A World Wide Web-based HCI-library Designed for Interaction Studies

Ketil Perstrup, Erik Frøkjær, Maria Konstantinovitz,
Thorbjørn Konstantinovitz, Flemming S. Sørensen, Jytte Varming

Department of Computing, Copenhagen University
Universitetsparken 1, 2100 København Ø, Denmark
Tel: +45 35 32 14 00, Fax: +45 35 32 14 01, E-mail: ketil@diku.dk

Abstract. The World Wide Web has the potential for making scientific information widely available even to people without access to scientific libraries, but using the World Wide Web for this is hard in practice. Possible causes for this are that there is no central repository for scientific papers, that the papers are often not indexed by the World Wide Web search engines, and when they are, novices have trouble using the services. We have developed HCILIB as an interface to a collection of scientific articles on Human-Computer Interaction available on the World Wide Web. HCILIB uses a scatter/gather inspired technique to display a browsable structure for the collection integrated with Boolean queries. It has facilities for searchers to restrict their view of the collection to the parts they consider interesting and reorganize these to display a personal classification of documents. This allows us to investigate usage patterns and differences in them for such a library including field studies of the interaction.

Keywords. Information search and retrieval, information interfaces and presentation, query formulation and search process, digital libraries, World Wide Web, scatter/gather, browsing.

1. INTRODUCTION

Novices often find that it is difficult to use the World Wide Web to locate scientific information. There is no central place to look for information on a specific subject, the search services don't index everything and the results from search services are hard to understand and use. This is partly because the current tools do not adequately support a broad selection of information needs and search behaviors. Meadow [Meadow 92] distinguishes among four categories of information needs and search behaviors: 1. Specific item search where information is sought about a specific item, such as the author of a book, 2. Specific information search where a specific piece of information is sought, 3. General information search where general information is sought on a subject, and 4. Exploration of a collection where the purpose is to survey the contents of the collection.

The query tools normally used for information retrieval are not good at supporting exploration. In libraries this is not a problem, because the shelves of libraries are excellent for exploration, but in purely digital collections of documents, such as the World Wide Web, there are no shelves to explore so it is hard to satisfy category three and four information needs. This may be an important omission as exploration through browsing is an important supplement to querying in small collections where it often yields results fast and efficiently [Hertzum 96].

This paper describes the design of the user interface of HCILIB, a searchable and browsable collection of documents on Human-Computer Interaction available on the World Wide Web. We have implemented the first version and are evaluating feedback from heuristic evaluation and thinking-aloud experiments and creating a new version based on this.

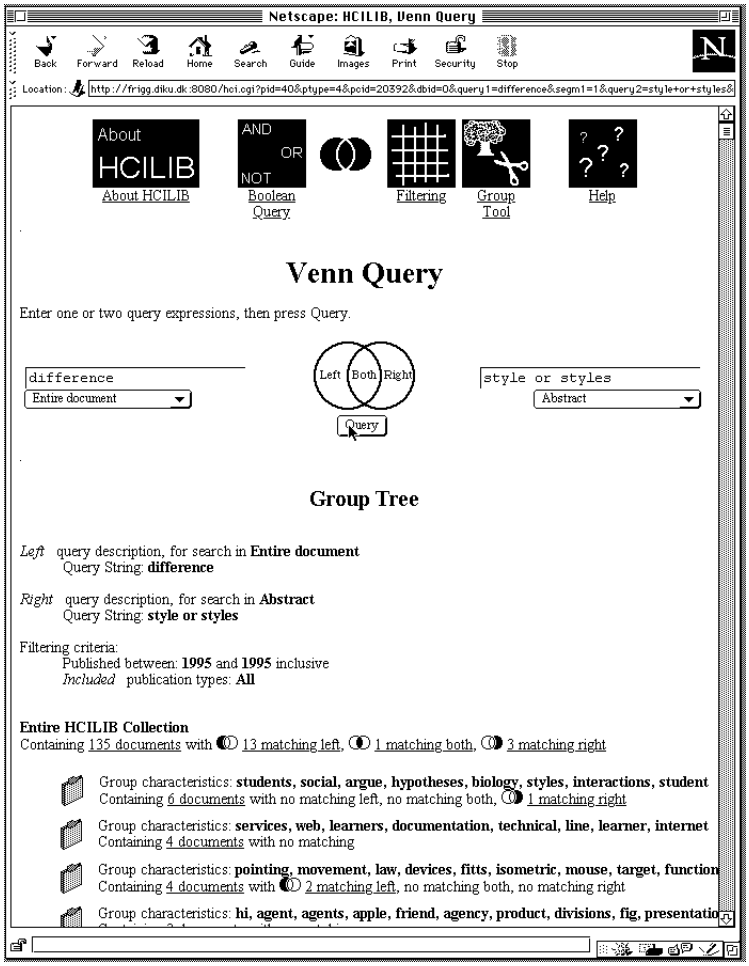

Figure 1: The Venn Query tool is used to search abstracts for *style or styles* and full texts for *difference*. A filter limits the search to documents published in *1995*. Clicking the folder symbols expands a group, the underlined link leads to lists of documents. The figure shows the results of a query to a small test database.

2. THE INTERFACE OF HCILIB

The interface of HCILIB combines search facilities based on Venn diagrams with browsing of an automatically generated tree generated from a *scatter/gather* process [Pirolli 96]. Figure 1 shows the Venn Query tool. The searcher enters queries in the top half and the bottom half displays the last query and the structure together with how many documents match each part of the query in each part of the structure. The searcher can specify which *semantic component* in the documents are queried, i.e., it is possible to restrict the query to match only the author's name, title, abstract, table of contents, references or the full text. A filter tool (not shown here) allows the searcher to restrict queries to publications published in a specific time period and to exclude certain types of publications, such as technical reports or master theses.

The searcher can examine a group closer, if desired, by clicking the folder icon next to the group which expands the group and shows the subgroups (or contracts it if expanded). The display also contains links to the contents of a group and the list of documents in the group matching the current queries.

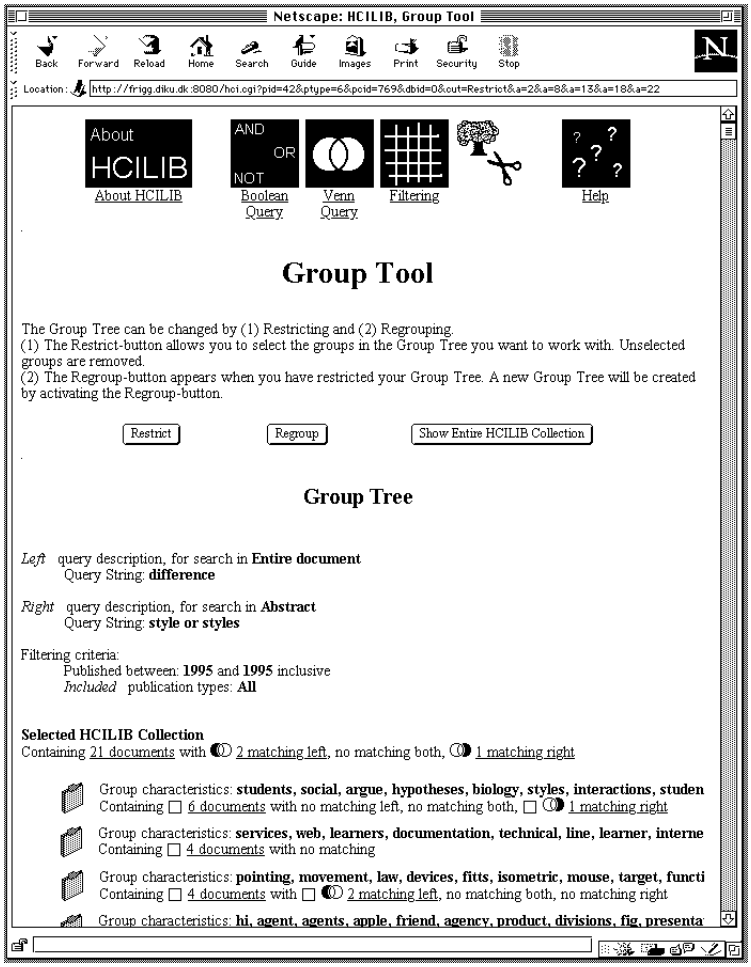

Figure 2: The group tool. Here the searcher can select the groups and query results to keep. The *Restrict* button removes groups and results from queries that are not selected from the display. The *Regroup* button creates a new set of groups from the selected parts. The searcher can include the whole collection by selecting *Show Entire HCILIB Collection.*

The group tool (Figure 2) allows the searcher to restrict the display to only show the parts of the collection that seem interesting. This tool allows the searcher to select groups or query results that seem promising for his or her need and remove the rest of the documents from the display. The group tool also allows the searcher to perform a scatter/gather operation on documents in selected groups and selected query results thereby creating a new tree containing only those documents.

Whenever the searcher submits a new query, filter or grouping, the display is updated to show the results of the new combination. The searcher can select another tool from the toolbar and use the new tool. This allows the searcher to explore the collection since it is always possible to undo the effects of actions by going back to one of the previous pages.

3. MAIN DESIGN IDEAS

HCILIB is an experimental tool designed to study interaction in a searchable and browsable collection of documents. The collection contains documents about HCI and Software Engineering. With HCILIB we want to investigate how different search tools are used and in par-

ticular which tools help satisfy which types of information needs as well as investigate whether the use of the tools varies between individuals or over time. We also want to investigate whether HCILIB is used differently from traditional information retrieval tools.

To aid in exploring the information in the collection, we have integrated a scatter/gather inspired browsing technique with query tools. In contrast to the way scatter/gather has been used previously [Cutting 92][Pirolli 96] we use scatter/gather to present an overview of the document collection or parts of it, instead of an alternative to query tools or a way to organize query results. This use of scatter/gather makes it possible to present each searcher with a personal classification of the parts of the collection that are interesting to him or her. In addition a tight integration of scatter/gather with query tools allows querying and exploration tools to supplement each other.

Lately others have emphasized the need to integrate classification structures in the display of query results to aid searchers to survey large sets of documents retrieved in response to queries and the use of classification hierarchies for browsing [Hearst 97]. Our work is unique in integrating query tools with a scatter/gather based classification, and by using scatter/gather to create a personal view of parts of the collection.

3.1 Creating Structure with Scatter/Gather

We expect that browsing will be an important supplement to querying, in particular as an aid in exploring the collection. Some experimental evidence suggests that scatter/gather aids in surveying the contents of a document collection [Pirolli 96], so to create a structure for browsing we implement scatter/gather as described by Cutting et al. in [Cutting 92]. Scatter/gather creates a set of groups through the text clustering algorithm described so that each group contains documents with a similar distribution of words frequencies. We have extended this algorithm to create a tree instead of a set of groups as described in [van Rijsbergen 79]. This tree arranges the documents in the tree so documents with high similarity are placed together in groups near the leaves in the tree and documents in different top-level groups are very dissimilar. The use of a tree allows us to present searchers with a hierarchical overview of the collection, so the searcher can examine groups more closely, if desired. In order to let the searcher judge what is in each group, each group is presented to the searcher with a set of words. These words are selected from the documents in each group as the words which occur most frequently in the documents but is not on a list of commonly used words (so-called “stop words”).

Initially the searcher is presented with the result of a scatter/gather operation on the whole collection, but the searcher can select groups or query results to include in a scatter/gather operation and create a new set of groups. This allows searchers to restrict themselves to customized views of those parts of the collection that currently have their interest.

3.2 Integration of Browsing and Querying

To assist in locating documents that are relevant to the searcher’s information needs HCILIB displays the results of queries in the scatter/gather tree. We place number of matching documents next to each group and subgroup and allow direct inspection of both all matching documents and matching documents in a group or subgroup.

This integration assists the searcher in locating the groups that may be relevant to his or her information needs, thereby using the query tools to aid in surveying the collection. In addition displaying the results of queries in the tree created by scatter/gather creates a visual representation of the information used in some automatic text retrieval to enhance effectiveness where the information displayed in the tree is used to automatically adjust queries based on the searcher's evaluation of the relevance of sample documents [Salton 86].

The results from queries can be used directly to select documents for inclusion in his or her personal view as the searcher can select and deselect the results from queries in each group for inclusion in the display independently of the group.

3.3 Search techniques

To support querying we have, in addition to Boolean search, implemented a search tool inspired by Venn diagrams [Michard 82]. This tool allows the searcher to enter two queries and see where matches occur for each of the queries and in which combinations. We display relationship between two query terms, as searchers rarely use more than two query terms, even when more can be used [Hertzum 96]. Showing the distribution for the queries independently allows the searcher to understand why a query results in very few or too many matching documents. Other efforts to solve this problem have been explored with Tilebars [Rao 95], Winquery's stacked histograms [Shneiderman 97] and the InfoCrystal [Spoerri 94].

We also include the possibility to distinguish between different semantic components of the texts to reflect the way we use different semantic components differently in everyday life. Titles convey the subject, and we often use the abstract or the table of contents to quickly gain a deeper idea about the content of a document. We want to discover whether this kind of information is usable by searchers, and whether different components are used differently.

4. EXPERIMENTAL USES FOR HCILIB

To make HCILIB widely available on the World Wide Web we have designed the interface for HCILIB in HTML. The HTML Forms interface presents some challenges for the designer compared to more traditional graphical user interfaces, but we have chosen HTML because the World Wide Web presents a unique environment to gather experimental data and to observe the use of a search tool, also outside the laboratory. With HCILIB we can, partly through field studies, explore the interaction process for the system over time by analyzing the usage patterns of the different tools in HCILIB as well as usage patterns for individual searchers. We expect to uncover differences in behavior between searches as suggested by the identification of different information needs described by Meadow.

We want to investigate whether browsing and scatter/gather techniques are usable as an aid for retrieving information—in particular whether it increases the searcher's satisfaction with the system, the searcher's performance in retrieving information and his or her confidence in the results of a search for information. In addition we want to see if scatter/gather will expose structures in document collections and whether it will be used for personal views of them.

We also want to discover we can detect differences in the way the different semantic components of documents are used in queries and whether the option of querying specific semantic components enhances the searcher's retrieval effectiveness.

Finally we hope that a central repository for HCI literature will enable researchers to reach a wider audience and make it easier to locate HCI literature on the World Wide Web. HCILIB will be available late in November at the URL “http://www.diku.dk/infosys/hcilib”

REFERENCES

[Cutting 92] D. R. Cutting, J. O. Pedersen, D. Karger, J. W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, in *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, 1992, pp 318-329.

[Hearst 97] M. A. Hearst, C. Karadi, Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results Using a Large Category Hierarchy, in *Proceedings of the 20th Annual International ACM SIGIR on Research and Development in Information Retrieval*, ACM Press, New York, 1997, pp 246-255.

[Hertzum 96] M. Hertzum, E. Frøkjær, Browsing and Querying in Online Documentation: A Study of User Interfaces and the Interaction Process, *ACM Transactions on Computer-Human Interaction*, 3(2), 1996, pp 136-161.

[Meadow 92] C. T. Meadow, *Text Information Retrieval Systems*, Academic Press, California, 1992, pp. 242-252.

[Michard 82] A. Michard, Graphical Presentation of Boolean Expressions in a Database Query Language: Design Notes and an Ergonomic Evaluation, *Behaviour and Information Technology*, 1, 1982, pp 279-288.

[Pirolli 96] P. Pirolli, P. Schank, M. Hearst, C. Diehl, Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Document Collection, in *Proceedings of CHI 96 Human Factors in Computing System*, ACM Press, New York, 1996, pp 213-220.

[Rao 95] R. Rao, J. O. Pedersen, M. A. Hearst, J. D. Mackinlay, S. K. Card, L. Masinter, P Halvorsen, G. G. Robertson, Rich Interaction in the Digital Library, *CACM*, 38(4), 1995, pp 29-39.

[van Rijsbergen 79] C. J. van Rijsbergen, *Information Retrieval*, Butterworths, London, 1979, pp 23-47.

[Salton 86] G. Salton, Another Look at Automatic Text-retrieval Systems, *CACM*, 29(7), 1986, pp 648-656.

[Shneiderman 97] B. Shneiderman, D. Byrd, W. B. Croft, Clarifying Search. A User Interface Framework for Text Searches, *D-Lib Magazine*, January 1997.

[Spoerri 94] A. Spoerri, InfoCrystal: Integrating Exact and Partial Matching Approaches through Visualization, in *Proc. of RIAO’94. Intelligent Multimedia Information Retrieval Systems and Management*. Rockefeller University, New York, 1994, pp 678-696.