A Graphic Interface for User Directed Clustering of Retrieved Documents

Morris Hirsch, MS, James Allan, PhD
Center for Intelligent Information Retrieval
University of Massachusetts at Amherst
{hirsch, allan}@cs.umass.edu

Introduction

Given a query that has returned hundreds of documents, all presumably "about" the query in one way or another, most retrieval systems display a ranked list of titles, and perhaps some sample text from each document. The user must read down this list until they find what they are looking for, or their patience gives out. What better tools can we provide to make sense of this list?

One well-known recent advance is the **Live Topics** feature of the Alta Vista web search service. Live Topics is activated whenever a user query returns too many matches, and suggests terms for an improved query, based on term cooccurrences in indexed documents.

Our approach is to analyze the set of retrieved documents for a list of **important terms,** which are found, in varying combinations, in the documents. By selecting some of these terms, and thus the documents that contain them, a user may partition the retrieval set.

Our primary goal is to help the user sift the documents they want from those they do not. A useful byproduct of the analysis may be lists of candidate terms for a new query, that would improve recall, precision, or both.

We have built a Java-based World Wide Web interface to the Inquery information retrieval system, using a visualization method called a Starfield, based on work by Ben Shneiderman.

Try it with

Medline (subset, approximately 140,000 citations)
http://ciir.cs.umass.edu/cgi-bin/hirsch/3.0.2/java_query
The Wall Street Journal (1987)
http://toowoomba.cs.umass.edu/cgi-bin/hirsch/3.0.2/java_query/usr/mel/data4/inquery/3.0.2/data/w
TREC Volumes 1 2 3.
http://toowoomba.cs.umass.edu/cgi-bin/hirsch/3.0.2/java_query/usr/mel/data4/inquery/3.0.2/data/tr

Please Note! A Java enabled browser is required. This currently works for Netscape on all Unix / X Window that we have tried, as well as Netscape on Mac and Windows 95 and NT. Other viewers have not been tested.

The Starfield Interface

As terms are selected by the user, they are placed around the perimeter of a document summary

Starfield, and any documents matching them are added to the summary field. Each document is represented by a small marker, positioned according to the weights of matched terms. Visiting one of the term markers causes matched documents to be highlighted. Visiting a document marker causes the title and an extract to be shown, while clicking the marker retrieves the full document text.

The system may suggest terms that the user finds unfamiliar or unexpected. To explain these suggestions, a Keyword in Context (KWIC) facility is provided, to show examples of the term as used in documents found by the query.

The user may add or delete terms, causing corresponding changes in the set of documents displayed, and in their positions. The user may click a document marker one or more times to learn progressively more about it, and may mark documents they judge to be relevant or nonrelevant.

Starfield displays are most useful for large and diverse document sets, resulting from an early general query. A traditional ranked list is useful once the results have been narrowed down, but there may not be a clear division between these cases; therefore, the interface supports both modes at all times.

We followed several GUI guidelines that we extracted from the literature:

- Make simple things easy,
- Make hard things possible,
- Minimize Reading, by providing graphic representations of relations and comparisons.
- Model-View-Controller (MVC) paradigm, providing multiple views of a shared data model, so that user actions in any view are reflected in all views. Although a given view may be intended for some data attribute, a view should show "other / hidden" axes as best it can.
- Rich display, as opposed to multiple displays. The cognitive burden of understanding a rich display, even one showing abbreviated or iconified information, appears to be less than required for switching between displays.
- Visible / Discoverable controls. All active elements are visible, and respond to mouse-overs.
- Two levels of response, providing quick summaries, and further information on request.
- Explanation. Document extracts and Key Words in Context (KWIC) explain system choices.

This version offers several Starfields, to present multiple views of the document and term spaces. The only remaining conventional list is a ranked list of titles.

Finding The Important Terms

We begin by producing a list of the terms in all the query documents, keeping only those which are proportionally more frequent in the query documents than in the collection as a whole. Any ratio greater than one suggests keeping this term, but we set the cutoff a bit higher in practice, to keep the number of key terms reasonable.

Related Work

Korfhage, R. R. To see, or not to see -- is that the query? *Proc. ACM SIGIR91 Conference*, pp 134-141.
Lyberworld system. Document icons are positioned according to query terms matched. Direct manipulation of query term weights helps visualize the relation of documents to terms.

Williamson, C., and Shneiderman, B., The Dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system, *Proc. ACM SIGIR92 Conference*, pp 338-346.

Direct incremental input and immediate evaluation in a small DB, with results displayed as map overlays.

Shneiderman et al Juvenile Justice System

Two stages of interaction, query evaluation followed by results analysis, required by the slower response of a large, remotely located DB. Local analysis allows user to make fewer but more effective requests to the remote DB.

Swan, R. & Allan, J. *University of Massachusetts TR CS 100*, Enhancing User Information Retrieval Performance by a 3D Interface.

Extension of clustering to show a network of similarities between documents, by icons positioned in a three-dimensional space such that their nearness approximates their similarity.

L. T. Nowell, R. K. France, D. Hix, L. S. Heath, E. A. Fox, Visualizing Search Results: Some Alternatives to Query-Document Similarity, *Proc. ACM SIGIR96 Conference*,

Displays search results as a matrix of icons, with layout under user control, so that icon position may represent document date, author, as well as ranking.

M A Hearst, J O Pedersen, Reexamining the Cluster Hypothesis: Scatter / Gather on Retrieval Results, *Proc. ACM SIGIR96 Conference* ,

Uses a scatter-gather algorithm to partition the result set into a user-chosen small number of clusers, then presents each cluster as a smaller ranked list headed by several discriminating terms from those documents. Controls allow the user to mark either individual documents or entire clusters as relevant or non-relevant.

Luhn, H., "The Automatic Creation of Literature Abstracts". IBM Journal, pp. 159-165 (1958).

Machine-generated concordance of important words with context.

Turtle, H. R. & Croft, W. B. Evaluation of an Inference Network Based Retrieval Model, *ACM Transactions on Office Information Systems 1991*,

The Inquery retrieval system.

Please suggest additional references to Morris Hirsch <hirsch@cs.umass.edu>

Implementation Issues

The web interface is written as a Java Applet communicating with a Common Gateway Interface (CGI) C program, which is bound to the search engine. As a matter of convenience, requests from the Applet to the CGI appear as if they were HTML form submissions, and replies appear as HTML pages. This allowed the Applet to de debugged against a static HTML file, and the CGI program to be debugged against a standard web browser. Programming details are available from the authors.

Conclusions and Future Directions

Our method is itself a victim of the synonym problem; we observe less than hoped-for term overlap among documents on the same topic. The result is that a high-frequency **concept** is often represented by several low-frequency **terms,** which are more likely to fall below the acceptance threshold. We are working to recognize such concept groups in real time.

In the case of Medline, the controlled vocabulary Medical Subject Headings (MeSH) provide an alternative. By design, these eliminate the synonym problem, and should provide the required term

overlaps.

Acknowledgements

A discussion with Ben Shneiderman lead to the idea of using Starfields for an Inquery interface.

Don Byrd of CIIR suggested providing a KWIC facility for explaining the choice of important terms.

Discussions with Leah Larkey of CIIR helped clarify the ideas of term richness and term clustering.

This work was supported in part by the National Science Foundation, the Library of Congress and the Department of Commerce under cooperative agreement number EEC-9209623, and in part by NRaD Contract Number N66001-94-D-6054.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsors.

Morris Hirsch home page, http://www.cs.umass.edu/~hirsch/
James Allan home page, http://hobart.cs.umass.edu/~allan/
CIIR home page, http://ciir.cs.umass.edu/