

t a p p i n g

t h e w e b :

a

f o u n t a i n

o f

i n f o r m a t i o n

c h a p t e r

t h i r t e e n

"We're not in Kansas anymore, Toto." This is Oz; we're on the World Wide Web. We should be able to find anything we want just by typing in the words we want to find!

It's not quite that simple, but several big engines are running out there, indexing tens of millions of pages, just so the user can type in words associated with the topic. It is commonly taken for granted that the computers know everything, and that we should be able to ask them simple questions and they will tell us everything we need to know.

Well, the World Wide Web is not the whole Internet, and the Internet itself is a single-digit fraction of all the knowledge in the world. But the fact remains that more information is now literally at our fingertips than was ever even dreamed of. Vast information resources are now available in a globally accessible library that defies the imagination.



Search engines, the combined hardware/software systems that provide the indexes for the ever-changing content on the Web, allow users to perform global research on the Internet.

t i p

Read the Help files at each site!

Searching the WWW: Tools To Target Your Information Hunt

<http://www.altavista.com>

t i p

Searching for plain old "Java definition" and the more tightly defined "Java NEAR definition" on AltaVista yields a vastly different number of hit documents, but the same document appears at the top of each search.

Such documents are worthy of study of their construction and contents. Check out the "Java definition" example.

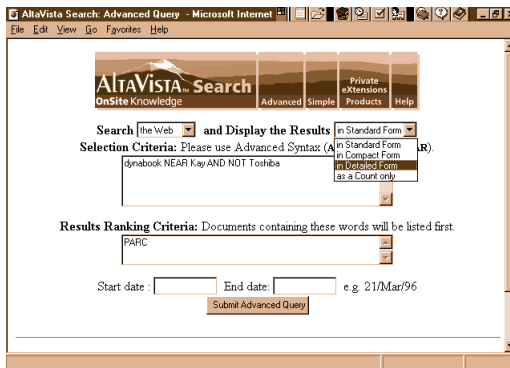
This is the summary of the top page on the hit list on both vague and precise searches for “Java definition” (about 100,000 hits) and “Java **Near** definition” (500 hits).

Java: Definition

Java: Definition. Java: Definition. Cool Object Oriented programming language designed from the ground up to be secure. Including garbage collection, <http://search.netscape.com/misc/developer/conference/proceedings/j4/tsld003.html> - size 737 bytes - 27 Mar 96

Considering the URL, it’s not so surprising that the Netscape Developer Conference Proceedings might have a clue about getting documents to rank high.

In this example, the words “Java” and “definition” comprise 42 characters out of a total file size of 737 bytes (characters) much of the remaining material is comprised of stopwords, and therefore unindexed and unranked in the hit list. This is a simple but extremely effective example of using the key-word and concept at the level of the Web document.



The AltaVista Advanced Search screen allows the user to specify where to search, the organization of the hit list and date ranges, and it has generous screens for complex Boolean queries and term weighting entries.

Excite

http://www.excite.com




Excite and Magellan merged, and their home page offers both concept searching and the Magellan editor's reviews.

Search Philosophy

Excite is designed to help users find information when they don't exactly know what they seek. They characterize searching in three generations: first, keyword searching; second, thesaurus-aided keyword searching; third, a concept knowledge base.

Paraphrasing the corporate explanation, the Excite index maps information in N-space to construct relationships between terms based on a probabilistic technique. A smoothing technique is used internally to clean up the matrix.

The resulting knowledge index points to statistically significant sentences because the Excite engine is never working directly with just words. Architext's engine concentrates on concept search rather than keyword search.



In layman's terms, N-space means that individual terms are related by their co-occurrences in the database. For instance, if the name "Kelly Johnson" often appears within the same sentences, paragraphs or sections as the terms "SR-71" and "Skunkworks," these terms will be close in N-space. Therefore, a concept search for any one of these terms may retrieve items that mention the others. A search for "SR-71" might also find articles on "Kelly Johnson's" other planes, such as the WWII twin-fuselage P-38 Lightning.

Compared to semantic term expansion, which is based upon fixed associations between terms, such as those found in dictionaries, thesauri and other reference sources, N-space term expansion is built upon the relationships computed between terms within a particular collection. Specifically, the indexing process creates these relationships between terms, which is referred to as N-space because it has unlimited dimensions or relationships between terms.

Smoothing may refer to processes that control or prevent overly wide term expansions.

Functional Description Of The Search Engine

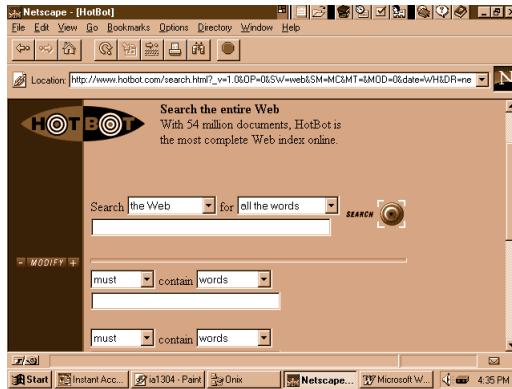
The user is presented with a field for natural language entries, with radio buttons to select either Concept or Keyword search.

The first 10 hits are served up, with automatically generated summaries of the hit documents. The Web Click-to-Go-To feature is used to view the documents. Users may change from Sort by Confidence to Sort by Site.

While Architext uses a natural language for queries and differentiates itself from the rest of the field through concept searching, suggested advance search techniques are built upon traditional information-retrieval techniques. All of the tips on the Boolean And, Or and Not search operators in Chapters 7 and 12 are valid in this concept engine.

HotBot

http://www.hotbot.com



HotBot offers sophisticated features such as term weighting and phrase searching in this painlessly simple menu, which appears upon clicking the Modify button on the initial search page. HotBot is also the first Web search engine to allow you to specify a media type as a Smiley Face: "%-)" and so on.

At the SAP Sapphire User Group meeting in Philadelphia in September 1996, Bill Gates shared the happy news that Microsoft was finally "totally pure." Mr. Gates was referring to the fact that the last mainframe-style mini-computer applications within Microsoft Corp. were now running on networks of micro-computers. This step was, of course, largely helped along by Microsoft Windows NT.

"Slurp the Web Hound" is the name of HotBot's robot that downloads 10 million Web pages per day. A joint production of HotWired magazine and Inktomi is interesting for both its simple user interface and the massively parallel architecture that it runs on. Inktomi is a company founded in February 1996 in Berkeley, California, to build massively parallel systems using large numbers of low-cost computers on high-speed networks.

In the case of HotBot, this NOW (network-of-workstations) architecture employs many PCs to perform both the indexing "Slurp the Web Hound" functions and the search functions required by Web users. The idea is that this architecture is vastly expandable with the simple addition of disk drives, cheap processors and network services, as needed.

Like Architext's Excite search engine, HotBot offers this excellent "free" service as its initial product. Of course, it generates revenues through advertising space on its pages, but Web users do enjoy an excellent free searcher.

tip

It always helps to register your URL with any engine you hope will find you. Most search engines include an "Add your URL" link on their search pages, which takes you directly to an HTML form to enter your info.

example

Four Searches On HotBot For "Quark to PDF":

Searching for the phrase "Quark to PDF" with the **Must Contain** modifier terms "Quark" and "Acrobat" retrieves 12 documents, most from the PDF-L Listserv Archives and Newsletters.

Without the **Must Contain** term-weighting operator, the search retrieves 19 documents, which include less-relevant items, such as Seiko's printer driver Web page.

If the **All Of The Words** operator is used (which is an **And** search) instead of **Phrase**, the search retrieves 2,122 documents.

Searching for **Any Of The Words** (which is an **Or** search) retrieves 214,021 documents.

Even in the last example, several of the same documents appear in the top 10 hits for all searches. Closer study of the hit summary for each one of these consistent hits would give insights into HotBot's relevancy ranking.

**From the HotBot FAQ:
Overview at:**

<http://www.hotbot.com/FAQ/faq-overview.html>:

HotBot's cryptic operating instructions also add:

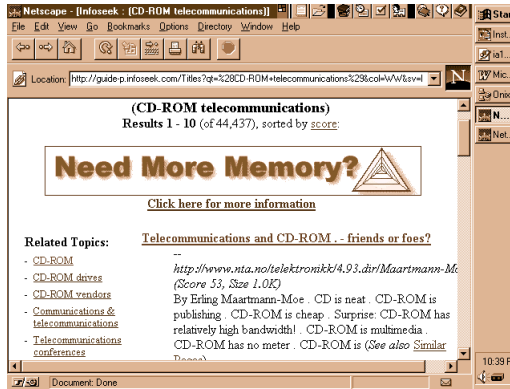
*HotBot is a search engine.
HotBot does not care which
realtor you use.*

*HotBot works for a magazine.
HotBot does not like long
documents.*

Do not taunt HotBot.

Infoseek

<http://www.infoseek.com>



The Frames presentation offers a main results list on the right two-thirds of the page, with conventional related topics on the left third. As in photography, the Rule of Thirds often leads to pleasing page layouts.

e x a m p l e

Searching for the “definition of ActiveX” on Infoseek, consider the results of the simple proximity operator:

Search for “ActiveX definition” with no query operator: *21,108 hits*

Search for “[ActiveX definition]” with **Proximity** operator: *10 hits*

It is much more likely that you will *find the definition of ActiveX in the focused 10 hits* than in the impenetrable mass of hits of the simple search. Note that the brackets “[]” are the **Proximity** (or **Near**) operators here.

T Big Iron On The Web

One of the most striking things about these engines is their impressive speed. While it is almost impossible to pinpoint the performance of any particular Web session, all of these search sites are noticeably quicker than the average site. And the indexes being searched must be very large. AltaVista claims to index 50 million pages and more than 3 million Usenet articles. Open Text states that it indexes every word on the WWW.

To accomplish such blazing performance, these search engines are running on high-end platforms. Understandably, DEC's AltaVista runs on the vaunted 64-bit Alpha processors. Since Sun Microsystems sponsors Excite, users enjoy the pleasures of the 8-CPU Sparc 1000E. Although the platform for Open Text is not disclosed, suffice it to say that its architecture is designed from the ground up to run in a multi-processing environment.

Steve Kirsch of Infoseek provided the big picture in the November 1995 issue of Boardwatch magazine: "We have a bunch of Sun machines, a T3 and a T1 coming in, a couple of routers, and about 350 GB of disk space. All together, there's 30 CPU's." So, if you want your Web site to be speedy, there's the blueprint.

tip

This is how Infoseek explains the basics of searching:

"Five Quick Secrets To Better Searching"

- 1. Capitalize names and titles, such as December and Star Wars.**
- 2. Use double quotation marks or hyphens to group words that are part of a phrase. This offers multi-word string matching.**
- 3. Use brackets to find words that appear within 100 words of each other, such as words you would expect to see in the same sentence or paragraph. This is the Proximity or Near operator.**
- 4. Put a plus sign (+) in front of words that must be in documents found by the search. This is the weighting element that affects relevance in the hit list.**
- 5. Put a minus sign (-) in front of words that should not appear in any documents found by the search. This is the Exclude function that allows a user to specifically list an entire set of erroneous "noise" hits for an otherwise effective query.**

If you're looking for information on OCR, you probably don't want to retrieve the OCR Orange County Registry, which is some sort of matchmaker service. By using the **Exclude** operator on "Orange County," articles that include both your intended string "ocr" and the excluded term will be ignored or ranked very low.

Open Text

<http://www.opentext.com>

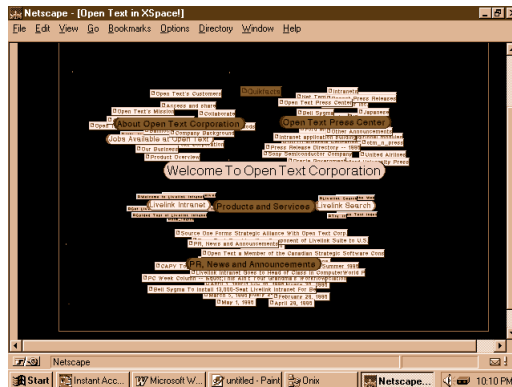
The Open Text Web index is unique in the focus upon exact words. In addition, queries can be directed to specific content:

Summary URL Title Hyperlinks First Heading Anywhere

The focus capability of this search engine is available via pull-down menus and gives users the tools to easily construct highly specific queries.

In addition to these field operators, the Open Text Web index also offers the standard Boolean operators, which include:

Near - Proximity **Or** - Loose Accrual **Followed By** - Order of Terms
And - Specific Accrual **Not** - Specific Exclusion



Here's the X Dimension view of complex collections, allowing users to "fly" through docs.

Lycos

http://www.lycos.com



The Lycos site shows its heritage in a very functional but still comfortable style. Lycos offers filtering and organizing services, and everything is easily accessible through a smart, simple interface.

Yahoo!

http://www.yahoo.com



Yahoo! is a classic secondary publisher, like BIOSIS (since 1926) and Lycos, and it provides the excellent service of reviewing and organizing raw info resources.

www.Search.com

http://www.search.com



A place on the Web that stables the horses of all the Web searchers. This site is crammed with searching options. Its raison d'être is to further the state of Web searching. "Choose your weapon," as the menu says.

Getting Noticed: Attracting Humans And Electronic Spiders

Everyone wants to take advantage of the "free advertising" by putting up billboards on the Information Superhighway, by posting pages on the World Wide Web. Just as with all advertising, it doesn't work if no one ever sees it. The Web search engines are being used every day by millions of people around the world, and well-designed pages should attract tons of interested readers. As in conventional advertising, the right product with the right packaging could be a gold mine on the Web.

The topic of how to attract spiders has become hot, and many articles and even books are being written on the subject. To show up at the top of the hit list on the big search engines is highly desirable. In effect, today's Web page designer must write for two audiences: the potential readers or customers (the humans), and the spiders (the robots), which determine the relevancy of the page, and determine whether a page is number 1 or number 100 on a results list.



Spider is a popular name for the software robots that constantly roam the Web in an effort to maintain up-to-date indexes for the search engines.

Each of the search engines offers advice on how it performs its indexing and hints on how to catch the spider's attention. Unlike conventional advertising, sex doesn't necessarily sell. Repeating the word "sex" on your page may not attract exactly the clientele you are hoping for, unless you are hoping for perverts and the FBI. So, Rule Number 1 is to pick the subject of your page and concentrate on it.

Though the various spiders go about their business in different ways, they will all start at the top of the page. The top of an HTML page contains the Title and Heading fields. Some spiders are only going to dip into the page a little bit, like maybe only the first few hundred characters, so it is important to use those characters well. So, Rule Number 2 is to use a Title and Heading that describe the core offerings of the page.

In addition to the HTML fields that are displayed by browsers, there are an additional class of tags called Meta Tags. As the name implies, these fields are used for information about the page. Specifically, the Meta Tag can be used to enter up to 1,000 characters of keywords, which may then be used by the spiders. So, Rule Number 3 is to use Meta Tags.

Finally, as Web users become more sophisticated, the content of the page becomes of paramount importance. The Web surfer is spending time and effort seeking valuable or interesting information. So, *Rule Number 4 is to concentrate on content and offer some real and unique value to readers of your page.*

How They Work: Spiders, Robots, Web Wanderers

The Web search engines depend upon a process of constantly indexing the ever-changing galaxy of information on a myriad of sites and pages. The World Wide Web is built on the HyperText Transfer Protocol, which provides instantaneous hypertext links between any two sites anywhere on the Internet, as long as they both support HTTP. This infinite interconnectedness spawned the image of a "web."

It's only natural that a web should have spiders walking on it. Spiders are also known as robots or wanderers that peripatetically follow hyperlinks around the Web and index the sites and pages they find.

Meta Tags



Meta Tags are elements in HTTP headers that can be included in HTML documents for search and management purposes. They do not display in normal view. Meta Tags provide custom index fields in the HTML environment, which can facilitate complex searching.

example

The following is the top of an HTML page, illustrating the use of Meta Tags. All comments are in italic.

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 2.0 plus SQ/ICADD  
Tables//EN" "html.dtd" >
```

Document type declaration, inserted by HTML editing program.

```
< HTML >
```

This signifies the beginning of the document, and a corresponding </HTML> tag is found at the end of the document.

```
< HEAD >
```

```
< TITLE > OCR Lab, Optical Character Recognition, Document Understanding,  
Text Searching </TITLE >
```

One of the elements that comprise a valid HTML document (head, title, body). This title is normally displayed in the browser title bar.

```
< META NAME = "keywords"
```

```
CONTENT = "OCR Lab, Optical Character Recognition, OCR, Document Un-  
derstanding, Text Searching, Information Retrieval, Web publishing, digital  
documents, PDF, Portable Document Format, text retrieval, search, Acrobat,  
Kofax, Cornerstone, Intrafed, Xerox, Caere, TextBridge, Omnipage document  
understanding, sgml, html, icr, forms" >
```

```
< META NAME = "description"
```

```
CONTENT = "Where Paper Documents become Digital Documents, how to  
get There from Here!" >
```

These two Meta Tags, Keywords and Description, are designed to attract Spiders to index the page under these words in the search engine index, especially Keywords.

```
</HEAD >
```

```
< BODY >
```

< H1 > From Books to the Web < BR > The On-line OCR Lab < /H1 >

This is the first line visible in a browser in normal mode.

< HR > < UL > < LI > < P > < EM > **Optical Character Recognition -_-
Document Understanding -_- Text Searching -_- Digital Libraries -
--**< /EM > < /P > < /LI > < /UL >

These few items are designed to let a reader immediately understand the type of material found at this site, and also to attract those spiders that index the first few lines of a page. Many search engines also include the first several words in a quote or summary in the hit list. It's important to make them meaningful so they stand out on the hit list that the humans read, no matter where they are ranked by the robots.

tip

On a search engine that allows field searching, all Meta Tags are searchable. Any search for "Keyword Field CONTAINS OCR" or "Description Field CONTAINS paper AND digital documents" would retrieve the document in the above example.

In this sense, a very sophisticated database can be constructed from simple HTML conventions, as long as all the users understand the particular conventions that have been added to a collection.

Smart HTML vs. Spamdexing

"Spamdex is a method used by a number of promotion companies in an attempt to push a site to the top of search results for certain keywords you specify," according to www.exploit.com. According to Lycos, spamdex is a "data manipulation trick ... we're happy to report no longer works in Lycos." Infoseek also advises against spamming your Meta Tag, saying that any tag containing a term repeated more than seven times will be ignored.

tip

Some spiders dip into the page by reading only Header 1, 2, or 3 information. Put your key words and concepts into your headers! They have to be readable by both humans and robots.

To understand the open nature of the Web, it's important to remember that HTML is a simplification of an earlier ideal representation of info called SGML, or Standard Generalized Markup Language. SGML was originally proposed decades ago as a universal format that would span all the end-user operating systems and all the communications media. Like UNIX and the Internet, HTML was designed to be "open" and easy to use, which opens the door to abuse.

For example, since Web browsers are designed to display both text and inline images, the user wouldn't be surprised to browse to a page that is mostly just a .GIF graphical image. It takes a little longer to download, but it just pops up in the browser. So, a "page" in Netscape Navigator or Microsoft Internet Explorer could display text and graphics but really only be showing a .GIF image displayed inline with no need for user intervention. The user, staring at the screen, is seeing words; it's a simple inline GIF image but it looks like text. When you View Source (in your browser) you will see that there is no information, only a clever inline Graphic Image Format file with tons of Hot Terms buried in invisible text. These spurious spider attractors garner a high ranking for the page, but the page is usually a dead end leading to some marketing deal, which almost always is a waste of time for the erstwhile information seeker on the Web.

The Excite engine serves as an excellent spam detector because it generates a "summary" of the page, and on these spamdex pages the search term is repeated ad infinitum. On the hit list, these spam pages stand out. The Excite summary exposes many excellent examples of how not to design Web pages.

t i p

Most Obvious Spamdex Move Number 1: Use "invisible text" to repeat search phrases.

Most Obvious Spamdex Move Number 2: Re-register and deliver the same info over and over.

In the long run, justice is done because the human Web searchers are seeking content, and when they are tricked into viewing an empty hole, they will not go back to it.

!Register - It!

This great resource on the Web allows you to directly register your page to many search engines in one fell swoop! Suddenly everyone in the world can find your page on the Web.

<http://www.register-it.com>

T The !Register-It! FAQ

This is an excerpt from the FAQ (Frequently Asked Questions archive) that illuminates the far-reaching effect of the !Register-It! Service:

Exactly which sites will I be registered with?

We maintain a database of over 1,000 sites on the Internet where your potential customers may find you. Depending on your industry, geography and several other factors, we register you with over 300 of these sites.

How long does it take you to register my site once I sign up?

We send your site out for registration with all relevant sites on our list within 24 hours of receiving your registration. But please note that several search engines and popular sites may take several weeks before registering your site.

Summary

These Web search engines offer the fulfillment of technology prophecy. From Vannevar Bush in his epic "How We May Think" article from the *Atlantic Monthly* in 1945 to Bucky Fuller in the Education Automation sermons he delivered in the 1950s, today's World Wide Web offers the true embodiment of a technological vision. What used to be just an unusually compelling and popular sci-fi prediction is now an increasingly popular reality. Anyone with a computer and a phone connection, from anywhere in the world, can navigate a vibrant and expanding global library of information.

