Structuring Collections with Scatter/Gather Extensions

Omar Alonso
Department of Computer Science
University of California, Davis
oralonso@ucdavis.edu

Justin Talbot
Department of Computer Science
Stanford University
jtalbot@stanford.edu

ABSTRACT
A major component of sense-making is organizing—grouping, labeling, and summarizing—the data at hand in order to form a useful mental model, a necessary precursor to identifying missing information and to reasoning about the data. Previous work has shown the Scatter/Gather model to be useful in exploratory activities that occur when users encounter unknown document collections. However, the topic structure communicated by Scatter/Gather is closely tied to the behavior of the underlying clustering algorithm; this structure may not reflect the mental model most applicable to the information need. In this paper we describe the initial design of a mixed-initiative information structuring tool that leverages aspects of the well-studied Scatter/Gather model but permits the user to impose their own desired structure when necessary.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Clustering. H.5.2 [**User Interfaces**]: Prototyping.

General Terms: Design, experimentation.

Keywords
Scatter/Gather, sense-making, clustering, text summarization.

1. INTRODUCTION
Given the enormous growth rate of information available from the Web and other sources, many authors have argued that the primary challenge for users changes from *finding* information to *making sense* of it. A typical information access tool provides a simple interaction model where the user enters a query and the engine returns results composed of title, snippet, and URL. This is sufficient for retrieving specific information, but not for building a holistic understanding of the information space. In the case of clustering search engines or Scatter/Gather [1], the system automatically imposes additional hierarchical structure on the search results. In Scatter/Gather, in particular, this additional structure has been shown to aid the development of a user's mental model [3]. However, this structure is directly determined by the underlying clustering algorithm. We believe that this is not sufficient.

In this paper, we consider the general scenario of a user given a collection of documents which must be collectively understood and navigated. Specific examples of such scenarios are: a finance analyst identifying emerging trends from a collection of news stories or an intelligence analyst assessing the consistency of newly discovered information with an existing corpus of accepted knowledge.

To effectively work with large document collections, the user must create a mental model of the contents of and relationships between elements of the collection. This organization is often task specific and it may encode significant domain knowledge. For this reason we believe it is helpful to explore tools that help users impose their own structure on the document collection while still enabling well-understood exploration techniques.

Our approach is an extension to the Scatter/Gather model in which the hierarchy of clusters is explicitly exposed to the user. We provide a small set of tools which allow the user to directly change the generated hierarchy. We further allow the user to add notes to the hierarchy and to incorporate automatically generated summaries into the hierarchy.

The main contributions of our research are:

- the extension of Scatter/Gather to explicitly reveal the hierarchical structure being created,
- a small set of tools which allow users to modify the created structure,
- and, support for including user-created notes and summaries directly in the hierarchy.

2. EXTENSIONS TO SCATTER/GATHER
Scatter/Gather is a technique for browsing a collection of documents. Static and runtime clustering produces a small set of document clusters. The user navigates through the collection selecting (gathering) a subset of the clusters, based on keyword summaries. Then the user can regenerate the cluster set (scattering) using only those documents in the selected clusters [1]. In this way the user can quickly narrow in on an interesting set of documents. Later research found that the Scatter/Gather model is effective in helping users build a "more coherent conceptual image of a text collection" [2], [3]. This fact makes Scatter/Gather a good basis for an information structuring tool.

The main extension is the addition of exposed hierarchical tree view that is incrementally built as the user explores the collections. The root node represents the entire document collection. Other nodes represent other clusters of documents produced by the Scatter/Gather operations.

When a single node is selected, the user can scatter the node's documents. This operation inserts new child nodes into the tree containing the new document clusters. (Scattering can also be done when multiple nodes are selected, in which case the selected nodes are automatically combined before the scattering occurs.)

SIGIR'08, July 20-24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

In the spirit of a faceted search interface, we provide multiple dimensions along which the user can scatter documents (e.g. standard term frequency clustering, grouping by publication date, by author, etc.). This permits the user to quickly explore common organizations of the documents.

By selecting multiple nodes in the tree and combining them, a user can gather the corresponding clusters. Unlike standard Scatter/Gather, the user is free to combine clusters from anywhere in the tree. This permits the user to correct incorrect automatic clustering where conceptual groups are split across multiple clusters.

Additionally, the user can use direct manipulation to move nodes around in the tree. This implicitly moves the corresponding documents around as well, updating the ancestors of the affected nodes. Finally, we provide a mechanism to discard nodes. This permits the user to hide any document clusters that may be irrelevant to the current information task. Discarded nodes can be retrieved at any time.

The use of the tree view introduces the challenge of selecting appropriate labels for the nodes in the tree. Space in the tree does not permit the use of extended keyword collections (as in [5]). We adapt a two-pronged approach. First, in a separate view we present extended summaries of just the currently selected nodes. Our tool generates a cluster summary at runtime that summarizes the main content in a short bullet list. Second, in the tree we use some simple heuristics (based on the scattering dimension) to generate a potentially meaningful label and then permit users to edit the labels as they desire. This allows users to capture the structure that is important to them as well as allowing the tree to be used as an in-context note-taking system. We encourage this use by allowing the user to copy portions of document summaries or titles into tree labels. Depending on the user's task, the resulting tree can end up looking like a hierarchical textual outline of the document collection. Figure 1 shows a schematic view of our proposed model.

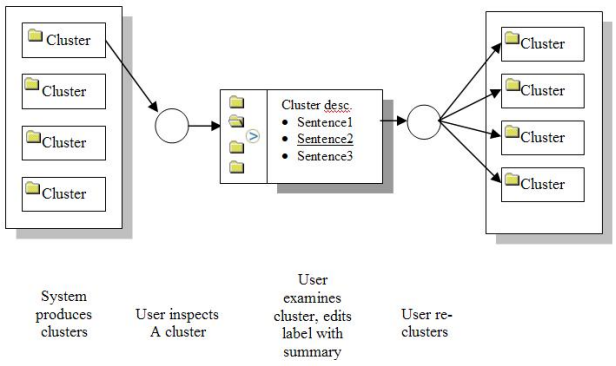

Figure 1: User interaction with Scatter/Gather with extensions

3. PROTOTYPE

We have implemented a prototype, called Napoli, of the organizational interaction extensions presented in this paper. Figure 2 shows Napoli used on a collection of papers from the SIGGRAPH conference. The left pane shows the current hierarchical organization of the document collection as captured by the tree view. The right pane shows the cluster summary for the currently selected node.

The user initially scattered the entire collection by publication date. Automatic labels containing the year ranges (since they were scattered by date) were generated for the clusters. The user then scattered the 1985-1989 cluster by topic and then scattered the first child cluster by topic again (k-means clustering of *tf-idf* vectors). The user has added a note "Papers about rendering" to the first child cluster capturing structure important to the user's information task.

The cluster summary contains the most common words and number of documents. The bulleted list is an automatic summary of the most important sentences in the cluster computed using a TextRank-based implementation [4].

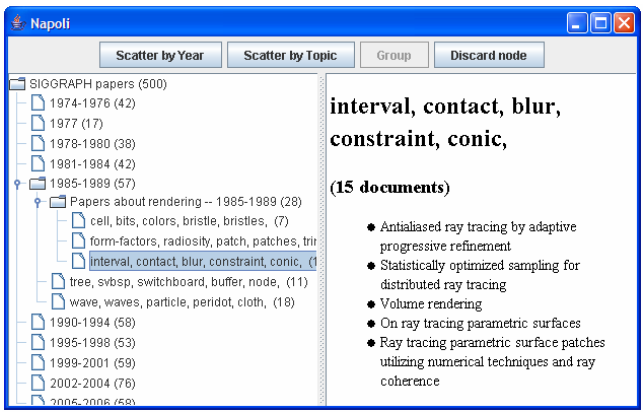

Figure 2: Exploring the SIGGRAPH collection.

4. CONCLUSIONS

We have described a number of interaction extensions to Scatter/Gather that provide functionality to capture user-centric, meaningful structure in document collections. Future work includes a user evaluation of the system.

5. REFERENCES

[1] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections". *Proc. Of 15th SIGIR* (1992).

[2] P. Pirolli. *Information Foraging Theory*. Oxford University Press (2007).

[3] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. "Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection", *Proc of SIGCHI* (1996).

[4] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts", *Proc. of EMNLP* (2004).

[5] M. Hearst, D. Karger, and J. Pedersen. "Scatter/Gather as a Tool for the Navigation of Retrieval Results", *Proc. Of AAAI* (1995).