

Information Doors
Where Information Search and Hypertext Link
May 30, 2000
San Antonio, Texas

A Commercial Perspective on Hypertext Search Results

Sally Kleinfeldt, Jaideep Baphna
Dataware Technologies, Inc.
Cambridge, Massachusetts, USA

Hypertext Search Results - The Status Quo

Thanks to Internet search sites such as Infoseek and AltaVista, the general public has become accustomed to a certain style of search results presentation, shown in figure 1. Each document returned for a query is described by:

- A number indicating its order in the results list (lower numbers indicate a better match).
- Its title hyperlinked to the document itself.
- A summary usually comprised of the first twenty or so words of text in the document's body.
- Additional information such as size of the document, its URL, when it was last modified, etc.



Figure 1: Typical Internet search results page.

This sort of presentation, although straightforward, can be frustrating for the following reasons:

- The order in which a search engine ranks documents is not always a reliable indication of their relevance to the user's information need. This misrepresentation is partly the fault of the search engine's algorithms, and partly the fault of the user, who typically provides a highly ambiguous one or two word query.
- The document summary - the snippet of text presented to indicate the document's contents - is often not sufficient to help the user assess whether the document is relevant or not. Therefore, users typically spend a frustrating period bouncing back and forth from the results page to the documents themselves (waiting for images to load and weeding out dead links), trying to find what they are looking for.

Clearly users would prefer their analysis of search results to be more efficient. In the remainder of this paper, we present several commercially available techniques that address this issue, and we discuss user acceptance of each technique. These techniques approach the problem from two perspectives:

- Helping the user formulate a question that better describes the information need.
- Giving the user better information that they can apply when judging which documents on the results page are most relevant to that information need.

Our assessment of user acceptance is based on: four years of experience supporting the commercial version of InQuery [7], a search engine originally developed at the Center for Intelligent Information Retrieval at the University of Massachusetts; 2 years of experience supporting KMS, a knowledge management product based on the BRS search engine[4]; and 2 years of experience supporting Query Server [11], a search the search engines product. An engineering staff of 5 to 15 people trained and worked with customers and developed an internal set of best practices based on our most successful applications.

Eliciting a Better Question

Although most search engines offer a query syntax (usually involving quotation marks, plus and minus signs, and Boolean operators) aimed at helping users craft a more specific query, evidence from query logs of major search sites indicates that a very low percentage of users actually use such a syntax. This indicates that expecting users to first articulate a complex and well-targeted query is not a viable strategy. Instead, the industry should seek alternatives that examine the content of the most relevant documents on the results page, and then suggest refinements to the query based on that content. Two such techniques have been developed: relevance feedback and concept mining.

Relevance Feedback

Relevance feedback provides the user with a button or link to "find more documents like this one". This button is usually located next to each document on the results page. The relevance feedback mechanism typically issues a new query that is created from the most frequent terms in the document. This feature, developed in the IR community [1,2] and present in the early days of Internet search engines, has disappeared from commercial usage. The problem is that from a results page, it is not yet known whether a document is relevant or even what it is about. For this reason, using relevance feedback from a results page is unpredictable, and negative user perceptions of this feature have led to its elimination.

Concept Mining

Concept mining is a newer technique not yet commonly available. One commercial implementation, in the InQuery search engine [7], is based on a technique called "local context analysis" [12]. LCA finds the most common concepts (such as people names, company names, noun phrases) found in the most relevant passages of the most relevant documents returned for a query. These relevant concepts can be presented to users in various ways, prompting them to refine, expand, or replace their query with one or several concepts.

Figure 2 illustrates a sample concept mining user interface. In this application, clicking on a concept's check box in the right hand frame adds the concept to the query with an "or" operator, expanding it. Clicking on the concept itself refines the query by adding the concept with an "and" operator. This explains what the concept has to do with the query – for example, answering the question "What do CD-ROMs have to do with Climatic Data Centers?"

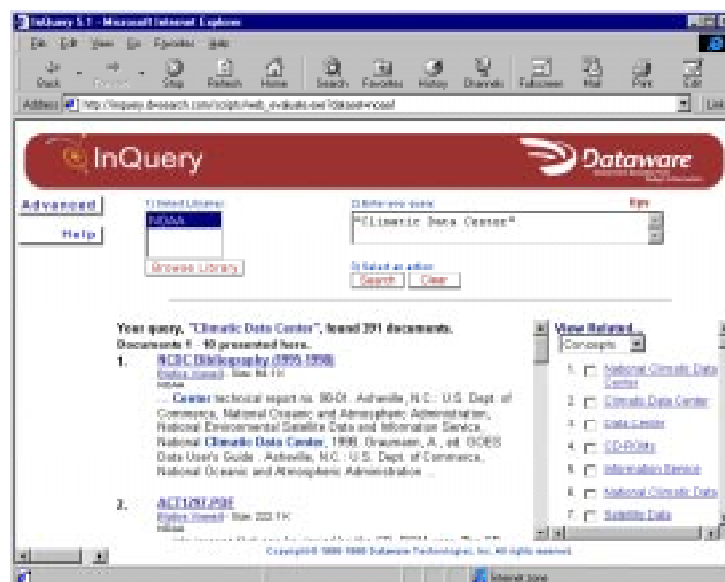


Figure 2: Sample concept mining user interface.

Although this feature inspires much positive comment on demonstration, we are only aware of one customer application currently making use of it (the search at <http://lotus.com>). What is holding people back from adopting this technique? Our impression is that concept mining offers so many possibilities that it is confusing for all but the expert searcher. Perhaps a better user interface design would lead to more widespread use.

Presenting Better Relevance Indicators

In this section, we discuss techniques for improving the information presented on a results page:

- Best passages
- Field-based summarization
- Trend analysis
- Results clustering by category
- Results clustering by content

The first two techniques deal with improved document summarization (the words chosen to represent the document on the results page). The last three techniques deal with improved document organization (how documents are grouped on the results page).

Best Passage

A document's best passage is the window of text that is most relevant to the user's query [5]. Rather than using the first few words to summarize the document on a results page, the best passage provides a better indication of the document's relevance by showing the "window" of the document where the most query terms were found. This feature is particularly effective when combined with query term highlighting. Figure 2 illustrates this approach.

Our customers are very positive about best passage summarization. They use it in preference to the other summarization techniques that are available, unless speed of the results display is critical (since the best passage calculation takes some additional processing time), or better abstract information is available from specific fields (see the next section). An Internet example may be found at <http://www.quicken.com> (use the search box at the bottom of the page).

Field-Based Summarization

Occasionally, excellent document summary information is already available in particular fields of the document source. An example is the THOMAS system at the Library of Congress (<http://thomas.loc.gov>), under the section for searching bill summary and status information. The source documents in this case contain fielded information highly relevant to this search.

Unfortunately, reliable fielded information is rare in most Web documents, aside from the occasional use of HTML meta tags such as "keyword". As standards such as XML are adopted by the Web community, better fielded information will become available. Until then, best passage summarization appears to be the best relevancy indicator in most situations.

Trend Analysis

An often-ignored dimension of document relevance is time or date information. It is not unusual for a search on a major Web search site to return, as the first result, a document that is several years old and no longer relevant simply because of its age. Simple techniques such as results sorting by date provide a way to explicitly increase the importance of time or date as a measure of relevance - but often at the expense of other relevance measurements, like document "score".

One alternative is to combine date and relevancy information to generate a histogram as a search result. An example is shown in figure 3. In this user interface, the scores of relevant documents for a given date are summed to generate an overall score for each date. These overall scores are presented in the histogram. Clicking on a date displays relevant documents from that day. The advantage of this technique is that the relevance scores generated by the search engine are not obscured, which allows the user to determine whether the score or the date is more important.

For time sensitive information such as news, organizing results by date should be quite helpful. This technique, however, is too new for us to have learned from any practical customer experience.

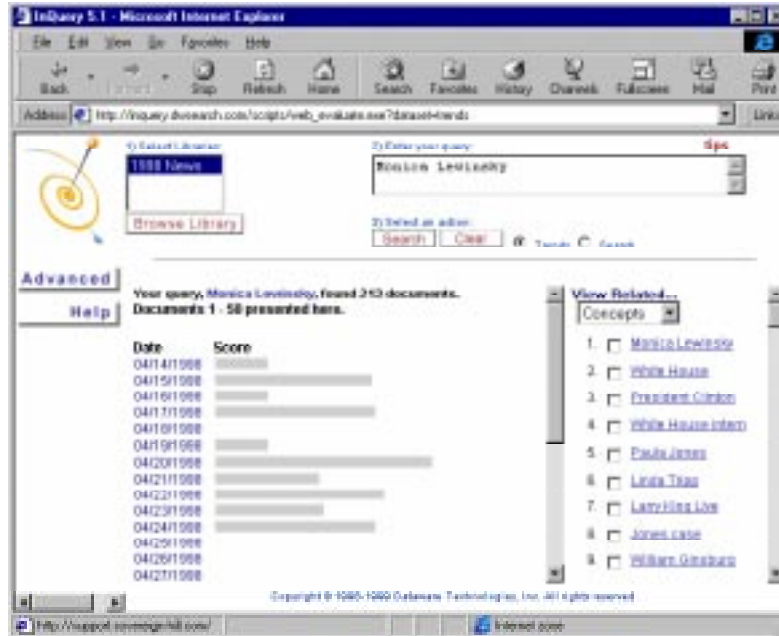


Figure 3: Sample trend analysis user interface.

Results Clustering by Category

For document collections without a strong date component, there may be other criteria useful for organizing results. In some applications the collection in which a document resides is a useful differentiator. In other applications there may be information in certain fields that can be used to assign documents to useful categories. One example is the "taxonomy" of a knowledge management system - where contributed documents are assigned to one or several categories, either interactively or through techniques such as auto-categorization. A user interface that clusters search results according to such a categorization scheme is shown in figure 4.

Departments	Rating	Type	Date	Authors	Title	Size
Departments (28)	★★★★★	Document	12/1 2/97	Steve Offsey,Dan	Knowledge Management Prospect Qu...	0
Corporate (4)	★★★★★	Document	12/1 2/97	Steve Offsey	RESUME	0
Marketing (28)	★★★★★	Document	12/1 2/97	Steve Offsey	Knowledge Management Suite Screen	46216
Product Development (1)	★★★★★	Document	12/1 2/97	Ann OLeary	Press Release: Internet & Electronic C	14336
Sales (3)	★★★★★	Document	12/1 2/97	Ann OLeary	Press Release: Dataware Technolog	17408
Consulting Services (0)	★★★★★	Document	12/1 2/97	Ann Hawkins	Press Release: Ace Hardware	16384
Customer Service (0)	★★★★★	Document	12/1 2/97	Ann Hawkins	Press Release: Adobe	16896
Human Resources (0)	★★★★★	Document	12/1 2/97	Ann OLeary	Press Release: Dataware Wins Appy A	17408

Figure 4: Sample user interface based on taxonomy clustering.

The easy navigability and rich interactivity of this interface helps the user conversant with the taxonomy to quickly find a document of interest. The tree can be expanded and collapsed, and columns can be resized and sorted. When a user drills down to another node on the taxonomy in the left pane, the interface displays the set of results found within the selected category in the right pane. The drop-down box located at the top left allows the user to switch between various categorization schemes. The entire category tree in the left pane would change if the drop down box were changed. Thus users can choose the categorization schemes with which they are most conversant. This results display component, implemented in Java to reduce server-side traffic, has proven popular among Dataware's knowledge management customers for whom taxonomies are very important.

Results Clustering by Content

In many situations, however, there is no information, such as date or category, around which to organize results. In this case, clustering can be done based on the content of the document. A simple example based solely on query term content is used by the THOMAS system at the Library of Congress. Results from THOMAS are clustered based on the exactness with which the query terms match document content (shown in figure 5). This method tends to compensate for short queries where the unstated objective is to find the exact phrase [6].

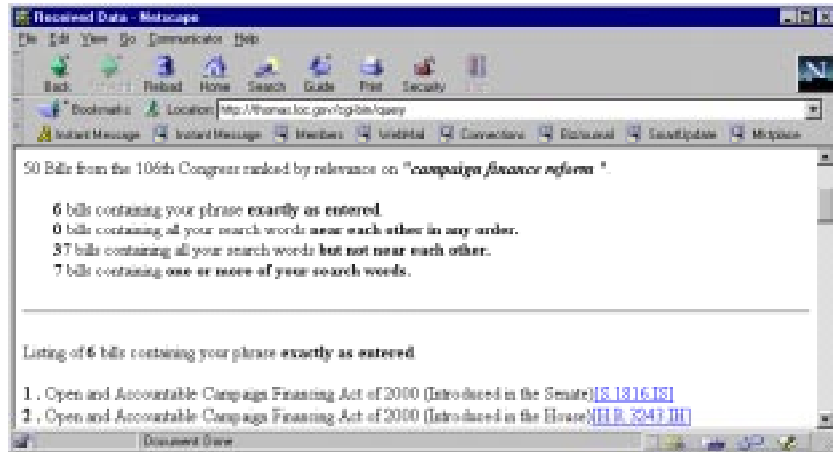


Figure 5: THOMAS system results clustering.

More complex examples of this approach use the full content of the returned documents for clustering. Many algorithms for document clustering [9] and many cluster visualization techniques exist [3]. Although the visualization techniques may generate a "Wow!" reaction, our experience is that they are not popular in real life applications because the graphics are still too unwieldy and unfamiliar on most users' workstations. A simpler cluster presentation technique is to group the documents on the results page with unique text headings that signify document content. This technique is implemented in Query Server [11] and shown in figure 6. Clustering by content has proven to be extremely popular with our customers, because relevant documents often fall within and are described by one of the clusters. Northern Light (<http://www.northernlight.com>) is an example of an Internet search site that organizes results in clusters, however the clusters do not appear to be based on pure content clustering.

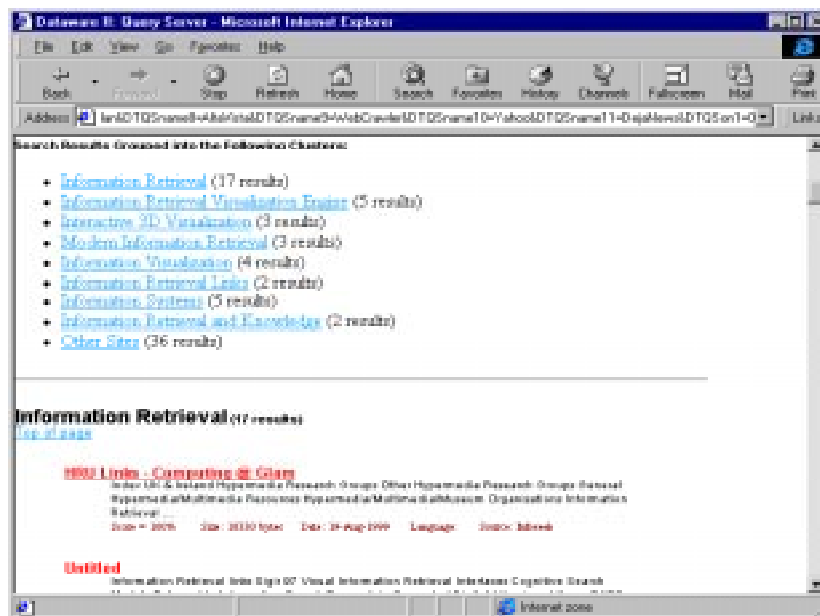


Figure 6: Sample content clustering user interface.

Looking Towards the Future

When workstation hardware and software support for 3D graphics becomes ubiquitous, more advanced cluster visualization techniques may become worthwhile supplements to text representations. One promising new development is the Lighthouse system [8,10], which combines visualization with relevance feedback in a complementary way that significantly reduces the time spent finding relevant documents in user trials.

Meanwhile, feedback from our customers clearly indicates that the search experience could be improved for most applications by using commercially available techniques that are based on purely textual information. The best approach today is to combine the following elements:

- Use summarization techniques, such as best passage, that give a better indication of a document's relevance.
- Organize the results page according to some appropriate criterion that clusters categories of related results.

In addition, using concept mining to help users improve their queries is a promising technique in need of further exploration.

References

1. J. Allan. Incremental relevance feedback. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, 1996, pp. 270-278.
2. J. Allan. Relevance feedback with too much data. In *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, 1995, pp. 337-343.
3. J. Allan, Leouski, A. and Swan, R. Interactive cluster visualization for information retrieval. CIIR Technical Report IR-116, Computer Science Department, University of Massachusetts, 1997, <http://cobar.cs.umass.edu/pubfiles/ir-116.ps.gz>
4. BRS Search. <http://www1.dataware.com/products/brs.htm>
5. J.P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994, pp. 302-310.
6. W.B. Croft, Cook, R. and Wilder, D. Providing government information on the Internet: experiences with THOMAS. In *Proceedings of the Digital Libraries Conference DL 95*, Austin, Texas, 1995, pp. 19-24.
7. InQuery. <http://inquery.dataware.com/> and <http://inquery.dwsearch.com/>
8. A. Leouski and Allan, J. Lighthouse: Showing the way to relevant information. Submitted to the *IEEE Symposium on Information Visualization 2000*.
9. A. Leouski and Croft, W.B. An evaluation of techniques for clustering search results. CIIR Technical Report IR-76, Computer Science Department, University of Massachusetts, 1996, <http://cobar.cs.umass.edu/pubfiles/ir-76.ps>
10. Lighthouse. <http://toowoomba.cs.umass.edu/~leouski/lighthouse/LighthouseApplet.html>
11. Query Server. <http://queryserver.dataware.com/>
12. J. Xu and Croft, W.B. Improving the effectiveness of informational retrieval with local context analysis. To appear in the January, 2000, *ACM TOIS*.