Hal Berghel's *Cybernautica*

"The Search is on ... the Web"

One of the most remarkable achievements of the World Wide Web has been the rapidity with which it has indexed and organized itself. And that's a good thing, for without the search tools that we discuss in this column, the Web would be as handy as a divining rod under water.

Earlier Internet document storage and retrieval protocols like Gopher and WAIS were, in comparison with the Web, relatively unsophisticated document retrieval systems which grew slowly due to the narrow educational niche markets to which they appealed. This enabled them to evolve hand-in-hand with appropriate searching and indexing resources needed for their successful use. The Web exploded into Internet ubiquitousness in just 3 or 4 years, and continues to tax its developer's abilities to harness its contents. That they have succeeded as far as they have is something of a miracle.

The root problem is that there are basically no standards when it comes to classifying, naming and defining individual Web documents. Now that there are more than 100,000,000 such documents, keeping track of them all is no small feat.

The underlying tactic is simple. Each document is named with a unique URL (uniform resource locator) which contains a server name, a path within that server's directory structure and a document name. The preprint of this article, for example, has the URL http://www.acm.org/~hlb/col_edit/cybernautica/may-june97/pcai_97c.html. Everything to the left of the three w's serves to identify the appropriate communications protocol to be used (hypertext transfer protocol in this case). "www.acm.org" is the domain name of the server on which the document resides. "pcai_97c.html" is the document name, and everything in the middle is the path structure to that document on the server.

The question for Web users is what is the content of this and scores of millions of other Web documents? I know what the content of this document is. You know what the content is. But for those who haven't read this column it's an unknown quantity. So how do we access this document if, say, we are interested in the topic of searching the Web? That's where the search engines come it. They help us find information that we want. They provide, as it were, indices for the Web. They are still pretty crude at this point, but without them all Web users would be lost in a tidal wave of useless information.

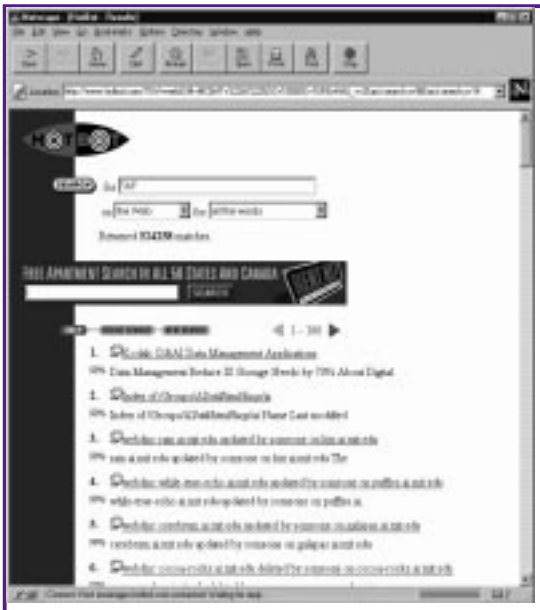
What are Search Engines?

The present species of search engine typically consists of an HTML, forms-based interface for submitting a query, an indexed database with an internal string matching routine, and some form of network indexer (which subsumes such entities as spiders, wanderers, crawlers, worms, ants and robots) which visits a specific set of network servers periodically and returns file or document-specific data for inclusion in the indexed database. Although the number and frequency of servers visited, the nature of the document extraction routines, and the robustness of the keyword-based, Boolean query interface varies by developer, all current search engines seem to target about the same level of search granularity.

The network indexer component provides the greatest computational challenge for the developer

because of the lack of universal standards for classifying the documents. Most indexers parse the contents of the document title which appears between the pair of tags <TITLE>...</TITLE> in the document's header. Many others also look to the <META> tag for such things as relevant keyword lists. Fewer others parse some or all of the document body for keywords. There are several problems with this approach.

First, a very large percentage of Web document creators either do not use the title and meta tags at all, or use them ineffectively. This means that the web indexers have less reliable information upon which to base their indexing. The search engines will then both retrieve too much unwanted information while at the same time retrieve too little wanted information. Information theorists describe this problem as a search result with both insufficient precision and inadequate recall. The end result is that the user, paradoxically, ends up retrieving much more information that they want, but still less than they need. Figure 1 illustrates a modern Web search engine in use.



if their Web authors applied header tags properly

The he and she of it is that modern search engines now index more chaff than wheat. As a data point, consider the number of hits produced in Figure 1. Veteran Web surfers will confirm that it is unlikely that more than a small fraction of these hits will yield much useful content There are, to be sure, important documents within the retrieved index, but finding the important documents amidst pages and pages of content-free screen gumbo is akin to finding the proverbial needle in the haystack.

The Search Engine Landscape

Modern, powerful search engines such as HotBot and Excite provide an example of how digital technology can be used to retrieve information. At this writing, well over 100 such search engines have been identified (128 have been identified by ugweb.cs.ualberta.ca/~mentor02/search/search-all.html), each with its own particular search characteristics. In addition, "meta"-level search engines have also been developed which utilize several object-level search engines in their operation, and then integrate their results. All4one integrates four such search engines, Highway61 seven, and SuperSeek ten, to name but a few.

Meta-level document indexers such as Yahoo and Galaxy are also available. Special purpose indexers also exist for such things as personal home pages, particular multimedia document types, academic interests, chemical structures, help-wanted ads, and so forth. In other words, there are search engines for virtually every appetite at this point. While they don't always work very well, given the size of the Web it is something of a miracle that they work at all.

Today's larger search engines now boast indices which include between 50 and 100 million URLs. Management of a database this size is a real challenge in itself, not to manage the networking aspect. Whether a document is a vanity or cosmetic homepage or a Pulitzer Prize winner, it still becomes grist for the indexer's mill. But regardless of the quality, if we end up with it and it's something that we don't want, it's still cyber-litter.

WHERE DO WE GO FROM HERE

Search engines are important tools for accessing Web resources. However, their effectiveness will decline over time as the Internet faces a tidal wave of new information.

While some additional effectiveness may be expected in such areas as indexing behavior - e.g., more sophisticated parsing and integration of <meta> and <title> tags with the index of the document body, they will not help enough. The Web and the Internet will continue to be over-indexed, and over-populated, for the foreseeable future.

Perhaps the next breakthrough in searching and indexing the Web will be personal software agents which would interface directly with the existing search engines on our behalf. Information customization tools, and the presence of Internet "brand names" will also be required to help deal with information overload. I'll return to these themes in subsequent columns.
