

Combining Computational Analyses and Interactive Visualization to Enhance Information Retrieval

Carsten Görg, Jaeyeon Kihm, Jaegul Choo, Zhicheng Liu, Sivasailam Muthiah,
Haesun Park, John Stasko
College of Computing, Georgia Institute of Technology, Atlanta, GA USA

ABSTRACT

Exploratory search and information-seeking support systems attempt to go beyond simple information retrieval and assist people with exploration, investigation, learning and understanding activities on document collections. In this work we integrate several computational text analysis techniques, including document summarization, document similarity, document clustering, and sentiment analysis, within the interactive visualization system Jigsaw in order to provide a flexible and powerful environment for people to examine sets of documents. Our focus is not on cutting edge algorithms for computational analysis but rather on the process of integrating automated analyses with interactive visualizations in a smooth and fluid manner. We illustrate this integration through an example scenario of a consumer examining a collections of car reviews in order to learn more about the car and understand its strengths and weaknesses.

Keywords

exploratory search, information seeking, sense-making, visualization, visual analytics

1. INTRODUCTION

We have been developing new interfaces and systems for information retrieval, in particular, for retrieval of collections of documents with a goal of understanding the many different dimensions and contents of those documents. Sometimes called Exploratory Search [11, 5], Information Seeking Support [6], or Sense-making [4], these processes go beyond the initial retrieval of data by providing environments in which a person can browse, explore, investigate, discover, and learn about the topics, themes, and concepts within the documents.

More specifically, the following situations provide examples of the types of processes we seek to support:

- A police investigator has a collection of case reports, evidence reports, and interview transcripts and seeks to “put the pieces together” to identify the culprits behind a crime.
- 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 wishes to buy a new digital camera but encounters a large variety of possible models to choose from, each of which with supporting documentation and consumer reviews.
- A family learns that their child may have a rare disease and they scour the web for documents and information about the condition.

Our approach combines two main components: automated computational analysis of the documents and interactive visualizations of the documents themselves and of the results of the analysis. Such a combination is described as a *visual analytics* approach [9, 3] and it attempts to leverage 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. The computer excels at complicated analyses of large data collections to determine metrics, correlations, connections, and statistics about the document collection.

Relatively few systems to date, however, have smoothly incorporated both automated computational analysis and interactive visualization while providing a tight coupling between the two. It is more common to encounter systems focused on one of the two capabilities that also add a few elements from the other capability. For instance, computational analysis tools sometimes provide rudimentary user interfaces to access analysis capabilities. Alternatively, interactive visualization systems may provide a few simple techniques such as filtering or statistical analysis of the data.

The system through which we have been exploring this coupling is Jigsaw [8], a tool for helping people explore document collections. Jigsaw is a relatively mature prototype system, and has seen initial use in the field by clients from law enforcement, investigative reporting, fraud detection, and academic research, among others. An initial user study with the system showed its potential in helping investigators and supporting different analysis strategies [2].

Until now, Jigsaw has provided more in the way of interactive visualization support of document exploration. In particular, Jigsaw visualizes connections between entities across documents to help investigators follow trails of information. Recently, we added enhanced computational analysis to the system. Jigsaw now also provides capabilities such as analysis of document similarity, document sentiment, document clusters by theme or content, and document summarization through a few words or sentences.

Our focus has not been about developing innovative new algorithms for computational analysis, however. Instead, we have been exploring methods for smoothly integrating the computational analyses into an interactive visual interface in a seamless manner that would provide a natural and fluid user experience.

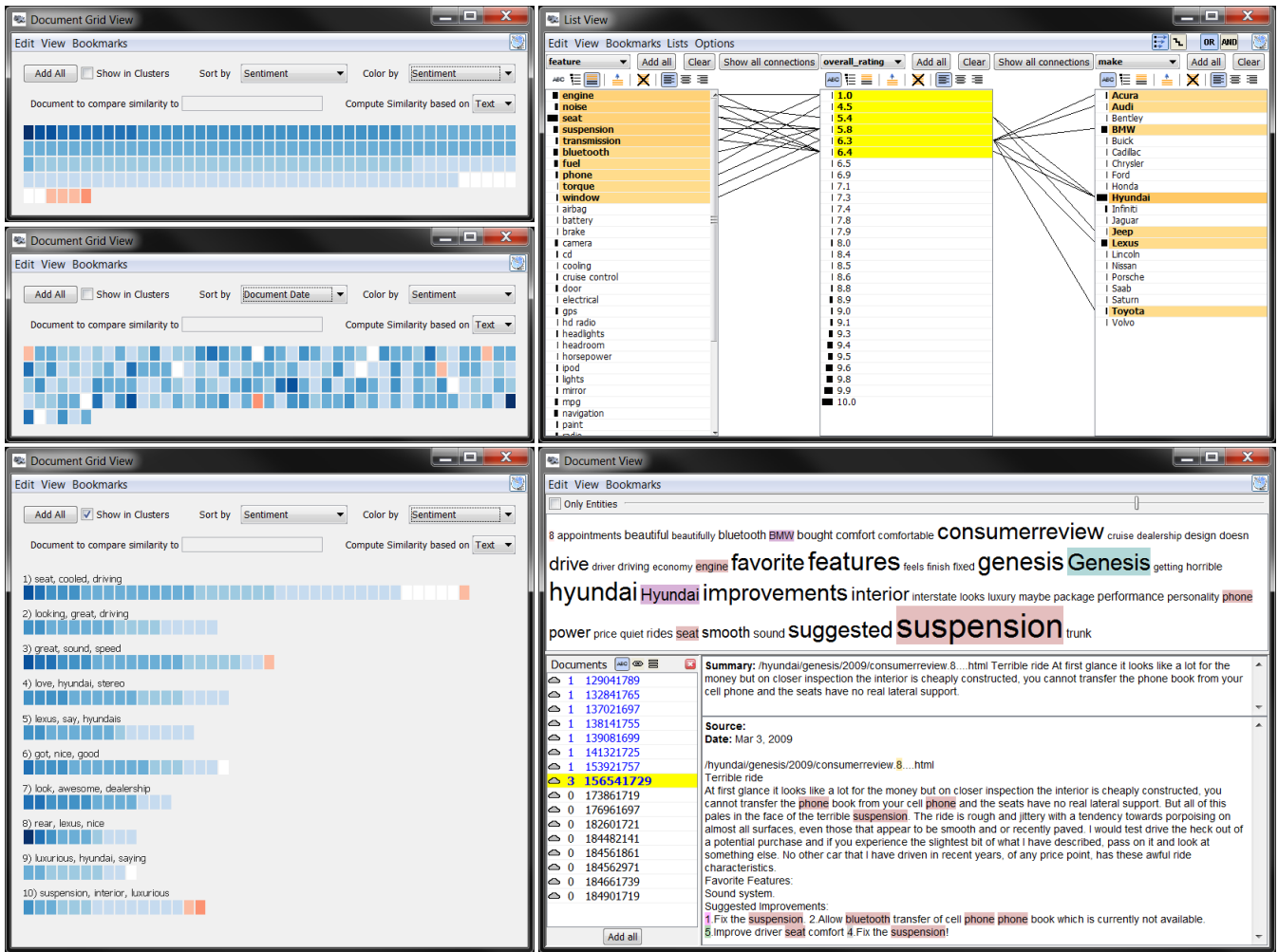


Figure 1: Jigsaw's Document Grid Views, List View, and Document View showing connections in and statistics of car reviews about the 2009 Hyundai Genesis retrieved from the edmunds.com website.

2. AN EXAMPLE INVESTIGATIVE SCENARIO: CAR REVIEWS

Jigsaw is a system for helping analysts with different kinds of investigative and sensemaking scenarios based on textual documents. It is a multi-view system, including a number of different visualizations of the documents in the collection and the entities (e.g., people, places, organizations, etc.) within those documents. Figure 1 shows some of the visualizations. Initially developed for use in intelligence and law enforcement scenarios, more recently Jigsaw has seen increased use in other domains and for many different kinds of document collections. More detail about Jigsaw can be found in [8].

To help illustrate how computational analyses and interactive visualization combine in Jigsaw, we present an example investigative scenario in which an example consumer, Mary, is shopping for a new car. To help her learn about a particular car, the 2009 Hyundai Genesis, that she is considering, Mary examines a document collection consisting of 178 reviews of the car from the edmunds.com website. She could, of course, examine these reviews sequentially from the website in the manner that anyone would do when exploring a topic using a collection of consumer reviews or webpages retrieved from a search engine. That can, however, be slow and not well illuminate the key themes and connections across the reviews.

For illustrating Mary's use of Jigsaw in this scenario, we scraped the 178 reviews from the edmunds.com website and imported them into Jigsaw. Each review is modeled as a document. The main textual content of the review is the text of the document. The document's entities include various rating scores (e.g., exterior design, fuel economy, overall, etc.) that the review author explicitly designated, and other car makes and models mentioned in the review's text. Additionally, we added an entity type "feature" for which we defined about 40 general terms about cars such as seat, trunk, and engine, and we look for mention of those terms in the review text. Figure 1 presents several Jigsaw views from the exploration session that will be used throughout our discussion.

To get an overview of the reviews, Mary begins her investigation by invoking the Document Cluster View (Figure 2) and examining the different key concepts across the reviews. The Cluster View shows each document as a small rectangle and it includes commands to cluster the documents based on either the document text or on the entities connected to a document. Here, Mary chose full document text as the basis for the clustering to achieve the broadest interpretation. Jigsaw then reorganizes the display and positions the documents into clusters based on the analysis. Mary notices clusters around concepts such as the sound system, the ride, fuel economy, and seating.

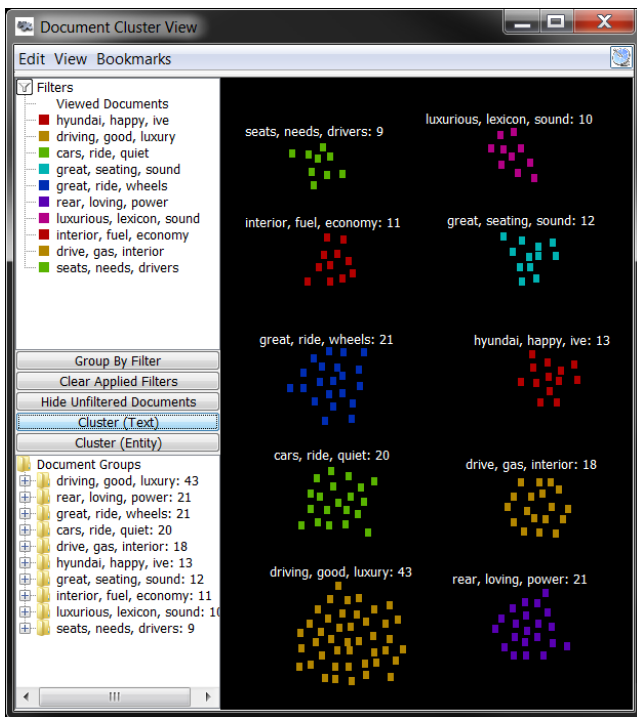


Figure 2: Jigsaw's Document Cluster Views showing clusters of car reviews about the 2009 Hyundai Genesis retrieved from the edmunds.com website.

Next, Mary wants to learn about the subjective opinions of the reviewers, so-called sentiment analysis [1], so she examines the Document Grid View. It displays all the documents as a grid of rectangles where the order and color/shading of the documents in the grid reflect different document metrics. Mary orders and colors the reviews by sentiment calculated by Jigsaw (see Document Grid View at the top left in Figure 1). Positive reviews are colored blue, neutral reviews are colored white, and negative reviews are colored red. Darker shades of blue and red indicate stronger positive and negative sentiment, respectively. At first glance, the reviews for the Genesis appear to be quite positive overall; there are only four negative reviews.

To double check the sentiment, Mary examines in the List View the connections of those four documents to the overall rating given by the reviewer and the car features mentioned by the reviewer. The List View organizes different types of entities into different lists and visually presents connections between entities through orange shading and connecting lines. Two entities are considered to be "connected" if they occur in at least one document together. The List View shows that those four reviews are indeed very negative. The consumers who wrote them assigned overall ratings of 1.0, 4.5, 5.8, and 6.5, respectively, far below the average rating of 9.4. The List View also shows that the features *phone*, *seat*, and *suspension* are most strongly connected to the four negative reviews.

Mary now changes the order of the reviews in the Document Grid View to be sorted by date (see the middle left in Figure 1). The most recent review from 09/28/2009 is the leftmost document in the first row, and the oldest review from 06/26/2008 is the rightmost document in the last row. This view indicates that the earlier reviews were slightly more positive than the more recent reviews. The strong positive reviews (dark blue) are in the lower rows, while most of the the neutral reviews (white) and three of the four neg-

ative reviews (red) are in the upper rows. This might indicate that some issues with the car were not apparent when it came out but were revealed over the course of the first year of use.

To learn more about the car's potential weaknesses, Mary displays features and overall ratings in the List View and selects all ratings with a score below 6.5. The terms *engine*, *noise*, *seat*, *suspension*, and *transmission* appear as the features most connected to the negative reviews (see List View in Figure 1, upper right). To put these results in context, Mary switches back to the Document Grid View and displays the reviews in ten clusters based on the review text as calculated by Jigsaw (see Document Grid View at the bottom left in Figure 1). The clusters are labeled with three descriptive keywords and the documents within each cluster are ordered and colored by their sentiment. Two of the four negative reviews are in cluster 10 mentioning *suspension* as a keyword. Cluster 1, mentioning *seat* as a keyword, also contains one negative and most of the neutral reviews. This suggests that the suspension and the seats may be weaker points of the 2009 Hyundai Genesis. Interestingly here, even though Jigsaw only performs document level sentiment analysis, the system also effectively presents a type of feature-level sentiment simply through its multiple views and brushing across views.

To examine more closely the reviews in cluster 10 containing two of the four negative reviews, Mary displays these reviews in the Document View (see Document View in Figure 1, lower right). The view shows a word cloud (at the top) of the loaded documents that helps the viewer to quickly understand the main themes and concepts within the documents by presenting the most frequent words across the documents. The number of words shown can be adjusted interactively with the slider above the cloud. Here, the word cloud shows that the suspension is indeed mentioned frequently in these reviews. Browsing through the reviews and reading their summaries reveals that the suspension is often described in a negative context, as shown in the selected review in the figure. To help with fast triage of a large set of documents, the Document View provides a one sentence summary (most significant sentence) of the currently displayed document above its full text. This one sentence summary of a document is available in all other Jigsaw views as well and can be displayed through a tooltip wherever a document is presented.

To learn more about the ride quality of the car, Mary displays the Word Tree [10] View for "ride" (Figure 3). A Word Tree shows all occurrences of a word or phrase from the reviews in the context of the words that follow it. The user can navigate through the tree by clicking on its branches. The Word Tree in Figure 3 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 "most impressive".

Not shown in this scenario is the document similarity computation and display within Jigsaw. Document similarity can be measured relative to complete document text or just to the entities connected to a document. These different similarity measures are of particular interest for semi-structured document collections, such as publications, in which metadata-related entities (e.g. authors or conferences) are not mentioned in the actual document text. The Document Grid View (top left in Figure 1) can provide an overview of all the documents' similarity (relative to a selected document) via the order and color of the documents in the grid representation. In all other views, the five most similar documents can be retrieved with a right mouse button command on a document representation.

Jigsaw also includes a Calendar View that presents documents and entities within the context of a calendar so that an investigator can see patterns, trends, and temporal orderings and a Graph

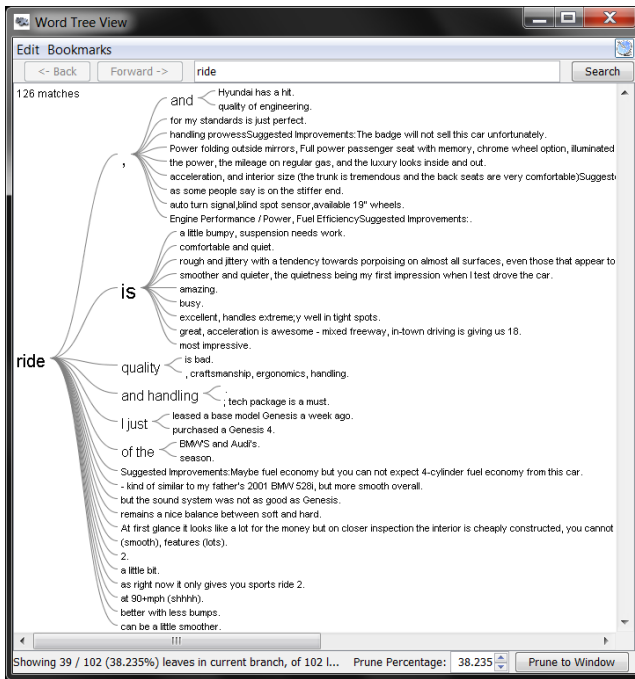


Figure 3: Jigsaw’s Word Tree View showing sentences using the word “ride” in car reviews about the 2009 Hyundai Genesis retrieved from the edmunds.com website.

View that shows a node-link network representation of documents and the entities within them. Investigators can choose two or more entities within the Graph View and Jigsaw will compute “related” entities, that is, entities in the local neighborhood of the selected ones, and it will show the shortest paths between all these entities.

All the Jigsaw views discussed above primarily assist investigators with “information foraging” activities, the first half of the investigative process model proposed by Pirolli and Card [7]. In this respect, we believe that Jigsaw is most useful in helping people determine which document(s) they should read next. To assist investigators with “sense-making” activities, the second half of the Pirolli-Card model, we have recently added a new window called the Tablet to Jigsaw. The Tablet functions much like an electronic notebook in which an investigator can drop in entities, documents, snapshots of views, or manually-generated notes and content. These items within the Tablet also can be connected with edges to help create structures like social networks or the items can be positioned along timelines. Essentially, the Tablet helps investigators to organize their thoughts, gather evidence, take notes, and develop ideas.

3. 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.

We speculate that simply performing rich computational analysis of the documents may not be sufficient – The analyst inevitably will think of some question or perspective about the documents that is not illuminated by the computational analysis. We also speculate that interactive visualization of the documents itself also may not be sufficient – As the size of the document collection grows, inter-

actively exploring the individual characteristics of each document simply may take too much time. Thus, through the combination of these two technologies, so-called visual analytics, we can develop systems that provide powerful exploratory, investigative capabilities that were unavailable before.

In this research, we have illustrated methods for doing just that: integrating automated computational analysis with interactive visualization for text- and document-based exploration, investigation, and understanding. We integrated a suite of textual analysis techniques into the Jigsaw system, showing how the analysis results can be combined with existing and new visualizations. Further, we provided an example analysis scenario that shows both the methodology and the utility of these new capabilities. Although the computational analysis techniques are not new, we have integrated them with interactive visualization in new manners to provide a system that we feel provides innovative and powerful exploratory search and sense-making capabilities.

4. ACKNOWLEDGMENTS

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

5. REFERENCES

- [1] M. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Workshop on Sentiment and Subjectivity in Text*, pages 23–30, 2006.
- [2] Y. Kang, C. Görg, and J. Stasko. The evaluation of visual analytics systems for investigative analysis: Deriving design principles from a case study. In *IEEE VAST*, pages 139–146, Oct. 2009.
- [3] 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*, pages 154–175, 2008.
- [4] G. Klein, B. Moon, and R. Hoffman. Making sense of sensemaking 1: alternative perspectives. *IEEE Intelligent Systems*, 21:70–73, 2006.
- [5] G. Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, Apr. 2006.
- [6] G. Marchionini and R. W. White. Information-seeking support systems. *IEEE Computer*, 42(3):30–32, Mar. 2009.
- [7] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *International Conference on Intelligence Analysis*, May 2005.
- [8] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.
- [9] J. J. Thomas and K. A. Cook. *Illuminating the Path*. IEEE Computer Society, 2005.
- [10] M. Wattenberg and F. B. Viégas. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, 2008.
- [11] R. W. White, B. Kules, S. M. Drucker, and M. C. Schraefel. Supporting exploratory search. *Communications of the ACM*, 49(4):36–39, Apr. 2006.