Revealing Collection Structure through Information Access Interfaces

Marti A. Hearst Jan O. Pedersen

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
(415) 812-4742
{hearst,pedersen }@parc.xerox.com

Information Access research at Xerox PARC focuses on amplifying the users' cognitive abilities, rather than trying to completely automate them. This framework emphasizes the participation of the user in a cycle of query formulation, presentation of results, followed by query reformulation, and so on. This framework is intended to help the user iteratively refine a vaguely understood information need. Since the focus is on query repair, the information presented is typically not document descriptions, but rather intermediate information that indicates relationships between the query and the retrieved documents. We have developed information access tools intended to supply some of this functionality, and describe two of these here.

As an illustration, suppose a user is interested in medical diagnosis software. Assume that initially the user has available a large, unfamiliar information source. In our example, this source is the 2.2 Gigabyte TIPSTER text collection [Harman, 1993]. Because the collection is unfamiliar, the user will be unsure whether it contains relevant information, and if so, how to access it.

To address this situation, we have developed a browsing method, called *Scatter/Gather* [Cutting et a/., 1992; 1993], that allows a user to rapidly assess the general contents of a very large collection by scanning through a dynamic, hierarchical representation that is motivated by a table-of-contents metaphor. Initially the system automatically *scatters,* or clusters, the collection into a small number of document groups, and presents short summaries of the groups to the user. These summaries consist of two types of information: topical titles (titles of documents close to the cluster centroid) and typical terms (terms of importance in the cluster). Based on these summaries, the user selects one or more of the groups for further study. The selected groups are *gathered)* or unioned, together to form a subcollection. The system then applies clustering again to scatter the new subcollection into a small number of document groups, which are again presented to the user. With each successive iteration the groups become smaller, and therefore more detailed. The user may, at any time, switch to a more focused search method. Figure 1 shows a portion of the top level clusters on the TIPSTER collection.

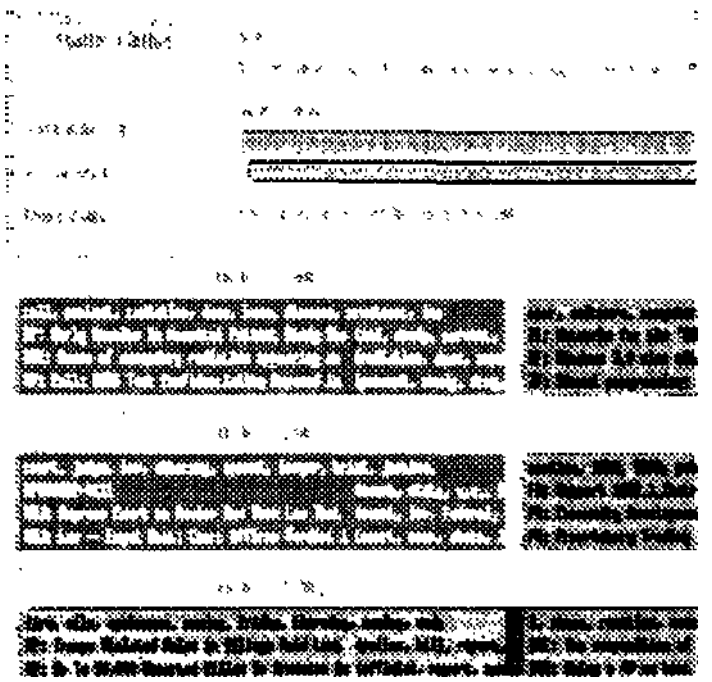

Figure 1: A portion of a top-level view of the Scatter/Gather algorithm over the TIPSTER corpus.

By browsing the collection in this manner, the user obtains an idea about the technical contents of the corpus, and can choose whether or not to further explore here or try another text collection. From the titles and terms retrieved, it becomes apparent that the collection contains commercially oriented discussions of technology, rather than predominantly academic ones. From this overview information, the user can conclude that this is indeed a promising collection for the user's information need.

Once a premising collection has been identified, the user can issue a search. In a typical information retrieval system, documents satifying the query are returned and are rank-ordered according to some function of the number of hits for each term [Salton, 1988]. But this kind of ranking is opaque to the user; it is not clear how well each term is represented in the retrieved documents.

To address these issues, the *TileBars* interface [Hearst, 1995] allows the user to make informed decisions

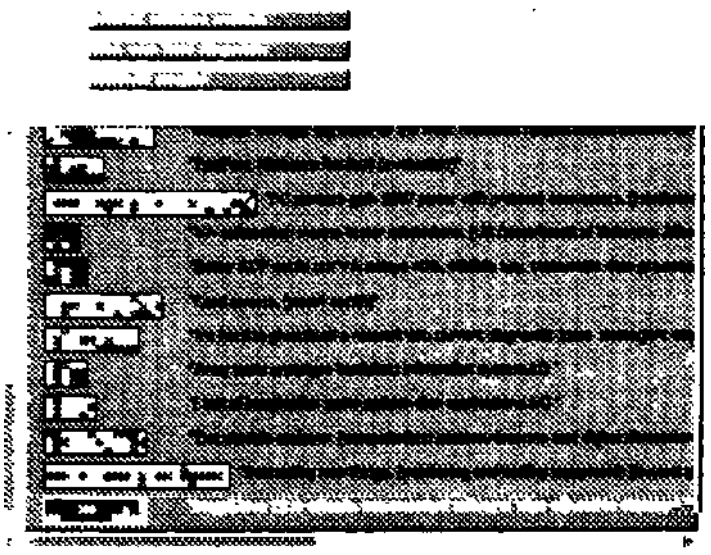

Figure 2: The TileBar Display on a query about automated systems for query diagnosis.

about which documents and which passages of those documents to view, based on the distributional behavior of the query terms in the documents. The goal is to simultaneously and compactly indicate (i) the relative length of the document, (ii) the frequency of the term sets in the document, and (iii) the distribution of the term sets with respect to the document and to each other. Each document is partitioned in advance into a set of multi-paragraph subtopical segments using an algorithm called *TextTiling* [Hearst, 1994].

Figure 2 shows an example run on a query about automated systems for medical diagnosis, run over the ZIFF portion of the TIPSTER collection. Each large rectangle indicates a document, and each square within the document represents a coherent text segment. The darker the segment, the more frequent the term (white indicates 0, black indicates 8 or more hits, the frequencies of all the terms within a term set are added together). The top row of each rectangle correspond to the hits for Term Set 1, the middle row to hits of Term Set 2, and the bottom row to hits of Term Set 3. The first column of each rectangle corresponds to the first segment of the document, the second column to the second segment, and so on.

The TileBars representation allows the user to sort the retrieved documents according to which aspects of the query are most important. For example, in the figure the query is formulated as: (patient OR *medicine* OR *medical)* AND (test OR *scan* OR *cure* OR *diagnosis)* AND *(software* OR *program).* This formulation allows the interface to indicate the role played by each conceptual part of the query: the medical terms, the diagnosis terms, and the software terms. In Figure 2, the user has indicated that the diagnosis aspect of the query must be strongly present in the retrieved documents, by setting the minimum term distribution percentage to 30% for the second termset. The document whose title begins *"VA automation means faster admissions"* is quite likely to be relevant to the query, and has all three term sets well-distributed throughout. By contrast, the document whose title begins *"It's hard to ghosibust a network ..."* is about computer-aided diagnosis, but has only a passing reference to *medical* diagnosis, as can be seen by the graphical representation. If the user decides that medical terms should be better represented, the constraint on this term set can be adjusted accordingly.

Note that a system that simply ranks the documents does not make these kinds of distinctions available to the user. The graphical representation allows the users to rapidly assess the structure of the retrieved documents with respect to the query, to better aid their decisions about which documents to view, or how to refine the query.

References

[Cutting *et ai,* 1992] Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/Gather: A cluster-based approach to browsing targe document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference,* pages 318-329, Copenhagen, Denmark, 1992.

[Cutting *et al.,* 1993] Douglass R. Cutting, David Karger, and Jan Pedersen. Constant interaction-time Scatter/Gather browsing of very large document collections. In *Proceedings of the 16ih Annual International ACM/SIGIR Conference,* pages 126135, Pittsburgh, PA, 1993.

[Harman, 1993] Donna Harman. Overview of the first Text REtrieval Conference. In *Proceedings of the 16th Annual International ACM/SIGIR Conference,* pages 36-48, Pittsburgh, PA, 1993.

[Hearst, 1994] Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics,* June 1994.

[Hearst, 1995] Marti A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems,* Denver, CO, May 1995. ACM.

[Salton, 1988] Gerard Salton. *Automatic text processing : the transformation, analysis, and retrieval of information by computer.* Addison-Wesley, Reading, MA, 1988.