Searching the Internet

How to find what you’re looking for

Brenda F. Bell, ACGNJ

Abstract

On the billion-page Web, there are thousands of sites that claim to be able to find what you're looking for – but only a few that will give you what you need, quickly, accurately, and in a language you can understand. We'll discuss search engines, metasearch sites, indexes, portals, vortals, and online databases, and techniques for finding the information you want quickly and efficiently.

Keywords: *Internet search, document retrieval, indexing, query structure, search engines*

Introduction

With over a billion distinct Web pages, finding anything – much less anything specific – on the Internet can be likened to searching for a sewing needle in a haystack the size of New Jersey – or a research library the size of North America. Like any library, it has a number of card catalogs and reference librarians to help you find what you’re looking for. Find the catalog, find the librarian, and you’ve found the key – or have you? Too often, search engines seem to come up with sites unrelated to the object of your search, while ignoring what you’re really looking for. In the following paragraphs, we’ll take a look at tools and techniques for increasing your ability to find what you’re looking for, while minimizing the time you have to spend looking for it.

Technical Terms

These are some technical terms relating to searching for information in a library, database, or museum (or on the Web):

- **collection** All of the books, articles, pamphlets, papers, etc. owned by a library, or available in a section of the library; all of the items (artifacts, documents, etc.) owned by a museum or a department of a museum; all of the pages on the Web e.g., *the reference collection, the music collection, the fine arts collection*
- **record** A record of a particular item in a collection, along with various sorts of key data about the item (title, author, origin, key words, etc.)
- **field** A specific type of key information about an item in a collection
- **descriptor** A key word or phrase that describes a physical item, or the content of a document or image

- **catalog** The accumulated records of all items in a collection, including information on each item's location), arranged in a manner that makes it easy to search based on the information in a specific field of the record (*subject catalog*, *title catalog*, etc.)

- **index** A list of key data in a single record field, linked to the items in a collection (*subject index*, *contract number index*, etc.) **Index** may also be used as a transitive verb, meaning "to create an index by assigning descriptors to the items in a collection"

- **acquisition** The process of purchasing, requesting, or otherwise gaining items to add to a collection

- **retrieval** The process of finding something specific within a collection and presenting it to the end user

- **scope** Topics about which the collection contains documents (e.g. "home improvement", "computers", "aerospace")

- **coverage** Depth and breadth of detailed information available on any given topic in the collection's scope

- **hit** An item in a collection that fits the criteria of a given search

- **relevancy ranking** In a selection of documents that meet the base criteria of a search, a numerical or ordinal description of how close each document comes to fulfilling the more complex criteria, or how many times a descriptor shows up in the document

- **granularity** Specificity – in other words, how precise is your search, and how closely do the items retrieved match your search criteria?

Three Keys to Effective Searching

There are a couple of basic things you need to know to be able to find something you're looking for:

1. What you're looking for.
2. Where to look for it
3. How to ask for it

It helps greatly if you can reduce what you're looking for to a couple of key words and concepts, and alternate wordings and phrasings for them. For example, if you're looking for information on fullerenes – a family of ball-shaped carbon molecules most frequently found in outer space – you might also want to look for "buckminsterfullerene", "buckyballs", and "C_{60}". If you knew there was a relation between fullerenes and nanotubes, you could consider looking for "nanotubes" as well.

If you look for up-to-the-minute materials on a site that specializes in Renaissance music, you're not likely to find what you're looking for. On the other hand, if you can find a place that specializes in carbon compounds, you might just find that buckyball waiting for you...

If you ask for "C60" on a site that deals with metals as well as organic materials, you'll come up with a lot of steel alloys as well as fullerenes. However, if you ask for "C60" and **not** "steels", or "C60" **and** "organic", you should cut out a lot of the stuff you're not looking for. If you know that the site uses "C60" exclusively for the steel alloy and "C(60)" for Buckminsterfullerene, you can search for "C(60)" and avoid the metals altogether.

Search Tools

Tools for searching the Internet have existed almost since its earliest, pre-Web days. Tools like **archie** and **WAIS** used a command-line interface while their successors **gopher**, **jughead**, and **veronica** used menus to simplify the commands the user would need to enter. There are still a few gopher sites around, but they are declining in popularity. Similarly, a number of Web sites that once used the WAIS engine to index their sites have since switched to something more end-user-friendly.

The most important thing to know about these search tools is that their search syntax carried through to many of today's search engines.

Types of Indexes

There are several types of Web index, not counting the "Find (on page)" command embedded in the Web browser, classified based on the information and types of documents in their collections, the way that information is indexed or categorized, and the way in which documents and information may be retrieved.

- A **full-text index** retrieves documents based on an index of (almost) every word in every document in a collection. "Stop words" such as "and", "the", and "it" are generally not included in the index.
- A **keyword-retrieval index** finds documents based on a limited number of descriptive words, phrases, or concepts assigned to those documents
- A **bibliographic index** retrieves documents based on, well, bibliographic information (author, publisher, place of publication, ISBN number, etc.)

The number and structure of catalogs and indexes on any site depends upon the site's search engine, and the site's design and purpose.

Indexing for Retrieval

An item in any collection is useless unless it can be retrieved upon demand – whether that demand is to read an article, study an artifact, or look at a medical record. The process of organizing a collection for purposeful retrieval is called indexing. Traditionally, this was a job done by professionals with a background in library science; today, much of it has become computerized or computer-assisted, and done by authors, publishers, and museum curators. Even today's PC operating systems have become indexers of a sort, looking at the data in our files and rearranging them for easy retrieval.

The most basic automated indexing and retrieval programs assign keywords and relevancy rankings based on the number of times a particular word will show up in a document. This is, by and large, inefficient, as the programs cannot interpret whether a phrase exists as a quote in an unrelated context, nor can they understand synonymous and related terms, and adjust relevancy rankings accordingly.

Larger collections usually need more complex indexing schemes than smaller collections in order to break down the data into chunks small enough for the average user to handle. Many managers of large collections first divide the collections by general interest area, and then index the subcollections with more specific descriptors. The general method of breaking down a collection into smaller and smaller (and better-defined) topics is known as **hierarchical indexing**, and the structure of the index known as a **taxonomy** (same as in biology). Another way to improve the retrievability of a large collection is to use **controlled-vocabulary** indexing, which limits content descriptors to a predefined list or taxonomy of acceptable words and phrases called a **thesaurus**.

Depending upon how the data are expected to be retrieved, a document may be indexed with only the narrowest (most precise) applicable descriptors, or it may be indexed with several hierarchically-related descriptors.

Searchable Sites and Online Indexes

Not all sites with a "search" form are "search engines" Strictly speaking, almost none of them are any more. A **search engine** is a set of software routines for finding items in a collection that meet user-defined criteria, and probably best relates to the programming behind the "search" button. Search engines may be licensed for use on a particular site and may be customized to that site.

Most of the sites we call "search engines" are in reality **portals** and **search portals**. A **portal** is, as its name implies, a "doorway" onto the Web. In addition to a search engine interface, most portals include limited directories of pre-vetted sites, access to Web-based e-mail, news, games, features, and shopping. A portal designed to meet the needs of a specific special-interest community is also known as a **special-interest portal**, or if it's aimed at a business, industrial, or commercial audience, a **vertical industry portal** or **vortal**.

Before the emergence of the World Wide Web, professionals who needed to research current publications in their fields would have to subscribe to **abstract journals** – indexes of titles, authors, subjects, and bibliographies of recent publications. During the mid-1980's, many of these were adapted to proprietary online formats known as **online databases** and marketed through database aggregators such as DIALOG, Engineering Information (Learned Systems, Inc.), Silver Platter, and Cambridge Scientific Abstracts. Many of these database aggregators have added Web interfaces to their database offerings. These providers might best be described as **database portals**.

Related to search engines is a class of programs known alternatively as **spiders**, **robots**, or **crawlers**. These are programs that systematically search the Web for new pages, automatically index them, and insert them into a site's search catalog. The engine behind *Lycos* is probably the most famous of this class of tools.

Choosing the Right Search Site

If you know of a site that is likely to have the specific information you're looking for, go there directly. There's no reason to rummage around a general-purpose portal for an HP printer driver if you know that Hewlett-Packard is located at http://www.hp.com. Of course, if you didn't know that in advance, you could search for the main HP site from a general-purpose portal, and then click over to get your drivers.

If you're looking for something a bit more technical or esoteric, you might have to do a two-stage search – use the general-interest portal to find a more specific search site, and search from the more-specific site. Sometimes it helps to go to a **metasearch** site – a site that searches several collections from a single submission, searching each collection in its native query mode. The disadvantage to that is, some of the search engines behind metasearch sites cannot properly replicate the query language for some search engines , particularly when you use their "advanced" syntax (that is, try to ask a complex question).

If you're constantly looking for highly technical information in a particular field – genetics, for one example, or microwave communications, for another – you might want to check your corporate or university library for subscriptions to one or more online databases or database portals in your field of research, or sign up for the (sometimes free) abstracts and alerts from one or more technical journals. (You will need to sign up for the journal, or use the library's subscription, to access the main articles online.)

Of course, if you're just looking to quickly access the text of the U. S. Constitution, you'll do just fine just entering "United States Constitution" in the text bar of your favorite general-interest search portal.

Ask a Simple Question…

Computer programmers have an axiom, "Garbage in, garbage out" – GIGO, for short. What it means is that if you give the computer bad data to start with, you'll get bad results. In terms of Internet searching, if you don't ask for what you want to find, in a way that the search engine can understand, the search engine will come back with everything **but** that which you're looking for. The process of deciding how to ask for something in an index is called **formulating a search strategy**. The most simple search strategy is to ask for a single word – "constitution", for example – and see what comes up from there.

> *Note:* While most search engines are not case-sensitive to all lower-case (miniscule) letters, they *are* sensitive to upper-case (majescule). In other words,
>
> **constitution** yields "constitution", "Constitution", and "coNStitUtIon";
> **Constitution** yields "Constitution"

This is all fine and dandy, but if you enter "constitution" and you're looking for the text of the Second Amendment of the U. S. Constitution, you may find yourself staring at pages of text about the various State constitutions and social clubs' constitutions-and-bylaws instead. Even if you enter "Constitution", you'll run into pages on "Old Ironsides". In the search industry, these are known as **false hits**, and having too many of them usually means that either your search probably wasn't specific enough or you didn't ask for what you were looking for in the first place.

As they say, GIGO.

Strategic Planning

Once your search goes beyond a single word, strategy begins to creep in. Is the search site's default an "and" search, an "or" search, or a "string" search? Is there a way I can get sites that are related, but which don't have the same word in them? How do I make sure my kids don't get back anything I don't want them to see? All these can be addressed with an appropriately-designed search stragegy.

Conjunction Junction and Boolean Operators

"'And' search?"

"'Or' search?"

The theory is quite simple, even if you've never studied logic before.

- an **and** search looks for "this **and** that"
- an **or** search looks for "this **or** that"
- a **string search** looks for the exact sequence of letters, numbers, and punctuation you entered in the "search" box.

AND and **OR**, along with their brother **NOT** ("and not"), are the most frequently used functions, or **operators**, of Boolean logic – the logic behind many search engines. Within the context of

Boolean logic, you can use regular algebraic expressions (parentheses-based groupings separated by operators) to refine your own search.

Let's say, for example, your search site consisted of scenes from the movie version of *The Wizard of Oz*, and you want to find the dialog that immediately precedes the Scarecrow's "If I Only Had a Brain". Since the Scarecrow is the first traveling companion Dorothy meets, you might look for "brain" in conjunction with "Dorothy" and "Scarecrow" – but you'll have to throw out every scene where the Scarecrow is in the company of, the Tin Man, the Cowardly Lion, and/or the Wizard himself. In logical terms, your search strategy might look like this:

> brain AND (Dorothy AND Scarecrow) NOT (Tin Man) NOT Lion NOT Wizard

Beyond "And" and "Or"

In order to properly find what you're looking for, most search engines go beyond pure Boolean logic. The most common non-Boolean operations you'll find are:

- exact-string searches
- specification of required terms
- proximity or "nearness" searches (operator **NEAR**) –"Dorothy" within five words of "Scarecrow", for example
- accrual or "fuzzy OR" searches (operator **ACCRUE**) – documents which have more specified terms, less frequently, rank higher than documents that have fewer of those tems, occurring more frequently
- wildcard or "truncation" searches ("simul*" brings up "simulate", "simulator", "simulators", "simulation", "simulations", etc.)

In addition, some search engines have provisions for limited-vocabulary searches (based on a predefined thesaurus), hierarchical indexing (use of broader and narrower terms, and related terms) and the ability to search by record field (URL, title, HTML tag, etc.), and searching by document date or document language.

Learning the Local Dialect

Not all search engines speak the same language.

While most search engines understand the concepts of "and", "or", and "not", they don't express them the same way. Just like French, Russian, and Hebrew all have nouns and verbs (but use them differently), different search engines use different symbols for indicating common Boolean and non-Boolean operators. In order to enter a complex search, you need to know how the search engine (and/or the search site) indicates these functions.

Fortunately, most search sites make this information available at the click of a mouse. Filed under headings like "Advance Search", "More Options", "Search Options", "Help", and so on, most search sites include either the raw syntax of their search engine or a user-friendly form interface. Most librarians like the power of raw syntax because they can type in the search options from the main screen's text box – but if you're not interested in learning a lot of esoteric markup, the forms will usually do just as well.

Some common markups include:

Site	AND	OR	NOT	String
Yahoo	“matches on all words (AND)” in search options screen	“matches on any word (OR)” in search options screen	-___	“___”
Lycos	“all the words (AND match)” in advanced search screen	“any words (OR match)” in advanced search screen	-___	“exact phrase (quoted query)” in advanced search screen
Alta Vista	AND in advanced search screen	OR in advanced search screen	NOT in advanced search screen	“___”
Google	(default)	(does not support)	-___	“___”
Air Force Link (uses Verity search engine)	AND	OR	-___	“___”
Web Crawler	AND	OR	NOT	“___”
Northern Light	AND	OR	NOT	“___”

Some other options these sites provide:

Site	required	Proximity	wildcard	parentheses
Yahoo	+___		___*	(no)
Lycos	+___			
Alta Vista	+___	NEAR (within 10 words)	___*	
Google	(default)		(none)	
Air Force Link	+___	<NEAR>		??
Web Crawler	+___		___*	yes
Northern Light	+___		___*___ (any length string) % (single character)	nesting

Putting it All Together

You’ve figured out a search strategy and learned how to write it out in the language of several search engines. The next thing to do is to ask the search engine is how to display it. You can usually do that from the Advanced Search screen. Some of the most common options for display are “sort by relevance”, “sort by date”, “sort by URL”, and “display/hide summaries”.

Paying the Piper

You may find a larger number of false hits in your search than you'd like – more to the point, these false hits may have absolutely nothing to do with any meaning of any of the words in your search, in any context (e.g., porn sites and wordlist sites). There are a couple of reasons why, despite your best efforts, this might happen:

- Pay for position – some search sites give preferential positioning to people who pay money to be listed in the top 50 sites for a given search.
- <META> tags have misleading information – some search sites index and position pages based exclusively on the keyword and description data embedded in their "meta" tags. Some Web writers capitalize on this by adding additional and/or unrelated keywords in their "meta" tags to try to ensure their high placement on returns to a given set of searches.
- "invisible" text – some Web writers put text outside of display tags, or text the same color as page background, for the purpose of trying to spoof spiders and crawlers.
- bad links – this implies that the database was not updated to weed out dead links. The better-serviced search sites will do this as a matter of course, but some of them require the page-owners to submit a request for removal instead.

A number of search sites have put in safeguards to limit at least some of these false hits

- site review before acceptance – people check out the submitted (or crawled) site to make sure it fits the search site's guidelines for appropriate content
- ignoring <META> tags
- blacklisting of sites that use inappropriate <META> tags
- full-text indexing – this makes it harder to ignore unrelated text and adult content

There are a few safeguards against adult content that you can add yourself:

- install a family-safe filter, such as *Net Nanny* or *CyberPatrol*.
- enable the Content Advisor on *Internet Explorer*, set content settings, and set a Supervisor Password that your young children are unlikely to guess. You can also use the Content Advisor to limit searching to a few sites that you approve of.

Note: some of these might overfilter, leaving your child without access to sites on breast cancer, for example.

- Use the "permissions" settings on your system to restrict your children to searching the Internet from child-safe sites (this may require the use of a more complex operating system such as Windows NT, Windows 2000 Professional, or Linux)
- And of course, keep aware of what your child does online, and keep the lines of communications open

Looking Ahead

Software companies, portal owners, and site owners are always looking to find new and better ways of providing you with information quickly, effectively, and efficiently. These are some trends to look forward to in the future:

More information will be available online. Between e-publishing and e-commerce, more people and more companies will be publishing more pages and more data, over more sites.

Better query processing. Software engineers are working on Artificial Intelligence techniques to make it easier to get good results from a natural-language query.

Natural language searching. A few search engines (most notably Microsoft's *Windows* Help engine and *Ask Jeeves*) say they support natural-language searching – that is, asking a question the way you normally would – but the technology still leaves much to be desired. Look for this to improve in the future.

There will be more searchable sites available to registered, paying users.

- Primary publishers will go online in a big way. Following in the footsteps of primary publishers such as Baen Books and Kluwer, more publishers will put up preview pages of advance publications, distribute certain works as e-books, put up their publication catalogs enabled for direct to the consumer sales of publications, and put up a searchable index of current and past publications. Much of this will be paid for by user subscription fees.
- Content aggregators such as *The Scientific* World will increase in number, and in the depth and breadth of services they provide. Aggregators take primary publications, break them up into easily digestible chunks (paragraphs of pertinent papers and articles, graphs, pictorial data, etc.) and index each separately. A paying subscriber can search the collection and retrieve only the paragraphs of each publication that meet his original search criteria.

There will be fewer for-free search sites, and their collections will grow less quickly than the for-pay sites. As primary publishers and aggregators realize the value inherent in their collections, they will design pages and sites that cannot be crawled and indexed by robots, and they may consider enjoining lawsuits against search sites that catalog their pages without permission. At the same time, mergers and acquisitions in the dotcom industry will cause even more consolidation among search sites, leading to fewer free search sites, or at least, fewer Web-based search engines. Since the major for-free search sites already contain millions of pages each, the influx of a thousand new pages will not create as large a percentage difference in catalog size as it will for a site which only contains ten thousand pages.

Summary

It is possible to find what you're looking for, even on today's crowded Internet, as long as you search effectively. To do this, you need to define what you're looking for, find a site likely to have that information, and execute a carefully thought-out search strategy.