	year	author	concept	keyword
InfoVis	1995	Stasko, J.	interaction	information visualization
VAST	1996	Keim, D.A.	graph	visual analytics
	1997	Heer, J.	network	visualization
	1998	Munzner, T.	visual analytics	graph visualization
	1999	Ribarsky, W.	evaluation	graph drawing
	2000	van Wijk, J.J.	text	focus+context
	2001	Carpendale, S.	case study	interaction
	2002	Ward, M.O.	clustering	treemap
	2003	Rundensteiner, E.A.	overview	evaluation
	2004	Wattenberg, M.	social	clustering
	2005	Yang, J.	hierarchy	parallel coordinates
	2006	Ebert, D.S.	navigation	design
	2007	Fekete, J.-D.	color	multiple views
	2008	Hanrahan, P.	experiment	user interface
	2009	North, C.	insight	geovisualization
	2010	van Ham, F.	document	multivariate data
	2011	Shneiderman, B.	parallel coordinates	text visualization
		Chang, R.	user study	user study
		Dykes, J.	geographic	data visualization

Fig. 1. List View showing conference, year, author, concept, and keyword, with the last three sorted by frequency.

we present two example use scenarios: A researcher exploring academic publications to learn about a research area and a consumer exploring product reviews to help make a purchase. The two scenarios involve relatively small document collections (in the hundreds) to make the presentation here more concise. We have used Jigsaw on larger collections numbering in the thousands of documents, however, and have found the new computational analysis capabilities to be even more useful at this larger scale. Because the static descriptions in this paper cannot adequately convey the dynamic nature of the investigator's interaction with the system, we refer the reader to the accompanying videos for further illustration and elaboration of similar scenarios.

4.1 Investigative Scenario: InfoVis and VAST Papers

In this scenario, we illustrate an investigation of a data set involving all of the IEEE InfoVis and VAST conference papers from 1995 to 2011; the InfoVis conference has run from 1995 to 2011 and VAST from 2006 to 2011. The data set contains 578 documents, one for each paper presented at either of the conferences; each document includes the title and abstract of the article it represents; its entities are the paper's authors, index terms, keywords, conference, journal, and year. Additionally, we added an entity type "concept" including 80 domain-relevant terms such as *interaction*, *brushing*, *network*, and *evaluation* to be found within the articles' titles and abstracts.

To generate this data set, we gathered information about the papers from the IEEE Digital Library. Throughout the data gathering process, we performed a few cleaning steps and we resolved aliases for authors. We unified each unique author to one specific name because it was not uncommon to find initialized names or inconsistent inclusion of middle names. For keywords, we unified terms effectively meaning the same thing to one common string identifier. For example, the terms "Treemap," "tree-map," "treemaps," all were changed to the string "treemap." Jigsaw's List View (Fig. 1) was very useful in this data cleaning phase as we could enumerate all the instances of any entity type in alphabetical order and easily check for similar strings. Additionally, we identified a set of documents to serve as seeds for clustering the documents.

Clearly, our domain knowledge helped in this initial data cleaning and entity resolution. Such transformations are typically necessary in any analysis of semistructured text document information [2]. Jigsaw allows the results of such a process to be saved as an XML data file for sharing with others. In fact, we have made this conference paper data set available on the web.[3]

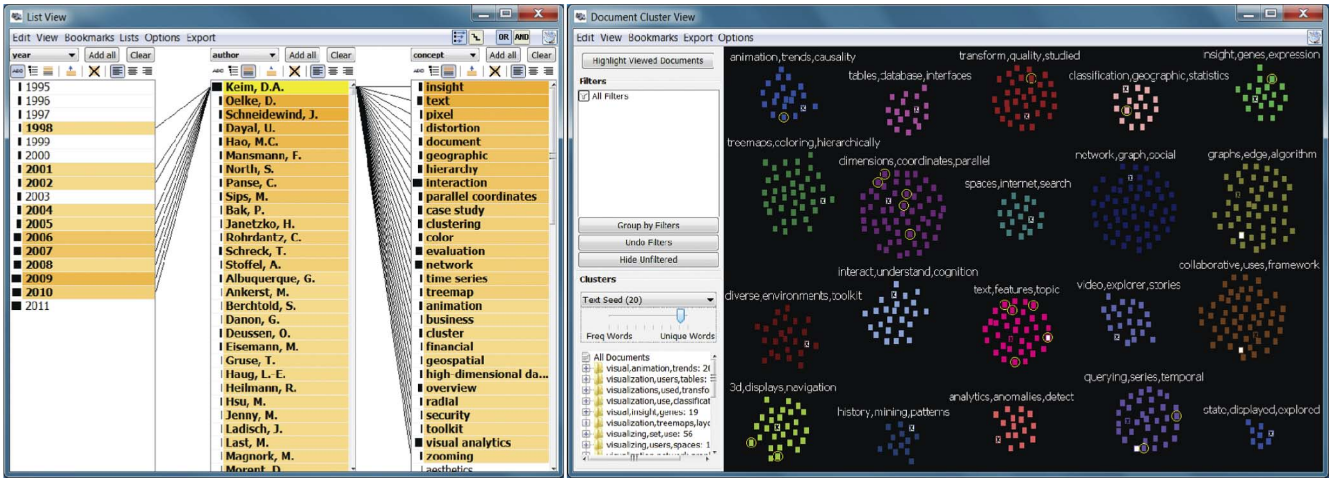
For the purpose of this scenario, we introduce a hypothetical academic researcher, Bill, who works in the database area. Bill has developed a new technique for representing database schemata as graphs or networks, and he has worked with a student to build a visualization of it. Bill knows a little about visualization research but not much detail about the IEEE InfoVis and VAST Conferences. He would like to learn whether one of these conferences would be a good fit for his paper, and if so, which one. Questions such as the following naturally arise in such an investigation:

- What are the key topics and themes of the two research areas?
- Have these topics changed over the history of the conferences?
- Who are the notable researchers in the different areas?
- Which researchers specialize in which topics?
- Are particular topics relating to his work present?
- Are there specific papers that are especially relevant?

Bill starts the investigation by examining statistics about the data set to gain an overview of the conferences and areas. Jigsaw's Control Panel (not shown here) indicates that 1,139 different researchers have contributed papers. These authors self-identified 1,197 keywords and IEEE designated 1,915 index terms for the papers. Seventy eight of the 80 concepts (we generated) appeared in at least one title or abstract.

After gaining a general overview, Bill wants to learn more specifics about the key topics and authors so he opens Jigsaw's List View (Fig. 1). He displays conference, year, author, concept, and keyword, then changes the list

[3] http://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html.



1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

Keim, D.A.
Oelke, D.
Schneidewind, J.
Dayal, U.
Hao, M.C.
Mansmann, F.
North, S.
Panse, C.
Sips, M.
Bak, P.
Janetzko, H.
Rohrdantz, C.
Schreck, T.
Stoffel, A.
Albuquerque, G.
Ankerst, M.
Berchtold, S.
Danon, G.
Deussen, O.
Eisemann, M.
Gruse, T.
Haug, L.-E.
Heilmann, R.
Hsu, M.
Jenny, M.
Ladisch, J.
Last, M.
Magnork, M.
Morent, D.

insight
text
pixel
distortion
document
geographic
hierarchy
interaction
parallel coordinates
case study
clustering
color
evaluation
network
time series
treemap
animation
business
cluster
financial
geospatial
high-dimensional da...
overview
radial
security
toolkit
visual analytics
zooming
aesthetics

All Documents
visual,animation,trends: 2(
visualization,users,tables:
visualizations,used,transfo
visualization,use,classificat
visual,insight,genes: 19
visualization,treemaps,layc
visualizing,set,use: 56
visualizing,users,spaces: 1

Fig. 2. List View (left) showing years, coauthors, and concepts connected to *Keim*. Document Cluster View (right) showing different clusters of related papers (small rectangles in different colors). Papers authored by Keim are selected (surrounded by a yellow circle).

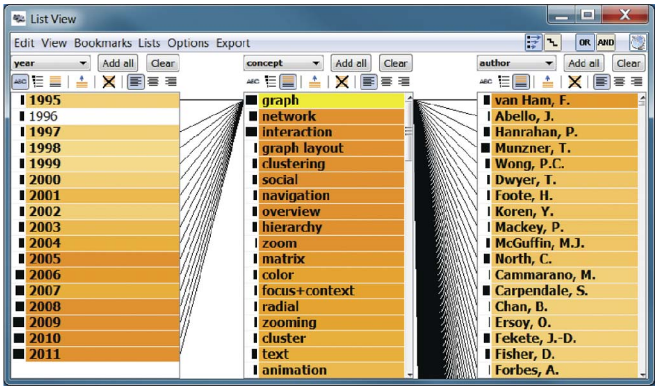
ordering from alphabetic to frequency-of-occurrence on the final three entity types to see the top-occurring entities. The small bar to the left of each entity denotes the number of documents in which it occurs. The general terms *information visualization* (101 occurrences), *visual analytics* (42), and *visualization* (40) are unsurprisingly the most frequent author-identified keywords. More interesting are the next most-common terms: *graph visualization* (18), *graph drawing* (17), *focus+context* (16), *interaction* (16), *treemap* (16), *evaluation* (14), *clustering* (13), and *parallel coordinates* (13). The term *interaction* (96) was the most frequent concept found in titles and abstracts, followed by *graph* (91), *network* (63), *visual analytics* (63), *evaluation* (55), and *text* (43). While these notions are likely familiar to someone within the field, they help a relative outsider such as Bill to understand some of the most important ideas in the research area.

Examining the author list, Bill notes that his old friend from database research, *Daniel Keim*, is one of the very top authors at the conferences. Bill is curious about Keim's papers at the conferences and decides to explore this further. He selects *Keim* in the List View and reorders the author and concept lists by strength of connection to that selection to see the entities most common with him (Fig. 2, left). Connections in Jigsaw are defined by document co-occurrence, either of identified entities in the document text, such as concepts, or of metaentities of the document, such as authors. Connection strength is defined by the number of document co-occurrences: More co-occurrences signify stronger connection. (Further details of Jigsaw's connection model are described in [55].) The List View highlights entities connected to the selection via an orange background, with darker shades indicating stronger (more frequent) connections. Entities with white backgrounds are not directly connected. The terms *insight*, *text*, *pixel*, *distortion*, *document*, and *geographic* are the most connected concepts. Keim's most frequent coauthors are *Oelke*, *Schneidewind*, *Dayal*, *Hao*, and *Mansmann*; he has published frequently from 1998 to 2010.

Bill now wants to explore ideas related to his own research. He notes that the concepts *graph* and *network* are the second and third most frequent, suggesting his work might be a good fit for these conferences. He selects the concept *graph* to learn which authors work on the topic. Jigsaw shows the most connected authors *van Ham*, *Abello*, *Hanrahan*, *Munzner*, and *Wong* and illustrates (dark shade of orange for recent years) that this has been a strong topic recently (Fig. 3). Selecting *network* shows the most connected authors *Brandes*, *Ebert*, *Fekete*, *Hanrahan*, *Heer*, and *Henry Riche* and that the topic also has been important recently. Surprisingly, the two author lists have many different names, which puzzles Bill because the two topics seem to be closely related.

To investigate further and gain a better understanding of the different topics within the conferences based on the articles' titles and abstracts, Bill switches to the Document Cluster View that displays each document in the collection as a small rectangle. Upon starting Jigsaw, Bill ran Jigsaw's automated computational analyses that calculated the similarities of all documents and a set of clusters based on these similarities.

The Document Cluster View (Fig. 2, right) shows the 578 papers divided into 20 clusters resulting from the cluster analysis. The groups are each assigned a different color and are labeled with three descriptive keywords commonly occurring in the titles and abstracts in each cluster. If the summary terms are selected based solely on

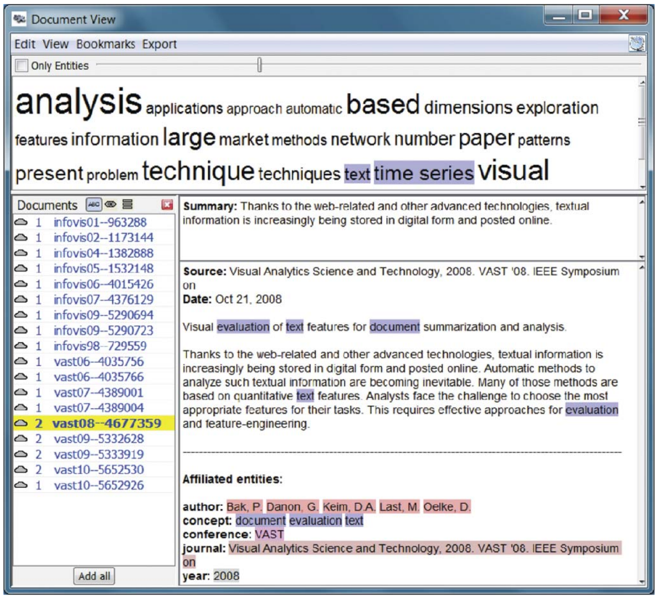


1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

graph
network
interaction
graph layout
clustering
social
navigation
overview
hierarchy
zoom
matrix
color
focus+context
radial
zooming
cluster
text
animation

van Ham, F.
Abello, J.
Hanrahan, P.
Munzner, T.
Wong, P.C.
Dwyer, T.
Foote, H.
Koren, Y.
Mackey, P.
McGuffin, M.J.
North, C.
Cammarano, M.
Carpendale, S.
Chan, B.
Ersoy, O.
Fekete, J.-D.
Fisher, D.
Forbes, A.

Fig. 3. List View with the concept *graph* selected, showing strongly connected years, concepts, and authors.


Document View
Edit View Bookmarks Export
Only Entities
analysis applications approach automatic based dimensions exploration
features information large market methods network number paper patterns
present problem technique techniques text time series visual
Documents
Add all

⛀	1	infovis01--963288
⛀	1	infovis02--1173144
⛀	1	infovis04--1382888
⛀	1	infovis05--1532148
⛀	1	infovis06--4015426
⛀	1	infovis07--4376129
⛀	1	infovis09--5290694
⛀	1	infovis09--5290723
⛀	1	infovis98--729559
⛀	1	vast06--4035756
⛀	1	vast06--4035766
⛀	1	vast07--4389001
⛀	1	vast07--4389004
⛀	**2**	**vast08--4677359**
⛀	2	vast09--5332628
⛀	2	vast09--5333919
⛀	2	vast10--5652530
⛀	1	vast10--5652926

Summary: Thanks to the web-related and other advanced technologies, textual information is increasingly being stored in digital form and posted online.

Source: Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on
Date: Oct 21, 2008

Visual evaluation of text features for document summarization and analysis.

Thanks to the web-related and other advanced technologies, textual information is increasingly being stored in digital form and posted online. Automatic methods to analyze such textual information are becoming inevitable. Many of those methods are based on quantitative text features. Analysts face the challenge to choose the most appropriate features for their tasks. This requires effective approaches for evaluation and feature-engineering.

Affiliated entities:

author: Bak, P. Danon, G. Keim, D.A. Last, M. Oelke, D.
concept: document evaluation text
conference: VAST
journal: Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on
year: 2008

Fig. 4. Document View showing all the papers authored by *Keim*. Above the selected document's text (right) is a one sentence summary and below are the affiliated entities. The word cloud (top) summarizes all documents loaded in the view.

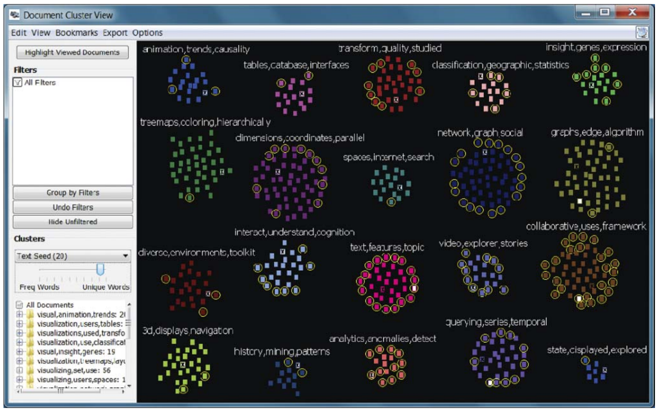

Document Cluster View
Edit View Bookmarks Export Options
Highlight Viewed Documents
Filters
All Filters
Group by Filters
Undo Filters
Hide Unfiltered
Clusters
Text Seed (20)
Freq Words
Unique Words
All Documents
visual,animation,trends: 2(
visualization,users,tables:
visualizations,used,transfo
visualization,use,classificat
visual,insight,genes: 19
visualization,treemaps,layc
visualizing,set,use: 56
visualizing,users,spaces: 1
animation,trends,causality
tables,database,interfaces
transform,quality,studied
classification,geographic,statistics
insight,genes,expression
treemaps,coloring,hierarchical y
dimensions,coordinates,parallel
spaces,internet,search
network,graph social
graphs,edge,algorithm
interact,understand,cognition
collaborative,uses,framework
diverse,environments,toolkit
text,features,topic
video,explorer stories
3d,displays,navigation
history,mining,patterns
analytics,anomalies,detect
querying,series,temporal
state,displayed,explored

Fig. 5. Document Cluster View with the VAST Conference papers highlighted. Note the clusters where they provide a strong presence.

their frequency, common terms such as "data" and "visualization" represent many clusters, which likely is not useful. The Cluster View provides a word frequency slider (left, lower center) for the investigator to interactively modify to show either more common or more unique terms affiliated with each cluster. Bill moves the frequency slider to the right, thus labeling clusters with terms more unique to that cluster. The resulting cluster labels represent important topics in these areas including toolkits, treemaps, text, animation, parallel coordinates, social networks, 3d, and databases (Fig. 2, right).

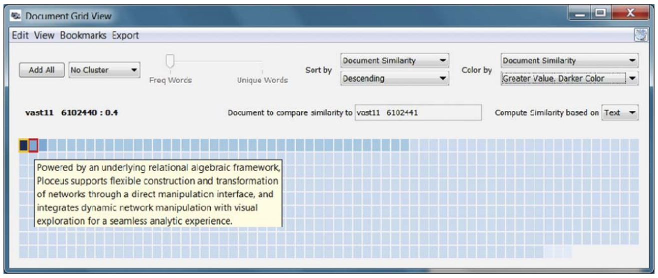
Bill is curious which clusters his friend Daniel Keim's papers fall into. He applies cross-view selection and filtering [65], one key capability of Jigsaw. It can, for example, show the topics (clusters) in which an author publishes simply by selecting that author in any other view. Selecting *Keim* in the List View (Fig. 2, left) immediately updates the Cluster View (Fig. 2, right) and highlights (yellow circles around document rectangles) the papers *Keim* has authored. As shown in the figure, his work is relatively focused with five papers each in the "dimensions, coordinates, parallel" and "text, features, topic" clusters, and eight other papers scattered among six other clusters. Knowing Keim's research, Bill is quite surprised to see none of his papers in the cluster with "database" as a descriptive word. He decides to load all of Keim's papers into a Document View to examine them more closely.

The Document View (Fig. 4) presents a list of documents (left) with the selected (yellow highlight) document's text and related information presented to the right. Below the text are the associated entities, and above the text is the one sentence summary of the document computed by Jigsaw's summary analysis (described in Section 5.2). The word cloud at the top shows the most common words (with highlighted keywords and concepts) in the abstracts of these loaded papers. Bill reviewed all the papers quickly and noticed that indeed none were about database research. He grows a little concerned about whether these conferences would be a good fit for his paper.

Next, Bill wants to understand the evolution of topics in the conferences over time to learn which have waned and which have been growing in importance recently. To do so, he selects the first four years (1995 to 1998, all InfoVis) in the List View and notices strong connections to the "internet", "toolkit," and "3d" clusters in the Cluster View; additionally, the List View shows strong connections to the concepts *interaction*, *case study*, *navigation*, and *animation*, with the concepts *network* and *graph* as the sixth and seventh most frequent. Selecting the most recent four years (2008 to 2011, both InfoVis and VAST) illuminates strong connections to multiple clusters but only connections to one document in the "3d" cluster and to two documents in the "internet" cluster. These topics clearly have waned over time. The terms *graph* and *network* are each in the top five connected concepts; thus, Bill sees how they have remained strong notions throughout the history of the conferences.

Bill next wants to better understand how the two conferences differ, so he explores the key concepts and ideas in each. He selects each conference, one at a time, in the List View and observes the connections. Among the 10 most common concepts for each conference, five terms appear in both: *interaction*, *network*, *evaluation*, *graph*, and *case study*; the five other unique terms for InfoVis are *overview*, *hierarchy*, *color*, *navigation*, and *experiment*, and for VAST are *visual analytics*, *text*, *collaboration*, *clustering*, and *insight*. As shown in Fig. 5, VAST papers (far fewer in number) occupied more than half of the "analytics, anomalies, detect," "video, explorer, stories," and "collaborative, uses, framework" clusters. These simple interactions help Bill begin to understand the subtle differences in the two conferences. His work still appears to fit well into either, however.

To learn more about the papers potentially related to his own work, Bill uses cross-view filtering in an opposite manner as he did earlier. He selects an entire cluster in the Document Cluster View and observes the resulting connections in the List View. For example, selecting the potentially related "network, graph, social" cluster shows that *Shneiderman*, *Fekete*, *Henry Riche*, *McGuffin*, *Perer*, and *van Wijk* are highly connected authors to its papers. Another potentially



No Cluster

Sort by: Document Similarity, Descending

Color by: Document Similarity, Greater Value, Darker Color

vast11 6102440 : 0.4

Document to compare similarity to: vast11 6102441

Compute Similarity based on: Text

Powered by an underlying relational algebraic framework, Ploceus supports flexible construction and transformation of networks through a direct manipulation interface, and integrates dynamic network manipulation with visual exploration for a seamless analytic experience.

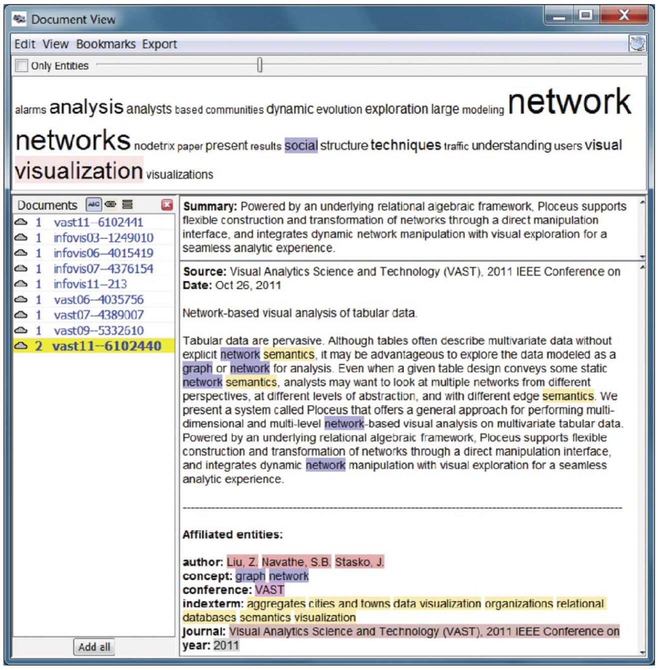
Fig. 6. Document Grid View with the document (small rectangle) order and shading set to correspond to the document's similarity to the selected Orion paper.

related cluster to Bill's work, "graphs, edge, algorithm" has top authors *Koren*, *Munzner*, *Abello*, *Ma*, and *van Ham*, all different than those in the previous cluster.

Bill decides to explore the papers in the "graphs, edge, algorithm" cluster. Since there are many, he moves his mouse pointer over the small rectangles in that cluster to quickly read a one sentence summary (tooltip) of each document. This document summary tooltip is available in other views such as the Document Grid View (Fig. 6) and the Graph View, where small iconic representations of documents are shown. None of the papers in this cluster seem to be relevant to his research; they are not about the general representation of structure and relationships in networks but about specific details of layout techniques and their mathematical optimizations. Therefore, he moves on to the "network, graph, social" cluster. Here, he discovers a paper whose summary sparks his interest: "Despite numerous advances in algorithms and visualization techniques for understanding such social networks, the process of constructing network models and performing exploratory analysis remains difficult and time consuming." Bill decides to load all the papers from this cluster into a Document View and selects this paper's icon in the Document Cluster View, thus also displaying it in the Document View. He reads the abstract of the VAST '11 paper by Heer and Perer about their Orion system and notices that it is definitely related to his work.

Bill now wants to know if papers similar to the Orion one have been published at the conferences. To find out, he uses Jigsaw's Document Grid View. The Document Grid View displays all the documents and is able to sort them by various text metrics, one being similarity to a base document. The Document Grid View in Fig. 6 shows the similarity of papers compared to the Orion paper.

Bill decides to examine the most similar papers more closely, so he selects the eight most similar ones and displays them in a Document View (Fig. 7). He observes that four of the eight papers are from InfoVis and four are from VAST. However, the paper most similar to the Orion paper is also from VAST '11 and is titled "Network-based visual analysis of tabular data." Upon reading the abstract, Bill learns that his work is quite similar to that done in this paper. Thus, he has both found some very relevant related work to explore further, and he has determined that his new paper likely would fit in either conference, but VAST may be a slightly better match.


Document View
Edit View Bookmarks Export
Only Entities
alarms analysis analysts based communities dynamic evolution exploration large modeling network
networks nodetrix paper present results social structure techniques traffic understanding users visual
visualization visualizations
Documents
Add all

1 vast11–6102441
1 infovis03–1249010
1 infovis06–4015419
1 infovis07–4376154
1 infovis11–213
1 vast06–4035756
1 vast07–4389007
1 vast09–5332610
2 vast11--6102440

Summary: Powered by an underlying relational algebraic framework, Ploceus supports flexible construction and transformation of networks through a direct manipulation interface, and integrates dynamic network manipulation with visual exploration for a seamless analytic experience.

Source: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on
Date: Oct 26, 2011

Network-based visual analysis of tabular data.

Tabular data are pervasive. Although tables often describe multivariate data without explicit network semantics, it may be advantageous to explore the data modeled as a graph or network for analysis. Even when a given table design conveys some static network semantics, analysts may want to look at multiple networks from different perspectives, at different levels of abstraction, and with different edge semantics. We present a system called Ploceus that offers a general approach for performing multi-dimensional and multi-level network-based visual analysis on multivariate tabular data. Powered by an underlying relational algebraic framework, Ploceus supports flexible construction and transformation of networks through a direct manipulation interface, and integrates dynamic network manipulation with visual exploration for a seamless analytic experience.

Affiliated entities:

author: Liu, Z. Navathe, S.B. Stasko, J.
concept: graph network
conference: VAST
indexterm: aggregates cities and towns data visualization organizations relational databases semantics visualization
journal: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on
year: 2011

Fig. 7. Document View with the eight most similar papers to the Orion paper loaded. Selected here is the most similar document.

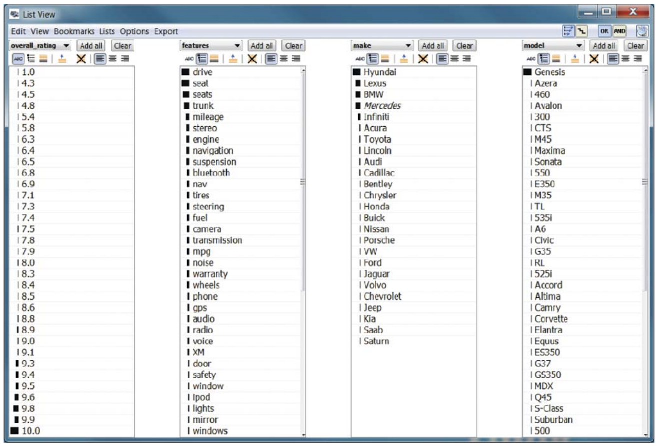
Through this abbreviated scenario, we illustrated how Jigsaw's analysis and visualization capabilities help analysts to gain quick insight on places to start an investigation, to learn about the key entities and topics in certain areas, and to explore connections and relationships in more depth. We also showed how it helps identify leaders, rapidly summarize sets of documents, compare and contrast information, find similarities and differences, and determine what should be investigated in more depth at a later point.

As shown in this scenario, investigative analyses of textual documents are often open ended and explorative in nature: Detailed questions or precise hypotheses may not be known at the beginning of an investigation but rather arise and evolve as the investigation unfolds. Analysts often switch back and forth between analyzing general trends, such as examining key topics, their relationships, and how they change over time, and more focused explorations about specific entities. Formulating new questions and finding supportive as well as contradictive evidence are fundamental tasks throughout these types of investigations.

4.2 Investigative Scenario: Car Reviews

The next scenario illustrates a different kind of investigation using documents—a consumer, Mary, who is shopping for a car. A colleague is selling his 2009 Hyundai Genesis, so to learn more about this particular model Mary examines a document collection consisting of 231 reviews of the car from the edmunds.com website. Mary wants to gain a general sense of consumers' views of the car and determine whether she should buy it. Specific concerns and goals that have arisen in her mind include:

- Identify and understand the important topics being discussed throughout the reviews,
- Learn the strong and weak points of the car,



overall_rating: 1.0, 4.3, 4.5, 4.8, 5.4, 5.8, 6.3, 6.4, 6.5, 6.8, 6.9, 7.1, 7.3, 7.4, 7.5, 7.8, 7.9, 8.0, 8.3, 8.4, 8.5, 8.6, 8.8, 8.9, 9.0, 9.1, 9.3, 9.4, 9.5, 9.6, 9.8, 9.9, 10.0

features: drive, seat, seats, trunk, mileage, stereo, engine, navigation, suspension, bluetooth, nav, tires, steering, fuel, camera, transmission, mpg, noise, warranty, wheels, phone, gps, audio, radio, voice, XM, door, safety, window, Ipod, lights, mirror, windows

make: Hyundai, Lexus, BMW, *Mercedes*, Infiniti, Acura, Toyota, Lincoln, Audi, Cadillac, Bentley, Chrysler, Honda, Buick, Nissan, Porsche, VW, Ford, Jaguar, Volvo, Chevrolet, Jeep, Kia, Saab, Saturn

model: Genesis, Azera, 460, Avalon, 300, CTS, M45, Maxima, Sonata, 550, E350, M35, TL, 535i, A6, Civic, G35, RL, 525i, Accord, Altima, Camry, Corvette, Elantra, Equus, ES350, G37, GS350, MDX, Q45, S-Class, Suburban, 500

Fig. 8. List View showing the overall rating, feature, make, and model entity types and their values from the reviews. The last three are sorted by frequency.

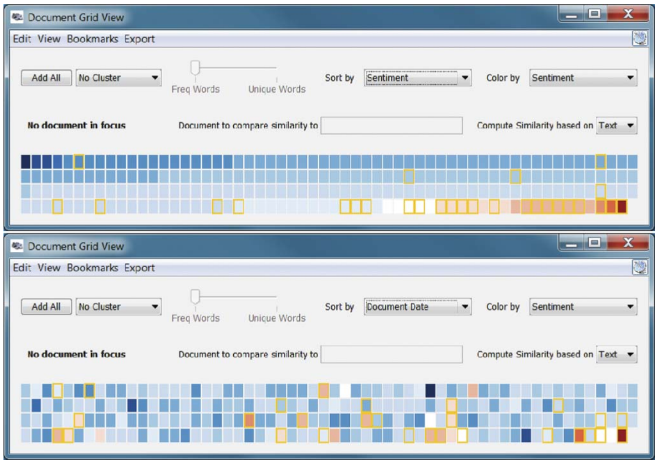
- Determine whether perceptions of the car have improved or weakened over time,
- Identify the key competitive makes/models of cars,
- Judge whether particular attributes of the car such as its gas mileage, power, sound system, and reliability are good.

Mary could, of course, examine these 231 reviews one-by-one from the website just as anyone could do when exploring a collection of consumer reviews or webpages retrieved from a search engine. However, this process is tedious and may not illuminate well the key themes and connections across the reviews.

For illustrating Mary's use of Jigsaw in this scenario, we scraped reviews of the 2009 Genesis from the edmunds.com website and imported them into Jigsaw. Each review, including its title and main narrative written by a consumer, is represented as a document. The document's entities include various rating scores (e.g., exterior design, fuel economy, and reliability) that the review author explicitly designated. We also calculated an overall rating that is the average of all the individual ratings. We added three other entity types to be found within the document text (title and review narrative): Car make (e.g., *Audi*, *Ford*, *Lexus*), car model (e.g., *525i*, *Avalon*, *ES350*), and car "feature," for which we defined 57 general terms about cars such as *seat*, *trunk*, *transmission*, and *engine*.

To get an overview of the reviews, Mary begins her investigation by invoking Jigsaw's List View (Fig. 8). She displays the overall ratings from consumers, as well as the features, makes, and models discussed in the reviews, each sorted by frequency. Mary notices that the review ratings are generally high (indicated by longer frequency bars near the bottom of the first list); *drive*, *seat(s)*, *trunk*, and *mileage* are the most mentioned features; *Lexus*, *BMW*, *Mercedes*, and *Infiniti* are the most mentioned makes (excluding *Hyundai* itself); and *Azera*, *460*, *Avalon*, *300*, and *CTS* are the most mentioned models (excluding *Genesis* itself). This is useful information to know about the key competitive cars and most commented-upon features of the Genesis.

Although the ratings are generally good for the car, Mary wants to know more details about reviewers' thoughts. An analysis of the sentiment [26] of the reviews is useful here.



Sort by: Sentiment
Color by: Sentiment
No document in focus
Document to compare similarity to:
Compute Similarity based on: Text

Sort by: Document Date
Color by: Sentiment
No document in focus
Document to compare similarity to:
Compute Similarity based on: Text

Fig. 9. Document Grid View showing all the reviews colored by sentiment: Blue indicates positive, white neutral, and red negative. The top view displays the documents sorted by sentiment as well, while the bottom view shows them ordered by date ranging from the top-left (oldest) to bottom right (newest).

To calculate sentiment, Jigsaw uses a dictionary-based approach, searching for positive or negative words throughout the document text. Here, Mary uses Jigsaw's capability to augment the dictionary by domain-specific words. For example, terms such as "quiet" and "sweet" are positive car sentiment words, while "lemon" and "clunk" indicate negative sentiment. Mary opens the Document Grid View and orders and colors the reviews by sentiment (Fig. 9, top). Positive reviews are colored blue and shown first, neutral reviews are colored white and appear next, and negative reviews are colored red and shown last. Darker shades of blue and red indicate stronger positive and negative sentiment, respectively. At first glance, the reviews for the Genesis appear to be positive overall, roughly mirroring the overall rating scores shown in the List View.

Mary once had a car that developed a number of problems after a year of driving it, so she is curious what the most recent reviews of the car express. Thus, she changes the order of the reviews in the Document Grid View to be sorted by date, as shown in Fig. 9 (bottom). The oldest review from 06/26/2008 is placed in the top-left position in the grid, and the most recent review from 07/24/2011 is in the bottom-right position. The view indicates that the earlier reviews were generally positive (shaded blue) but the more recent reviews begin to show more negative (red) perceptions. The most recent review is, in fact, the most negative, which is a concern. This trend might indicate that some issues with the car were not apparent when it first appeared but were revealed over time as the car matured.

To learn more about the car's potential weaknesses, Mary sorts the feature entities in the List View by their strength of connection to these negative reviews with overall rating below 8 (Fig. 10). The terms *seat*, *tires*, *transmission*, *steering*, and *suspension* appear as the features most connected to the negative reviews, and Mary wants to investigate perceptions of these particular car features further.

For this task, document clustering by concept in Jigsaw is useful. Mary switches back to the Document Grid View and

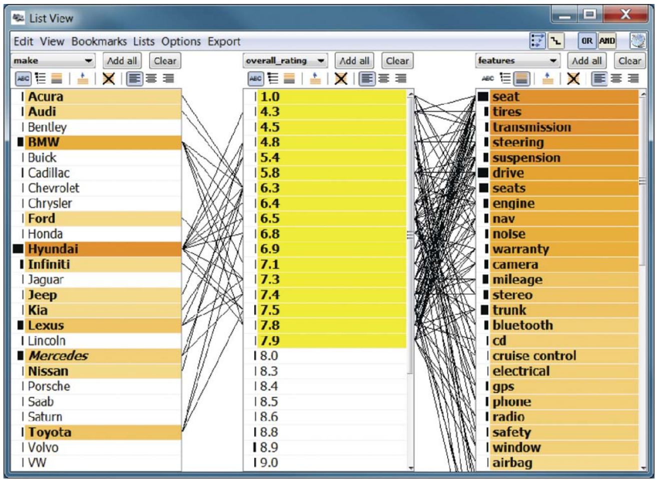


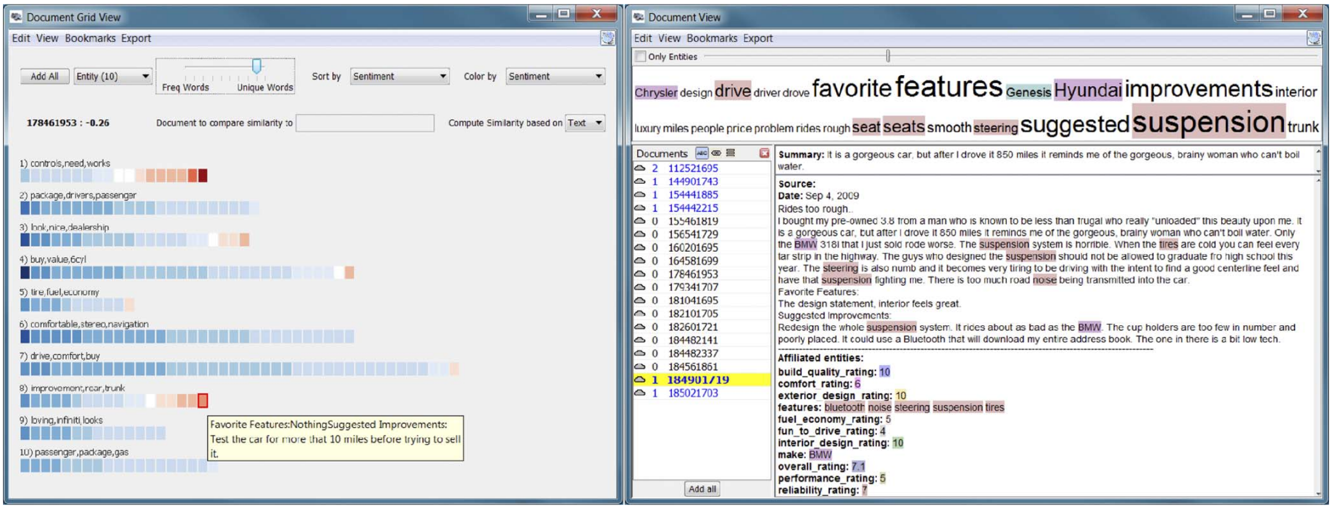
Fig. 10. List View showing the make, overall rating, and feature entity types. Low-rated reviews from 1.0 to 7.9 are selected and the feature list is sorted by connection strength to these selections.

sorts the reviews into 10 clusters, where document similarity is calculated by Jigsaw based on the set of entities connected to each review. The clusters are labeled with descriptive keywords and the documents within each cluster are ordered and colored by their sentiment (Fig. 11, left). The majority of the negative reviews aggregate into clusters 1 and 8 described by the terms "controls, needs, works" and "improvement, rear, trunk," respectively. It is not clear what each of these clusters is describing, so Mary loads the documents from each into a separate Document View to learn more.

The word clouds from each view highlight the most common words found in each review. The terms "suggested improvements" and "favorite features" are found in every review, so they are expectedly large. Similarly, the words "Hyundai" and "Genesis" also are common. However, the first cluster's word cloud also shows the word "transmission" in a large size, as does the second cloud for the word "suspension" (see Fig. 11, right). This observation and the earlier similar finding from the List View suggest these may be key problems with the car. Mary decides to investigate further and reads all the reviews in cluster 8. She finds that the suspension is often described in a negative context, as shown in the review in Fig. 11 (right). She concludes that the suspension may indeed be a weak point of the 2009 Hyundai Genesis. Even though Jigsaw only performs document level sentiment analysis, Mary was able to also determine a type of feature-level sentiment analysis by combining the results of multiple computational analyses and coordinating their results across different visual representations of the document collection.

Mary now recalls that far more reviews were positive than negative, so she decides to examine the good aspects of the car. She selects all of the reviews giving the car a perfect overall rating of 10.0 in the List View (48 reviews in total, shown in Fig. 12). The features *drive*, *seat(s)*, *stereo*, *fuel*, and *navigation* show up as being most connected. The terms *drive* and *seat(s)* occur in many documents overall as indicated by the long bar in front of the terms in the List View, so they may not be as useful. Mary now loads the documents mentioning *stereo*, the next highest term, into a Document View and reads these reviews. She learns that the Genesis' sound system is a 17-speaker Lexicon system and the reviewers typically rave about it, a definite plus to her.

Mary also wants to learn what are the other top, competitive brands of cars to consider as alternatives. She is curious about reviews mentioning other makes of cars. Thus, she sorts the car make entity by frequency in the List View and selects the top four other mentioned makes (all luxury cars), *Lexus*, *BMW*, *Mercedes*, and *Infiniti*, one by one. She notices that, overall, the connected reviews for each receive high ratings, suggesting that the Genesis is being compared favorably with these other makes. The reviews mentioning *BMW* exhibit slightly lower overall ratings, however. Perhaps prior BMW owners are not quite as favorably impressed as owners of the three other car brands. She reads the reviews also mentioning BMW and confirms that this is true.



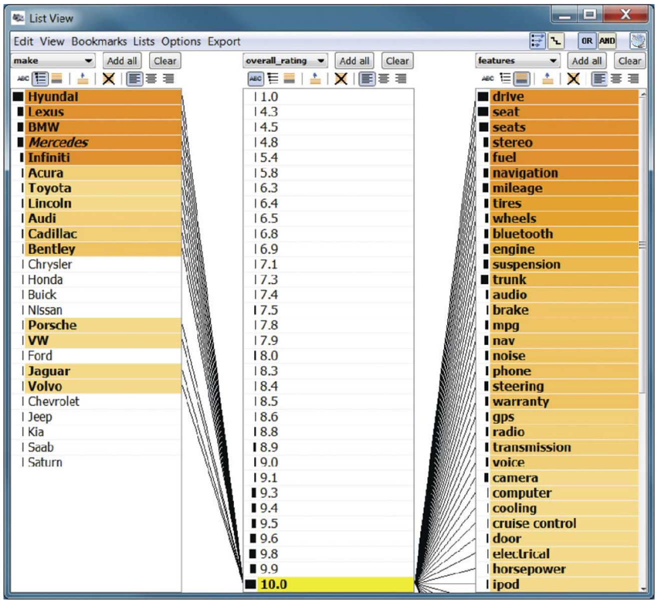
Favorite Features:NothingSuggested Improvements: Test the car for more that 10 miles before trying to sell it.

Summary: It is a gorgeous car, but after I drove it 850 miles it reminds me of the gorgeous, brainy woman who can't boil water.

Source:
Date: Sep 4, 2009
Rides too rough..
I bought my pre-owned 3.8 from a man who is known to be less than frugal who really "unloaded" this beauty upon me. It is a gorgeous car, but after I drove it 850 miles it reminds me of the gorgeous, brainy woman who can't boil water. Only the BMW 318I that I just sold rode worse. The suspension system is horrible. When the tires are cold you can feel every tar strip in the highway. The guys who designed the suspension should not be allowed to graduate fro high school this year. The steering is also numb and it becomes very tiring to be driving with the intent to find a good centerline feel and have that suspension fighting me. There is too much road noise being transmitted into the car.
Favorite Features:
The design statement, interior feels great.
Suggested Improvements:
Redesign the whole suspension system. It rides about as bad as the BMW. The cup holders are too few in number and poorly placed. It could use a Bluetooth that will download my entire address book. The one in there is a bit low tech.

Affiliated entities:
build_quality_rating: 10
comfort_rating: 6
exterior_design_rating: 10
features: bluetooth noise steering suspension tires
fuel_economy_rating: 5
fun_to_drive_rating: 4
interior_design_rating: 10
make: BMW
overall_rating: 7.1
performance_rating: 5
reliability_rating: 7

Fig. 11. Document Grid View (left) with the reviews grouped by similarity and ordered and colored by sentiment. Clusters 1 and 8 have the most negative sentiment. Document View (right) with the reviews from cluster 8 loaded. The word "suspension" is noteworthy within the word cloud at the top. The selected document illustrates an example of the views from reviews in this cluster.

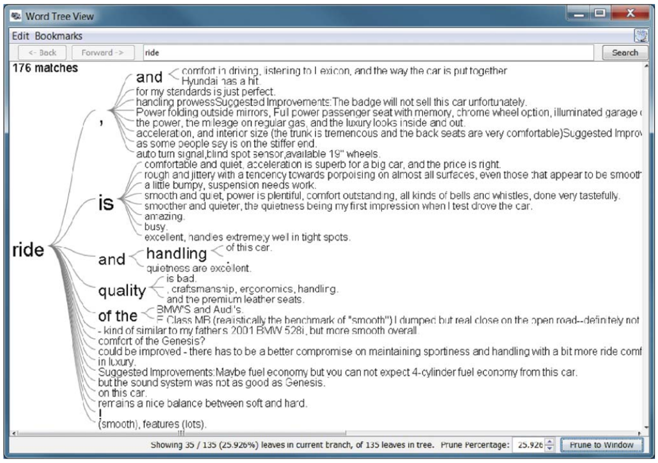


Hyundai
Lexus
BMW
Mercedes
Infiniti
Acura
Toyota
Lincoln
Audi
Cadillac
Bentley
Chrysler
Honda
Buick
Nissan
Porsche
VW
Ford
Jaguar
Volvo
Chevrolet
Jeep
Kia
Saab
Saturn

1.0
4.3
4.5
4.8
5.4
5.8
6.3
6.4
6.5
6.8
6.9
7.1
7.3
7.4
7.5
7.8
7.9
8.0
8.3
8.4
8.5
8.6
8.8
8.9
9.0
9.1
9.3
9.4
9.5
9.6
9.8
9.9
10.0

drive
seat
seats
stereo
fuel
navigation
mileage
tires
wheels
bluetooth
engine
suspension
trunk
audio
brake
mpg
nav
noise
phone
steering
warranty
gps
radio
transmission
voice
camera
computer
cooling
cruise control
door
electrical
horsepower
ipod

Fig. 12. List View showing the make, overall rating, and feature entity types. All the reviews with an overall rating of 10.0 are selected. The make list is sorted by overall frequency within the document collection and the feature list is sorted by connection strength to the 10.0 overall ratings.



ride
, and comfort in driving, listering to Lexicon, and the way the car is put together
Hyundai has a hit.
for my standards is just perfect.
handling prowessSuggested Improvements:The badge will not sell this car unfortunately.
Power folding outside mirrors, Full power passenger seat with memory, chrome wheel option, illuminated garage
the power, the mileage on regular gas, and the luxury looks inside and out.
acceleration, and interior size (the trunk is tremendous and the back seats are very comfortable)Suggested Impro
as some people say is on the stiffer end.
auto turn signal,blind spot sensor,available 19" wheels.
is comfortable and quiet, acceleration is superb for a big car, and the price is right.
rough and jittery with a tendency towards porpoising on almost all surfaces, even those that appear to be smooth
a little bumpy, suspension needs work.
smooth and quiet, power is plentiful, comfort outstanding, all kinds of bells and whistles, done very tastefully.
smoother and quieter, the quietness being my first impression when I test drove the car.
amazing.
busy.
excellent, handles extremely well in tight spots.
and handling of this car.
.
quietness are excellent.
quality is bad.
, craftsmanship, ergonomics, handling.
and the premium leather seats.
of the BMW'S and Audi's.
E Class MB (realistically the benchmark of "smooth") I dumped but real close on the open road--definitely not
- kind of similar to my fathers 2001 BMW 528i, but more smooth overall.
comfort of the Genesis?
could be improved - there has to be a better compromise on maintaining sportiness and handling with a bit more ride comf
in luxury.
Suggested Improvements:Maybe fuel economy but you can not expect 4-cylinder fuel economy from this car.
but the sound system was not as good as Genesis.
on this car.
remains a nice balance between soft and hard.
!
(smooth), features (lots).

Showing 35 / 135 (25.926%) leaves in current branch, of 135 leaves in tree.

Fig. 13. Word Tree View showing occurrences of the word "ride" and the most common phrases that follow the word in sentences within the review collection.

To learn more about the ride quality of the car, an important feature to her, Mary displays Jigsaw's Word Tree View for "ride" (Fig. 13). A Word Tree [64] shows all occurrences of a word or phrase from the reviews in the context of the words that follow it, each of which can be explored further by a click. The Word Tree View shows that reviewers have different opinions about the quality of the ride, ranging from "a little bumpy" and "rough and jittery" to "comfortable and quiet" and "excellent."

Mary's investigations of the Genesis' reviews have helped her understand overall perceptions of the car and what the most recent impressions are. The computational analyses, the sentiment analysis and document clustering in particular, facilitated the identification of the car features perceived most favorably and unfavorably by the reviewers, and Mary learned more about other competitive makes and models of cars. An important part of such an exploration is reading the individual reviews of note, which we have not emphasized here for obvious reasons of brevity. However, we must stress that this activity is a key aspect of any document corpus investigation like this. The newly integrated computational analyses in Jigsaw help to more rapidly identify the documents of note for any of a variety of attributes or dimensions.

5 Computational Analysis Algorithms

In this section, we provide a brief discussion of the text analysis algorithms we implement in Jigsaw, primarily for the reader interested in more detail. We integrate well-known algorithms for the different computational analyses, practical algorithms that can be readily implemented in Java (Jigsaw's implementation language) and that run in a "reasonable" time on computers that real clients would have. These descriptions and our experiences in designing and implementing the capabilities may be beneficial for other researchers who wish to integrate enhanced automated computational analysis in their visual analytics systems.

5.1 Preprocessing

To apply computational analyses, text documents are typically converted to a certain form of numerical vector representation. We use the standard "bag of words" encoding scheme, where each dimension corresponds to a unique term, and the value represents the term count in the document. In Jigsaw, the vocabulary that constitutes the entire set of dimensions can be based on either all the terms occurring in the document corpus or only the entities that are identified within the documents. Thus, we obtain either a term-document or an entity-document matrix.

Then, we follow standard preprocessing procedures for text data such as stemming and stop word removal. For stemming, we use the Porter Stemmer [51] implementation in the Lingpipe library. Additionally, we exclude the terms and entities that appear less than three times throughout the entire document set. (The terms and entities are only excluded from the computational analyses; they are not removed from the data set.) Based on empirical experiments, we determined that these terms do not affect the results of the computational modules significantly while the vocabulary size is reduced drastically, often up to 40 percent, which improves both the computation time and memory usage.

After building the term-document matrix, we apply TF-IDF weighting and normalization [1]. TF-IDF weighting penalizes the terms that broadly appear in many documents because they would not contribute to the differentiation of one document from another. Normalization transforms each document vector to a unit norm to overcome the dependency on the document length.

Based on this numerical encoding of textual documents, we integrate three text analytical modules into Jigsaw: Document summarization, document similarity, and document clustering. Document sentiment analysis, our fourth module, operates directly on the original document text.

5.2 Document Summarization

This module summarizes documents by extracting significant sentences. It first computes the importance scores (described below) for all the terms and for all the sentences within a single document, and then ranks the sentences with respect to the scores. The sentence with the highest importance score is determined to be the most representative sentence in the document and chosen as a summary sentence. The scored and ranked terms are used to summarize multiple documents with keywords (described in Section 5.4). To determine the sentences and the terms in a document, we use a sentence splitter and a tokenizer from the Lingpipe library.

To implement the summarization algorithm, we apply the mutual reinforcement learning method [68]. This method first decomposes each document into a set of all the terms $T = \{t_1, \ldots, t_m\}$ and a set of all the sentences $S = \{s_1, \ldots, s_n\}$. A weighted bipartite graph between T and S is built with a weight matrix $W = \{w_{ij}\} \in \mathbb{R}^{m \times n}$, where w_{ij} is the frequency of the term t_i in the sentence s_j. Then, we randomly initialize two vectors, $u \in \mathbb{R}^{m \times 1}$ and $v \in \mathbb{R}^{n \times 1}$, of the importance scores of terms and sentences, and perform a power iteration, i.e., $u = Wv$ and then $v = W^T u$, normalizing after every step. This iteration continuously passes the importance scores between terms and sentences until they converge.

5.3 Document Similarity

This module computes all the pairwise similarity scores for the documents in the corpus. The computation of similarity between two documents can be based on various measures. Although the most widely used measure is the euclidean distance, semantically, cosine similarity can be a better choice for textual data [56] and, therefore, we use it in our implementation.

To obtain semantically better results, we do not compute the similarity based on the original document vector. Instead, we first reduce its dimension by applying the latent semantic analysis (LSA) technique [17] and then compute the similarity in the resulting reduced dimensional space. By grouping semantically similar terms, LSA improves similarity scores against polysemy and synonymy problems. After experimenting with different values, we chose to set the number of reduced dimensions to 20 percent of the number of dimensions after the preprocessing step of removing terms that occur in less than three documents. LSA requires the computation of the singular value decomposition (SVD) of the term-document matrix. We use the JAMA library[4] for matrix computations such as SVD.

Using the term-document or the entity-document matrix, we can compute document similarity based on either the entire document text or on the entities identified in the documents.

5.4 Document Clustering

This module groups the documents into a given number of clusters, where similar documents fall into the same cluster. The similarity can be based on the document text or on the entities identified in the documents. We adopt the spherical k-means clustering algorithm [18], which uses cosine similarity as a distance measure.

The clustering algorithm requires the number of clusters as an input parameter. Theoretically, it is crucial to choose the "right" number of clusters to get optimal clustering results. Although there exist methods to quantitatively evaluate clusters [29], semantically it is difficult to determine the right number of clusters and achieve satisfactory results on noisy real-world data. Thus, by default, we choose 20 as the number of clusters. Our reasoning behind this choice is that, on the one hand, if the document set has fewer clusters, then our result would show a few similar clusters that can be merged into a single true cluster by humans' further analysis. On the other hand, if the document set has significantly more clusters, e.g., 50 clusters, analysts might have difficulties in understanding their structure due to an unmanageable number of clusters, even if they represent the correct clustering result. However, we also provide a user option to specify the number of clusters in case an analyst is familiar with a specific document set and has some knowledge about its structure.

In addition, the algorithm requires a list of initial seed documents for the clusters as an input parameter. Although the algorithm is not sensitive to this parameter if the document set has a clearly clustered structure, we observed that the results can vary significantly depending on the initial seeds for most real-world document sets that do not have well-defined clusters. Thus, we carefully choose initial seed documents using a heuristic in which seed documents are recursively selected such that each seed document is the least similar document to the previously selected seed document. Due to space limitation we do not discuss details, such as optimizations and exceptions of this heuristic. We also provide a user option to choose initial seed documents in case an analyst is interested in specific topics in a document set and wants to steer the cluster analysis by choosing the seed documents according to the topics of interest.

To enhance the usability of the clustering results, we summarize the content of each cluster as a list of the most representative terms within its documents. We use the algorithm described in Section 5.2 after aggregating all documents in a cluster into one single document. However, instead of the most representative sentence, we use three high-ranking terms as the summary of the cluster. We compute a number of alternative term summaries for each cluster. One summary is based only on the term frequency within a cluster, whereas another summary also takes term uniqueness across clusters into account and eliminates any terms that would occur in multiple summaries; additionally, we compute summaries that are gradually more strict on the uniqueness of summary terms (i.e., eliminating any terms that occur in 10 percent, or 20 percent, . . . , or 90 percent of the cluster summaries). As described in Section 3 the analyst can interactively switch between the different cluster summaries to gain different perspectives on the clustering result, either examining the content of individual clusters or understanding differences among the clusters.

[4] http://math.nist.gov/javanumerics/jama.

5.5 Document Sentiment Analysis

This module provides two different implementations to characterize the text in a document on a positive-to-negative scale. It does not apply the preprocessing steps discussed in Section 5.1 but operates directly on the original document text.

One implementation is based on the classifier provided in the Lingpipe library. It applies the hierarchical classification technique described in Pang and Lee [48] and requires two classifiers: One for subjective/objective sentence classification and one for polarity classification. The technique involves running the subjectivity classifier on the document text to extract the subjective sentences first and then running the polarity classifier on the result to classify the document as positive or negative. We trained the subjectivity classifier with the data provided in [48]. To train the polarity classifier, we used 2,000 product reviews (1,000 positive and 1,000 negative) extracted from amazon.com. We considered all reviews with a rating of 4 or 5 as positive and those with a rating of 1 or 2 as negative. We did not use reviews with a rating of 3.

An alternative implementation computes a quantitative sentiment score for each document on a scale from $+1$ (positive) to -1 (negative) via a dictionary-based approach, identifying "positive" and "negative" words in documents. We developed the list of words by creating two initial sets of negative and positive words and then iterating and checking results against known positive and negative documents. The results have been surprisingly good, particularly when characterizing documents strong in expected sentiment such as product reviews. We also allow the user to provide domain-specific dictionaries of positive and negative words to classify the documents. This feature was very useful for a collaborative analysis in which we examined wine reviews with a wine expert; we developed dictionaries of words that describe "good" and "bad" wines to classify the reviews.

5.6 Computation Time

The runtime of the computational analyses depends on the characteristics of the document collection (number of documents, average document length, and number of entities per document) and the available computational power (processor speed and memory size). The InfoVis and VAST papers data set in our case study has 578 documents, the average length of a paper's title and abstract is 1,104 characters (min: 159; max: 2,650), and the average number of entities per paper is 17 (min: 4; max: 58). On a desktop computer with 8 GB of memory and two 2.4-GHz Quad-Core processors, computing the summary sentences and the sentiment analysis each took 2 seconds, the text-based similarity computation took 47 seconds, the entity-based similarity computation took 10 seconds, the text-based cluster computation took 20 seconds, and the entity-based cluster computation took 15 seconds, resulting in a total computation time of less than 2 minutes. The car reviews data set is much smaller (231 documents) and all analyses finished in 12 seconds.

We also ran the analyses on other data sets with different characteristics. From a practical point of view, the computation time can be divided into three categories. For small data sets (about 500 documents), the computation time is a few minutes (coffee break), for medium-sized data sets (about 1,500 documents) the computation time is less than 1 hour (lunch break), and for larger data sets (about 5,000 documents) the computation time is several hours (overnight).

6 Discussion

Investigations on document collections proceed with the analyst gathering nuggets of information while forming new insights and deeper understanding of the document contents. Especially, when the documents are unfamiliar, an investigator may not know where to start, what is related, or how to dive more deeply into analysis. We believe that fluid integration of computational analyses with interactive visualization provides a flexible environment that scaffolds the investigator's exploratory process.

In exploration and sensemaking, investigators likely want to ask a broad set of questions and also develop new questions throughout the investigation process. Interactive visualization supports this dynamic conversation or dialog between the investigator and the data, and it makes the results of powerful computational analyses more easily accessible and contextually relevant. Our efforts to integrate enhanced computational analysis support into Jigsaw have taught us a number of lessons about this process (resulting both from an implementation perspective and from working with users of the system [7], [35]), but five in particular stand out:

1. *Make different computational analysis results available throughout the system in a variety of different contexts and views, not in just one canonical representation.* Chuang et al. [11] identify *interpretation* and *trust* as two key issues to the success of visual analytics systems. With respect to the results of computational text mining, trust seems to be a primary concern. We have found that portraying the results of mining algorithms under different perspectives better allows the analyst to inspect and interpret the algorithm's results. In particular, multiple analyses within Jigsaw appear in several different views and can be examined under different perspectives. For example, the single sentence document summaries are shown above the corresponding full document text in the Document View as one might expect, but they also are available as tooltips anywhere a document is represented iconically or by name. Clusterings are shown (naturally) in the Document Cluster View but also in the Document Grid View that simultaneously can show similarity, sentiment, and summary analysis results. Furthermore, clusters are easy to select and, thus, inspect the member documents under other analysis perspectives and views. Given any set of documents resulting from a text analysis, one simple command allows those documents to be loaded into a Document View for further manual exploration.

2. *Flexibly allow analysis output also to be used as input.* Investigators using Jigsaw can select individual documents from any analysis view and can then request to see that document's text or see related documents. Jigsaw presents the results of similarity, clustering, and sentiment analyses visually (output), but such results can be clicked on or selected by the analyst (input) to drive further exploration. This capability is pervasive throughout the system—any document or entity can be acted upon to drive further investigation. We believe this design, which helps to facilitate the core, iterative sensemaking cycle of visual

analytics [36], enables smoother, more flexible interaction with the system, ultimately leading to deeper inquiry, exploration, and increased knowledge.

3. Integrate different, independent computational analysis measures through interactive visualization to extend functionality and power. A deep integration of automated analysis with interactive visualization results in capabilities beyond each of the two components ("the whole is greater than the sum of the parts"). For example, Jigsaw provides document-level sentiment analysis, but does not analytically provide sentiment with respect to specific terms, concepts, or features within a document. However, as illustrated in the car review scenario, by first performing content-based clustering that divides the car reviews into sets of documents discussing different car features, and then visualizing the sentiment on the resulting clusters, one achieves a type of feature-based sentiment. The scenario showed how the reviewers felt negatively about the car's suspension and transmission.

4. Provide computational support for both analysis directions: Narrowing down as well as widening the scope of an investigation. Many investigations take the form of an hourglass: An analyst first confronts a large amount of data (top of the hourglass), iteratively filters and searches the data to discover a small number of interesting leads (middle of the hourglass), and then expands the data under investigation again by following connections from those identified leads (bottom of the hourglass). These new data points then represent the top of another hourglass and the analyst repeats the process. Cutting et al. [16] describe this narrowing widening, iterative process as the Scatter/Gather method. To smoothly move through the different stages of an hourglass investigation, a visual analytics system should provide support for narrowing down as well as widening the scope of analysis. Jigsaw provides a variety of analysis support for both tasks. Document clustering and sentiment analysis help narrow down the scope by limiting it to one (or a few) clusters or taking only positive or negative documents into account; document similarity and recommending related entities help widen the scope by suggesting additional relevant documents; and identified entities help with both directions: They can be used to determine a germane subset of a document set (containing one or more identified entities) or to suggest other related documents (containing the entity of interest).

5. Expose algorithm parameters in an interactive user-accessible way. The effectiveness of many computational analyses depends on the choices of their parameters. Whenever possible, visual analytics tools should provide users intuitive access to the parameter space of the underlying analyses. In Jigsaw, we expose parameters in a number of different ways. For the k-means clustering, we expose the corresponding parameters directly because they are quite intuitive. Users can either choose default values or define the number of clusters, specify whether the clustering should be based on the document text or only the entities connected to a document, and provide initial seed documents for the clusters. They then can display different clusterings from different parameter choices in multiple Document Cluster Views to compare and contrast them. We take a different approach for the cluster summarization algorithm. Instead of exposing the summarization parameters directly, we precompute a set of summarizations and let users explore the parameter space by selecting cluster summaries via an interactive slider (based on uniqueness versus frequency of the summary words). Users have preferred this approach more than exposing the (not so intuitive) parameters of the summarization algorithm directly. For the dictionary-based entity identification and the sentiment analysis, users can provide their own domain-specific dictionaries. This flexibility has proven to be very useful in various domain-specific investigations that we and others have conducted with Jigsaw (e.g., investigating wine reviews, car reviews, scientific papers, and Java code). User requests for exposing additional algorithm parameters, such as regular expressions for the rule-based entity identification approach, confirm the importance of this lesson.

7 Conclusion

Helping investigators to explore a document collection is more than just retrieving the "right" set of documents. In fact, all the documents retrieved or examined may be important, and so the challenge becomes how to give the analyst fast and yet deep understanding of the contents of those documents.

In this paper, we have illustrated methods for integrating automated computational analysis with interactive visualization for text- and document-based investigative analysis. We implemented a suite of analysis operations into the Jigsaw system, demonstrating how to combine analysis results with interactive visualizations to provide a fluid, powerful exploration environment. Further, we provided two example sensemaking scenarios that show both the methodologies and the utility of these new capabilities. We included brief descriptions of the computational analysis algorithms we chose to help readers seeking to implement similar operations in their systems. Finally, we described our experiences in building the new system and the lessons we learned in doing so.

The contributions of the work are, thus, as follows:

- Techniques for integrating computational analysis capabilities fluidly with different interactive visualizations, and realization of those techniques in the Jigsaw system.
- Illustrations of the benefits of this approach via two example sensemaking scenarios. These scenarios provide sample questions and tasks, methods to resolve them, and the analysis and insights that result.
- Guidance for HCI/visualization researchers about the implementation of practical, text-focused computational analysis algorithms.
- Design principles for the construction of future document analysis and visual analytics systems.

A particular strength of Jigsaw is its generality for analyses on different types of documents. Many other systems have been tailored to a specific style of document or content domain and, thus, provide sophisticated capabilities only in that area. Jigsaw has been applied in the domains here (academic research and consumer product reviews) and in other diverse areas such as aviation documents [49], understanding source code files for software analysis and engineering [53], genomics research based on PubMed articles [25], and investigations in fraud, law enforcement, and intelligence analysis [35], [69]. The system is available for download.[5]

Many avenues remain for future research. We admittedly have not conducted formal evaluations or user studies of these new capabilities within Jigsaw. Determining the best methods to evaluate systems like this is a research challenge unto itself. Our earlier user study involving Jigsaw [34] identified the potential benefits of the system, so we believe that the addition of the new computational analysis capabilities will provide even further value. In particular, the new capabilities address analysis needs identified in the user study and determined through earlier trial use of the system by clients.

We also plan to explore newer, more powerful methods and algorithms for calculating analysis metrics. The areas of computational linguistics, dimensionality reduction, and text mining are ripe with analysis methods such as topic modeling [6] and multiword expressions [5] that could be integrated into Jigsaw. Furthermore, allowing user-driven interactive feedback to modify and evolve the computational analyses would provide an even more flexible exploration environment.

Finally, we made a claim that to achieve its fullest potential within visual analytics, a system must deeply and seamlessly combine automated computational analysis with interactive visualization. Actually, according to the definition of visual analytics introduced in *Illuminating the Path* [58], we omitted the third key piece of the equation: Integrated support for analytical reasoning. Systems such as Jigsaw seeking to provide comprehensive analytic value also should include facilities for supporting human investigators' analytic reasoning processes and goals.

We are encouraged that the vision of visual analytics is beginning to be realized. The system and experiences described in this paper illustrate the potential of such an approach: Fluidly integrating computational data analysis algorithms with flexible, interactive visualizations provide investigators with powerful data exploration capabilities and systems.

ACKNOWLEDGMENTS

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

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search,* second ed. ACM Press, 2011.

[2] C. Bartneck and J. Hu, "Scientometric Analysis of the CHI Proceeding," *Proc. ACM Conf. Human Factors in Computing Systems (CHI),* pp. 699-708, 2009.

[3] M. Berry and M. Castellanos, *Survey of Text Mining II: Clustering, Classification, and Retrieval,* vol. XVI. Springer, 2008.

[4] E.A. Bier, S.K. Card, and J.W. Bodnar, "Principles and Tools for Collaborative Entity-Based Intelligence Analysis," *IEEE Trans. Visualization and Computer Graphics,* vol. 16, no. 2, pp. 178-191, Mar./Apr. 2010.

[5] D. Blei and J. Lafferty, "Visualizing Topics with Multi-Word Expressions," arXiv:0907.1013v1, technical report, 2009.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[7] E. Braunstein, C. Görg, Z. Liu, and J. Stasko, "Jigsaw to Save Vastopolis - VAST 2011 Mini Challenge 3 Award: 'Good Use of the Analytic Process'," *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST),* pp. 323-324, Oct. 2011.

[8] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "FacetAtlas: Multifaceted Visualization for Rich Text Corpora," *IEEE Trans. Visualization and Computer Graphics,* vol. 16, no. 6, pp. 1172-1181, Nov./Dec. 2010.

[9] A.J.B. Chaney and D.M. Blei, "Visualizing Topic Models," *Proc. Sixth Int'l AAAI Conf. Weblogs and Social Media (AAAI ICWSM),* pp. 419-422, 2012.

[10] J.-K. Chou and C.-K. Yang, "PaperVis: Literature Review Made Easy," *Computer Graphics Forum,* vol. 30, no. 3, pp. 721-730, 2011.

[11] J. Chuang, D. Ramage, C.D. Manning, and J. Heer, "Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis," *Proc. ACM Conf. Human Factors in Computing Systems (CHI),* pp. 443-452, 2012.

[12] C. Collins, S. Carpendale, and G. Penn, "DocuBurst: Visualizing Document Content Using Language Structure," *Computer Graphics Forum,* vol. 28, no. 3, pp. 1039-1046, 2008.

[13] C. Collins, F.B. Viegas, and M. Wattenberg, "Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* pp. 91-98, Oct. 2009.

[14] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z.J. Gao, X. Tong, and H. Qu, "TextFlow: Towards Better Understanding of Evolving Topics in Text," *IEEE Trans. Visualization and Computer Graphics,* vol. 17, no. 12, pp. 2412-2421, Dec. 2011.

[15] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6),* 2011.

[16] D.R. Cutting, D.R. Karger, J.O. Pedersen, and J.W. Tukey, "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections," *Proc. ACM SIGIR 15th Ann. Conf. Conf. Research Development in Information Retrieval,* pp. 318-329, 1992.

[17] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *J. Soc. Information Science,* vol. 41, pp. 391-407, 1990.

[18] I.S. Dhillon and D.S. Modha, "Concept Decompositions for Large Sparse Text Data using Clustering," *Machine Learning,* vol. 42, no. 1/2, pp. 143-175, 2001.

[19] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant, "Discovering Interesting usage Patterns in Text Collections: Integrating Text Mining with Visualization," *Proc. ACM Conf. Information and Knowledge Management (CIKM),* pp. 213-222, 2007.

[20] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Parallel Topics: A Probabilistic Approach to Exploring Document Collections," *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST),* pp. 229-238, Oct. 2011.

[21] S.G. Eick, "Graphically Displaying Text," *J. Computational and Graphical Statistics,* vol. 3, no. 2, pp. 127-142, 1994.

[22] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge Univ. Press, 2007.

[23] M.J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi, "The Topic Browser: An Interactive Tool for Browsing Topic Models," *Proc. Neural Information Processing Systems (NIPS) Workshop Challenges of Data Visualization,* 2010.

[24] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko, "Jigsaw Meets Blue Iguanodon - The VAST 2007 Contest," *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST),* pp. 235-236, Oct. 2007.

[25] C. Görg, H. Tipney, K. Verspoor, W. Baumgartner, K. Cohen, J. Stasko, and L. Hunter, "Visualization and Language Processing for Supporting Analysis Across the Biomedical Literature," *Proc. 14th Int'l Conf. Knowledge-Based and Intelligent Information and Eng. Systems,* pp. 420-429, 2010.

[26] M. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner, "User-Directed Sentiment Analysis: Visualizing the Affective Content of Documents," *Proc. Workshop Sentiment and Subjectivity in Text,* pp. 23-30, 2006.

[5] http://www.cc.gatech.edu/gvu/ii/jigsaw.

[27] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A.U. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual Analysis of Large Text Corpora with Topic Modeling," *ACM Trans. Intelligent Systems and Technology,* vol. 3, no. 2, pp. 23:1-23:26, 2012.

[28] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: Visualizing Theme Changes over Time," *Proc. IEEE Symp. Information Visualization (InfoVis),* pp. 115-123, Oct. 2000.

[29] J. He, A.-H. Tan, C.L. Tan, and S.Y. Sung, "On Quantitative Evaluation of Clustering Systems," *Clustering and Information Retrieval,* pp. 105-134, Springer, 2003.

[30] E. Hetzler and A. Turner, "Analysis Experiences using Information Visualization," *IEEE Computer Graphics and Applications,* vol. 24, no. 5, pp. 22-26, Sept./Oct. 2004.

[31] "i2 - Analyst's Notebook," http://www.i2inc.com/, 2013.

[32] D. Jonker, W. Wright, D. Schroh, P. Proulx, and B. Cort, "Information Triage with TRIST," *Proc. Int'l Conf. Intelligence Analysis,* May 2005.

[33] H. Kang, C. Plaisant, B. Lee, and B.B. Bederson, "NetLens: Iterative Exploration of Content-Actor Network Data," *Information Visualization,* vol. 6, no. 1, pp. 18-31, 2007.

[34] Y.-a. Kang, C. Görg, and J. Stasko, "How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation," *IEEE Trans. Visualization and Computer Graphics,* vol. 17, no. 5, pp. 570-583, May 2011.

[35] Y.-a. Kang and J. Stasko, "Examining the Use of a Visual Analytics System for Sensemaking Tasks: Case Studies with Domain Experts," *IEEE Trans. Visualization and Computer Graphics,* vol. 18, no. 12, pp. 2869-2878, Dec. 2012.

[36] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual Analytics: Definition, Process, and Challenges," *Information Visualization: Human-Centered Issues and Perspectives,* pp. 154-175, Springer-Verlag, 2008.

[37] *Mastering the Information Age - Solving Problems with Visual Analytics,* D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, eds. Eurographics Assoication, 2010.

[38] D.A. Keim and D. Oelke, "Literature Fingerprinting: A New Method for Visual Literary Analysis," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* pp. 115-122, 2007.

[39] G. Klein, B. Moon, and R. Hoffman, "Making Sense of Sensemaking 1: Alternative Perspectives," *IEEE Intelligent Systems,* vol. 21, no. 4, pp. 70-73, July/Aug. 2006.

[40] B. Lee, M. Czerwinski, G. Robertson, and B.B. Bederson, "Understanding Research Trends in Conferences Using Paperlens," *Proc. Extended Abstracts ACM Conf. Human Factors in Computing Systems,* pp. 1969-1972, 2005.

[41] S. Liu, M.X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, "Tiara: Interactive, Topic-Based Visual Text Summarization and Analysis," *ACM Trans. Intelligent Systems and Technology,* vol. 3, no. 2, pp. 25:1-25:28, Feb. 2012.

[42] Z. Liu, C. Görg, J. Kihm, H. Lee, J. Choo, H. Park, and J. Stasko, "Data Ingestion and Evidence Marshalling in Jigsaw," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* pp. 271-272, Oct. 2010.

[43] G. Marchionini, "Exploratory Search: From Finding to Understanding," *Comm. ACM,* vol. 49, no. 4, pp. 41-46, Apr. 2006.

[44] G. Marchionini and R.W. White, "Information-Seeking Support Systems," *Computer,* vol. 42, no. 3, pp. 30-32, Mar. 2009.

[45] D. Oelke, P. Bak, D. Keim, M. Last, and G. Danon, "Visual Evaluation of Text Features for Document Summarization and Analysis," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* pp. 75-82, Oct. 2008.

[46] D. Oelke, M. Hao, C. Rohrdantz, D. Keim, U. Dayal, L.-E. Haug, and H. Janetzko, "Visual Opinion Analysis of Customer Feedback Data," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* pp. 187-194, Oct. 2009.

[47] W.B. Paley, "TextArc: Showing word Frequency and Distribution in Text," *Proc. IEEE Symp. Information Visualization (INFOVIS) (Poster),* 2002.

[48] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Ann. Meeting Assoc. for Computational Linguistics,* pp. 271-278, 2004.

[49] O.J. Pinon, D.N. Mavris, and E. Garcia, "Harmonizing European and American Aviation Modernization Efforts Through Visual Analytics," *J. Aircraft,* vol. 48, pp. 1482-1494, Sept./Oct. 2011.

[50] P. Pirolli and S. Card, "The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis," *Proc. Int'l Conf. Intelligence Analysis,* May 2005.

[51] M.F. Porter, "An Algorithm for Suffix Stripping," *Program,* vol. 14, no. 3, pp. 130-137, 1980.

[52] L. Ratinov and D. Roth, "Design Challenges and Misconceptions in Named Entity Recognition," *Proc. Conf. Computational Natural Language Learning (CoNLL),* pp. 147-155, 2009.

[53] H. Ruan, C. Anslow, S. Marshall, and J. Noble, "Exploring the Inventor's Paradox: Applying Jigsaw to Software Visualization," *Proc. ACM Fifth Int'l Symp. Software Visualization (SOFTVIS),* pp. 83-92, Oct. 2010.

[54] D.M. Russell, M.J. Stefik, P. Pirolli, and S.K. Card, "The Cost Structure of Sensemaking," *Proc. ACM Conf. Human Factors in Computing Systems (CHI),* pp. 269-276, 1993.

[55] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: Supporting Investigative Analysis through Interactive Visualization," *Information Visualization,* vol. 7, no. 2, pp. 118-132, 2008.

[56] A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," *Proc. Workshop Artificial Intelligence for Web Search (AAAI),* pp. 58-64, 2000.

[57] V. Thai, P.-Y. Rouille, and S. Handschuh, "Visual Abstraction and Ordering in Faceted Browsing of Text Collections," *ACM Trans. Intelligent Systems and Technology,* vol. 3, no. 2, pp. 21:1-21:24, Feb. 2012.

[58] J.J. Thomas and K.A. Cook, *Illuminating the Path.* IEEE CS Press, 2005.

[59] F. van Ham, M. Wattenberg, and F.B. Viégas, "Mapping Text with Phrase Nets," *IEEE Trans. Visualization and Computer Graphics,* vol. 15, no. 6, pp. 1169-1176, Nov./Dec. 2009.

[60] F.B. Viégas, S. Golder, and J. Donath, "Visualizing Email Content: Portraying Relationships from Conversational Histories," *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems (CHI),* pp. 979-988, 2006.

[61] F.B. Viégas, M. Wattenberg, and J. Feinberg, "Participatory Visualization with Wordle," *IEEE Trans. Visualization and Computer Graphics,* vol. 15, no. 6, pp. 1137-1144, Nov./Dec. 2009.

[62] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar, "What's Being Said Near 'Martha'? Exploring name entities in literary text collections," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* pp. 107-114, Oct. 2009.

[63] M. Wattenberg, "Arc Diagrams: Visualizing Structure in Strings," *Proc. IEEE Symp. Information Visualization (INFOVIS),* pp. 110-116, 2002.

[64] M. Wattenberg and F.B. Viégas, "The Word Tree, an Interactive Visual Concordance," *IEEE Trans. Visualization and Computer Graphics,* vol. 14, no. 6, pp. 1221-1228, Nov./Dec. 2008.

[65] C. Weaver, "Cross-Filtered Views for Multidimensional Visual Analysis," *IEEE Trans. Visualization and Computer Graphics,* vol. 16, no. 2, pp. 192-204, Mar. 2010.

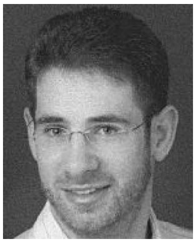
[66] R.W. White, B. Kules, S.M. Drucker, and M.C. Schraefel, "Supporting Exploratory Search," *Comm. ACM,* vol. 49, no. 4, pp. 36-39, Apr. 2006.

[67] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort, "The Sandbox for Analysis: Concepts and Methods," *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems (CHI),* pp. 801-810, Apr. 2006.

[68] H. Zha, "Generic Summarization and Keyphrase Extraction using Mutual Reinforcement Principle and Sentence Clustering," *Proc. ACM 25th Ann. Int'l Conf. Research and Development in Information Retrieval,* pp. 113-120, 2002.

[69] C. Görg, Y. Kang, Z. Liu, and J. Stasko, "Visual Analytics Support for Intelligence Analysis," *Computer,* pp. 30-38, July 2013.

[70] C. Görg, Z. Liu, and J. Stasko, "Reflections on the Evolution of the Jigsaw Visual Analytics System," *Information Visualization,* 2013, to appear.

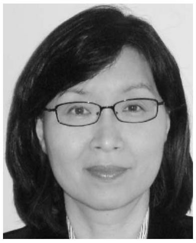


Carsten Görg received the PhD degree in computer science from Saarland University, Germany in 2005. He is an instructor in the Computational Bioscience Program in the University of Colorado Medical School. His research interests include visual analytics and information visualization with a focus on designing, developing, and evaluating visual analytics tools to support the analysis of biological and biomedical data sets. He is a member of the IEEE and the IEEE Computer Society.

Zhicheng Liu received the BS degree in computer science from the National University of Singapore in 2006 and the PhD degree in human-centered computing from the Georgia Institute of Technology in 2012. He is currently a postdoctoral scholar at Stanford University. His current research interests include visualizing big data and developing novel interaction mechanisms in visual analysis.

Jaeyeon Kihm received the BS degree from the Illinois Institute of Technology in 2009, the MS degree from the Georgia Institute of Technology in 2011, and is currently working toward the PhD degree in information science at Cornell University. He is currently developing an energy-efficient user interface system for mobile information appliances.

Jaegul Choo received the BS degree in electrical engineering from Seoul National University, Seoul, Korea in 2001, and the MS degree in electrical engineering from the Georgia Institute of Technology in 2009, where he is currently a research scientist as well as working toward the PhD degree in computational science and engineering. His research interests include visualization, data mining, and machine learning with a particular focus on dimension reduction and clustering methods. He is a student member of the IEEE.



Haesun Park is currently a professor in the School of Computational Science and Engineering and the director of the NSF/DHS FODAVA-Lead (Foundations of Data and Visual Analytics) Center at the Georgia Institute of Technology. She has published more than 150 peer reviewed papers in the areas of numerical algorithms, data analysis, visual analytics, bioinformatics, and parallel computing. She has served on numerous editorial boards including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *SIAM Journal on Matrix Analysis and Applications*, *SIAM Journal on Scientific Computing*, and has served as a conference cochair for the SIAM International Conference on Data Mining in 2008 and 2009. She is a member of the IEEE.

John Stasko received the PhD degree in computer science from Brown University in 1989. He is a professor and an associate chair of the School of Interactive Computing at the Georgia Institute of Technology. His research interests include human-computer interaction with a specific focus on information visualization and visual analytics. He is a senior member of the IEEE and an ACM Distinguished Scientist.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.