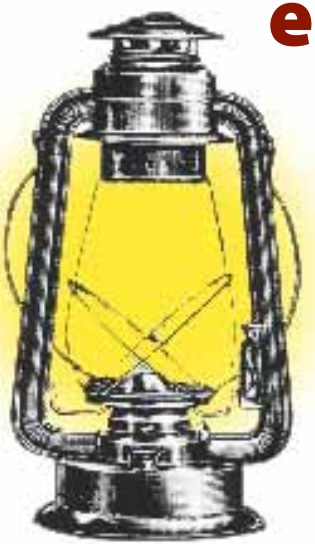
Behind the search engines (Part 2 of 2)



How they find the sites you're looking for

Search engines make the Web go. They provide a forum for advertising; they direct people to sites; they're a Baedecker and traffic cop all in one.

But how do they actually work? The search sites can't store all the information on the Internet—the task of finding material would be just as difficult as it is now. And storing other people's information would probably violate copyright.

Instead, search engines go one better: They store an abstract idea of where words live and what relationship they have

to each other. This information can be reconstituted on demand to create a hazy hologram of what the Web is at any given moment.

In last issue's column (October 15, 1996), you learned how to make good searches. Here, you'll see what you're making your searches with.

What's in an index

The search engines use homegrown servers called "spiders" that have a simple function. Taking off from the root of any Web site (that is, any so-called home page), the spider follows all links from the page, comparing them with a vast list it keeps of pages it has already indexed. In this way, a hot-links page that links to 40 other sites will cause the spider to ultimately retrieve all the pages off those 40 other sites. Or, a site that uses its home page to link to all the major areas of its site will lead the spider down to the bottom of each section.

The spider does a recursive search of all Web structures it finds—that is, a search down every path and link—until all links have been explored. Since virtually all Web sites are linked from somewhere else, exploring all hierar-

chies below any given page on the Web ultimately will result in retrieving every page. The time a spider takes to accomplish this becomes relevant, given the ever-increasing size of the Web. According to Alta Vista, the Web currently comprises 40 million pages.

Pages not linked to or from anywhere must have their top-level URL (the highest level on the area that links to other areas) manually submitted to the various engines, all of which provide a mechanism that allows Webmasters to submit those missing URLs. If these sites fail to submit their top levels, they could be called dead branches on the family tree.

Each search site has its own algorithm for what it does, but the simplest explanation is that every unique word encountered anywhere on the Web is in turn referenced to every page on which it occurs. Sites that use a "proximity" index (see my column in the October 15, 1996 issue for more on proximity) also assign values between words; two words occurring within ten words of each other, for instance, are given an association so that a user can search for phrases or pairings, like "virtual" and "reality."

The spiders generally retrieve only pure text and HTML files (which are like extended text files, with easy-to-filter codes). PDF files and other documents that don't have clear stretches of text require special extensions for searching their contents. (Adobe is encouraging the major search companies to incorporate PDF indexing; you can see a local example of that here on *adobe.mag* at our past articles page.)

Coding your pages right

Myths abound that putting "IBM" 14,000 times on your home page will make your home page come up first when people search anywhere for IBM. This is untrue, and imputes a degree of stupidity to the very clever people who write and tweak the spider software that scours the Net. (In fact, it's easy enough to detect patterns like that and exclude either those words or those pages.)

You *can* improve the indexibility (if you'll forgive my using that word) of your site with a few simple tags and techniques.

TITLE tags. Always include <TITLE> tags naming each file uniquely and

appropriately. If the HTML file in question lives on "zenauto.com" and describes part 3 of four parts on repairing an automobile engine, include something like

```
<TITLE>ZenAuto - Repairing an Automobile Engine -
Part 3 of 4</TITLE>
```

In PageMill, this information goes into the "title" field at the top of the document.

META tags. The HTML tag <META> allows you to insert special information read by servers and searchers. The two values you'll use are keywords and description, in this syntax:

```
<META name="description" content="Beef Yakky
Corporation catalog page 1. Taste the full flavor of
smoked, salted, and boiled yak, shipped in
permafrost.">
```

```
<META name="keywords" content="beef,jerky,yak,smoked
meat,boiled meat,mail order,permafrost,yummy">
```

The description is a summary of what's on the page; the keywords are a comma-delimited list of terms associated with the contents of the page. Some search engines will use these META descriptions and others won't, but you can bet this or something similar will be used increasingly as the Web grows larger, so work of this kind won't be wasted.

A good, concise explanation of these and other tags can be found at *http://www.tiesoft.com/kwiug/search41.htm* and in the slightly more bizarre tutorial at *http://www.cnw.com/~drclue/Formula_One.cgi/HTML/META/META.html*.

Submit URLS. You can use the automatic engines like Submit It! at *http://www.submit-it.com*, but you still have to go through and check that everything made it in, and in the right form. A colleague of mine, Eric Ward, runs a service called NetPOST at *http://www.netpost.com* that's the paid, expert equivalent of Submit It—you're paying for a human being to compose the

entries and follow up for you, repeatedly if needed.

Generally, sending in URLs to the top search sites is good enough, but if you want to make yourself really integrated into the fabric of the Web, your next task is to research all sites related to yours and offer to exchange links. Film.com, a company I've worked with from near the beginning of the commercial Web, has an extensive links section that gets a fair amount of use. But the key factor is that the sites linked to also link back. This is a subtle and effective way to generate more traffic while simultaneously producing more matches in search engines. If Film.com, for instance, is listed on 40,000 other pages, that produces a lot more positive matches at Alta Vista.

Personal spiders and Webmaps

One recent development is personal spiders, where individuals can build a site map of a remote site by recursively exploring links "local" to that site—links with the same base host name. In the same vein, link verifiers use spider-like activity to follow a chain down and check the links as they go.

An example of each genre, respectively, is NetCarta's WebMapper and Pacific Coast Software's SiteCheck. There are more of each, with many variations; strangely, though, Yahoo doesn't have a category for client-based spiders yet!

WebMapper takes a home URL and builds an outline of the whole site; it can check for bad links and other kinds of behavior as well, and even allows you to open the page (if the server it's on can be mounted locally) and edit the bad links. Primarily, though, WebMapper creates a comprehensive overview of the entire structure of a site. It's available for UNIX and Windows 95/NT.

SiteCheck is an inexpensive, single-minded application for the Macintosh that takes a URL and follows and checks all links from that level down, generating a list of responses. It's an easy way to check what's working on a site, both locally and in terms of links to offsite locations.

Danger! Danger, Will Robinson!

If you look at your Web logs, you have certainly noticed queries for a file called "robots.txt" at the top level of the site. If you log browser information, you've also seen thousands of queries from browsers that call themselves spiders or robots. These queries are generated by the spiders making their journeys over the Net. The "robots.txt" file is a simple list for incoming spiders of what to index or what to avoid; it's also called a "robot exclusion file."

Most sites prefer that spiders don't send information to Common Gateway Interface (CGI) programs, since the potential input to many programs is unlimited, and spiders can go out of control sending queries that generate new links that they then query.

A good example is the Architext search engine, which excite gives away to Webmasters. Every search on Architext generates not only a response with matches, but also links to produce new searches on similar matches. If a spider were allowed to follow that kind of search, it could be endlessly recursive, given that each search generates new queries that may appear to the spider

to be unique files to retrieve.

The "robots.txt" file is a standard followed by all "well-behaved" spiders; that is, all spiders that have been programmed to obey it. Unfortunately, since the Web is an open medium, you can't restrict spiders automatically, and someone could write a spider that ignored this entirely. The only solution for Webmasters in that case is to lock out a remote site entirely via a router filter. This is an extreme step that's rarely used.

There's a syntax to the "robots.txt" file that allows you to request that certain spiders only index certain parts of the site; it also allows you to create a general directive saying what's off-limits to all spiders. The syntax for these techniques, and the full technical story behind spiders, is neatly laid out at *http://info.Webcrawler.com/mak/projects/robots/norobots.html*.

A book worth noting

A new book may you help sort out some of the meaning here: *Getting Hits*, written by my good friend and colleague Don Sellers and due to be pub-

lished by Peachpit Press *(http://www.peachpit.com)* in early winter. It covers many of these search engine issues at length, as well as Web advertising and other related means of publicizing a site.

Searching for meaning

The fundamental basis of search engines is to provide windows of direction into a mass of undifferentiated information. One trend may affect this: Sites trying to reach large audiences with timely information increasingly feed information live from databases or through structures that are dynamically driven, even customized to each user's desires. Search engines can't update fast enough to contain accurate links to sites like this; hotwired, at least, has provided a permanent archive location for all articles and features, allowing at least part of its dynamic Web site to have a hardwired (sorry …) location.

URLs

Getting Listed on the Search Engines
http://www.tiesoft.com/kwiug/search41.htm

Submit It!
http://www.submit-it.com

HTML Guide: Meta Tags
http://www.cnw.com/~drclue/Formula_One.cgi/HTML/META/META.html

NetPOST (WebMapper)
http://www.netpost.com

NetCarta (SiteCheck)
http://www.netcarta.com

Pacific Coast Software
http://www.pacific-coast.com

A Standard for Robot Exclusion
http://info.Webcrawler.com/mak/projects/robots/norobots.html

©1996 Adobe Systems Incorporated