

Search Engines in General

A VERY BRIEF HISTORY

Web search engines have a very brief history, less than a decade, and this brief section is a very brief summary of that brief history.

Before there were Web search engines, there was chaos. If you wanted to find something on the Internet you needed to know its exact address. The first really significant step out of that chaos and toward a degree of organization of Internet content was the development of “gophers,” server-based collections of Internet addresses arranged in a menu format. (The term “gopher” comes from the mascot for the University of Minnesota, from whence the first Internet “gopher” emerged.) Gophers were non-HTML-based and typically indexed not much more than file titles or very brief descriptions, but if you knew how to get to a gopher it would allow you to download selected files. Gophers begat Archie (which searched gophers) and Archie begat Veronica (which searched all of “gopherspace”) and Veronica begat Jughead, but by that time they had become less relevant than even the comic strip characters after which they were named and few people even got around to figuring out what Jughead was.

The gopher lineage was barely more than a couple of years old when it was overshadowed by the rapid development of the World Wide Web, which allowed exploitation of hyperlinks, full-text searching,

graphical browsers, and other easy-to-use and highly interactive technology—and the development of Web search engines.

The first successful Web search engine to emerge was WebCrawler, which came from the University of Washington and made its public debut in April 1994. Within a year three competitors were on the scene: Lycos, Infoseek, and OpenText. In late 1995 AltaVista and Excite appeared. Interestingly, much, maybe most, of the actual searching technology of use to the serious searcher today was already present in varying degrees in these earlier search engines, including features such as Boolean, truncation, etc. Unfortunately—and the impact of this continues into the present—none of these search engines took advantage of the heavy-duty searching technology and approaches found in online services such as DIALOG and LEXIS-NEXIS. Additionally, neither the search engines nor their cousins, the Web directories, took advantage of the extensive subject classification theory and practice of the last hundred or so years. These points are relevant in a very practical way in that the serious searcher must recognize that most Web search engines were and are developed for the more casual searcher, not for those who are anxious to take advantage of more sophisticated approaches and techniques.

HotBot came along in 1996 and Northern Light in 1997. HotBot brought a more sophisticated yet easy-to-use interface coupled with a very large database (by the end of 1997, it was the largest available). Northern Light brought an integration of Web searching and searching of proprietary information. Google appeared in 1998, and its “popularity-based” ranking of records and an ultra-simple interface were effectively combined to produce an engine that quickly achieved popularity among both casual and longtime searchers. Meanwhile, the race to be the largest search engine had abated somewhat until the appearance in 1999 of Fast Search, which claimed a database of over 200 million records. This impetus, along with other competitive factors, meant the race for size was on again, with four engines having hit the 200-million-record mark by January 2000.

Among the “early” search engines, Open Text was the first to bite the dust. By early 1998 it was no longer available. There will probably be more disappearances over the next two or three years, and probably the appearance of at least one or two more major search engines. In the meantime, the changes within current engines continue, though many of these are largely either fairly superficial or more a part of the “portal” nature of the service than an integral part of the “searching” aspect. (More on the portal aspect later.) We can hope that the producers of these tools will continue working on enhancing search capabilities, and there are indications that the competitive aspects will continue to nudge this along. In a few cases, it will be a step in the right direction if the engine just begins to fulfill its promises.

As with the rest of the business world, search engine companies are extremely susceptible to fads. In 1996 and 1997, the fad was to make sure that your engine had an “advanced” version, regardless of whether the advanced version really did anything more sophisticated or whether the same things could not have been incorporated into the main home page.

Of more significance in terms of benefits, 1998 brought “personalization” and “portalization.” The personalized portal or “Web gateway” idea manifested itself in localized and user-selected news categories appearing on the home page, local weather and TV listings, personal stock portfolio tracking, personal calendars, etc. (Yes, horoscopes, too.) Nourished by the search engine producers’ desire to follow the lead of others and the realization that this approach was something that could attract advertising revenues, these two closely related models quickly became the almost-universal business model for the major search engines. Though many users had not yet realized it, this portalization/personalization approach was a major step forward in terms of really bringing the Web to the level of a household and desktop “appliance”—one that’s always at hand, uncomplicated, used frequently, and, most importantly, providing concrete and obvious benefits.

The years 1999 and 2000 brought a more subtle and less heralded, but very powerful, corollary to the portal concept. In the

first year or so of portals, the added tools (such as directories, etc.) were mainly just laid out on the home page with the hope that people would use them. In 1999 there was a major shift toward automatically incorporating the content of these “add-ons” into the results pages—at the same time the search engine’s Web database is searched, it searches the subject directory, the company directory, etc., and presents those results along with the regular search results. This integration of resources has significantly improved the quality of search results by seamlessly providing the searcher with output that’s highly relevant and that comes without having to perform the search separately in several tools. For the low price of nothing you can get a search not just of the Web index, but a Web directory search, a company directory search, a dictionary search, etc.—a little bit like the “cross-file” searching in some of the older, commercial online database services.

The next step is up to the users as much as to the search engine producers. The tools that receive user attention will be retained, enhanced, copied, and valued. The problem, as from the beginning with Web search engines, is that the person likely to be reading this book (the extreme searcher), and who needs the features and tools emphasized by this book, is not the typical search engine user. The “typical” user could care less about the more sophisticated and research-oriented features. The degree to which this is true is very evident if you look at typical searches. Lycos provides an interesting, though sometimes depressing, list of favorite searches. In a typical week, the top 50 searches include 46 that are in the entertainment, sports, or games categories. The relevance of this is not an issue of elitism, or information snobbery, but the need to face the reality that the main place most search engines make money is not with the researcher using the Web for professional purposes. The good news is that the overall audience is increasing, and the number of people who use search engines for professional purposes, for investing, and for increased literacy on such topics as science, humanities, business, and medicine, is perhaps increasing more rapidly. The number of searches for

“Worldwide Wrestling Federation” isn’t likely to decrease. However, the number of, shall we say, “more intellectually valuable” searches is increasing. There are more reasons for the search engine producers to pay attention to the extreme searcher. But the serious searcher also needs to use an engine’s more serious features so that those features will stay around and be enhanced.

HOW SEARCH ENGINES ARE PUT TOGETHER

Since discussions of search engines naturally lead dangerously close to an automotive metaphor, we might as well give in and go with that metaphor briefly. A danger is that some readers already may be saying to themselves, “I don’t care what’s under the hood of my vehicle, I just want to know how to drive it.” Quite honestly, this book is not intended for the “driver” who doesn’t care to know how to check the oil. It’s intended for the researcher who wants to know at least a little more than the basics, who cares about taking a few extra steps that may very significantly improve the performance of his or her searching. To do that, it’s necessary to understand some things about how search engines are put together.

Before we can talk about the structure of search engines, it’s important to address the context in which they are now more often than not placed: the portal. The idea behind portals is that there can be a primary page (site) on the Web that a user automatically goes to first and that provides an easy gateway to that user’s most-needed tools. This gateway (portal) lays out a collection of frequently needed information and tools that save the user from having to look in several different places. For example, by using a personalized Excite page as my browser’s “start page,” in one place I can see selected categories of news headlines, my local weather forecast, my stock portfolio, my calendar of upcoming engagements, etc. Most importantly, in the context of this book, I see the query box for the site’s search engine, the box that allows me to query the database of over 200 million Web sites. We’ll be looking primarily at that part of these sites, the search engine

itself, but not ignoring the other portal features, especially when they contribute significantly to better results for a search query.

Unfortunately, in common usage the term “search engine” has, because of its origins, come to refer to both the service’s entire site and the part of that site that accepts queries and searches the large Web database. In most cases, the term “search engine” here will be referring to the latter, and “service” or “portal” will refer to the entire site. “Portal features” will be used to refer to the other tools and information provided (directories, weather, etc.). Maybe we’d better run through that one time: The AltaVista *service* provides a *portal* that includes a *search engine* and other portal features such as news, a Web directory, and other tools.

The search engine itself can be considered to have five main functional parts: (1) the engine’s “crawlers,” which go out and find Web sites and pages; (2) the database of information gathered about those pages and about other pages that have been gathered from other sources; (3) the indexing program, which indexes the content of the database; (4) the “retrieval engine,” the algorithm and associated programming, devices, etc. that, upon request, retrieve material from the index/database; and (5) the graphical (HTML) interface, which gathers query data from the user to feed to the retrieval engine.

Because of the increased degree to which portal features are being integrated into the searching process, it actually would be legitimate to consider some portal features as a sixth main part.

Crawlers

Crawlers, or *spiders*, are the programs that go out to the Web to (1) identify new sites that are to be added to the search engine and (2) to identify sites already covered that have changed. Crawlers gather information about the content of pages from sites and feed that information to the search engine’s database. Much could be said about how this happens, but for the searcher just a few points are relevant and provide an understanding of why some engines find certain pages and other engines miss those

pages, even when the page is in the second engine's database. For many engines, more popular sites (such as those that are clicked on frequently by users and those that have lots of links *to* them) are probably crawled more thoroughly and more frequently than less-popular sites. Crawlers can be programmed for *depth* or for *breadth*, or both. Those programmed for depth not only identify main sites, but identify the subsidiary pages to the main page, the subsidiary pages of those pages, etc. Crawlers programmed for *breadth* of sites are typically concerned with finding more main sites, but not necessarily identifying all the subsidiary pages of a site. As search engines have matured and become even more competitive, there has been a tendency to see a greater melding of both depth and breadth.

The Engine's Database

The total collection of information that's stored about all the individual Web pages constitutes the search engine's database. The collection includes pages that have been identified by crawlers but increasingly also includes pages identified by other sources or techniques. A very large number of sites added to search engines come from direct submissions by Web page publishers. If you examine any search engine's home page, you will probably find a link that allows you or anyone else to submit a page to the search engine. As long as the page is not just a case of "spamming," pages submitted will probably be added to the database. All or most search engine producers examine submitted pages for spam (nasty little tricks used by nasty little programmers to illegitimately increase a page's chances of being retrieved). A service may also apply other criteria but, with the exception of spam, chances are very good that a submitted page will end up in the engine's database.

Other sources may also feed into the search engine's database. The database may, for example, include pages and/or subject headings from a directory such as Open Directory or Yahoo!.

(Note: In this discussion we're using the words "site" and "page" somewhat interchangeably. Technically speaking, a "site," usually thought of as corresponding to a particular domain name, can have many pages—even thousands of them.)

It's sometimes easy to forget that when we're using a search engine, we're not directly searching the Web, but rather searching a database that contains records describing a portion of those pages that exist on the Web. Remembering this can help avoid unrealistic expectations about what a search engine can actually accomplish.

The Indexing Program and the Index

In terms of which pages will actually be retrieved by a query, indexing can be even more critical than the crawling process. The indexing program examines the information stored in the database and creates the appropriate entries in the index. When you submit a query, it is this index that's used in order to identify matching records.

Most search engines claim to index "all" of the words from every page. The catch is what the engines choose to regard as a "word." Some have a list of "stop words" (small, common words that are considered insignificant enough to be ignored) that they don't index. Some leave out such obvious candidates as articles and conjunctions. Some leave out other high-frequency but potentially valuable words such as "Web" and "Internet." Sometimes numerals are left out, making it difficult, for example, to search for "Troop 13." The good news is that over the last couple years, in general, search engines have been treating fewer words as stop words and the "Troop 13" search will work in more engines than previously.

All major engines index the "high value" fields such as the title and the URL. Metatags are usually indexed, but not always. (Metatags are words, phrases, or sentences that are placed in a special section of the HTML (Hypertext Markup Language) code as a way of describing the content of the page. Metatags are not displayed when you view a page, though you can view them if you wish by telling your browser to show the "page source." For those who don't know HTML, viewing the page

source for a page or two can be an informative and worthwhile exercise.) Without much imagination, it's easy to see how useful the content of metatags is for information retrieval. However, some engines purposely do *not* index some metatags because metatags are the part of the page that's most susceptible to abuse by spammers. This caution is taken at the considerable expense of ignoring extremely valuable indexing information.

Those familiar with HTML know that frames are used in millions of sites. (Frames are an HTML device that treats different parts of a page as somewhat independent “windows” or window “panes.”) Some search engines do not index frames, thereby causing the searcher the possible loss of some relevant sites. This weakness is somewhat compensated for by the fact that the astute Web page developer will create a “no frames” version of the site as well as the frames version. In addition, with the evolution of Web page building, frames are being used less frequently than they were in the past.

Some search engines index the words in hypertext anchors and links (e.g., “Click Here”), names of Java “applets,” links within image maps, etc. Other search engines do not. Understanding that there are these variations in indexing policy goes a long way toward explaining why relevant pages, even when in the search engine's database, may not be retrieved by some searches. It also explains why a page may be retrieved by one engine and not another, even when the same page is in both engines.

The Retrieval Engine

This is the program that receives your query and then searches the index to identify and deliver the records that match your query. In effect, two major things happen as part of this process: (1) the retrieval engine identifies the matching records by means of a “retrieval algorithm,” and (2) the engine then arranges the retrieved items in a particular order to be displayed to the user. These may happen more or less simultaneously, or they may be fairly distinct operations.

Retrieval algorithms are discussed in some detail later on. For the moment, we will just say that these programs utilize matching criteria to determine which records contain particular words, phrases, or combinations thereof. They may also match other user-specified criteria, such as whether a particular page contains audio or image files.

The part of the search engine that estimates relevance of records may be closely integrated into the retrieval algorithm or it may be a separate process. Even when it's a fairly separate process, the separateness may not be obvious to the user, and usually doesn't need to be. In some cases, the fact that two processes are occurring may be obvious, such as with AltaVista's Advanced Search, where the user must specify, in the separate "Sort by" box, that relevance ranking should occur.

The HTML Interface

What users see when they connect with a search engine is the HTML-based *interface*. This interface gathers query data from the user, and sends that data to the search engine for it to do the retrieval. Its most obvious function is to provide a means for the user to specify the query. However, the interface also serves several other functions, including providing a space for advertisers (which consequently generates revenue for the search engine company), providing access to the various portal features, and providing links to "Help" pages and other information about the service.

THE DATABASES BEHIND THE DATABASES

Having described the preceding parts of a typical search engine, it's now necessary to complicate the picture a bit and point out that not all search engines create their own databases. Some search engines rely on databases created by third parties, then add their additional special content, features, ranking algorithms, interfaces, etc. Most prominently, several search engines (such as HotBot and MSN Search) make use of Inktomi. Inktomi (with 500 million

records) has done the crawling and indexing, and access to the resultant database is sold to HotBot and others. Those search engines then can manipulate the database, provide varying points of access (field searching), and, if they wish, meld the results of the Inktomi database search with results from other sources. Consequently, searching two search engines, both of which may use Inktomi, may produce different results.

Fast Search also provides its database to others, and for the time being (unlike Inktomi) also enables access directly through its own site. The first major search engine to make use of the Fast Search database is Lycos, but expect others to follow.

PORTAL FEATURES

In the first edition of this book, this section was labeled “Add-Ons”—and therein lies an important point. The features we’re referring to are those additional tools and information items appearing on the service’s interface that are not necessarily a part of the Web “searching” function—Web directories, news, company directories, stock information, maps, weather, etc. (For our present purposes, we’re defining the “searching” function as the process where a user enters specific criteria and the service searches a database to identify and return Web pages that match the criteria.)

When the portal concept first began to be developed by Web search services, most of the non-searching features were pretty much just “added-on.” They weren’t very closely integrated with the searching function and many of the benefits they provided could be obtained in better form elsewhere.

Perhaps the first good example of effective *integration* of Web database searching with one of these other tools is Yahoo!, where the searching function and the directory functions were integrated early on. Yahoo! is more often thought of as a directory (a browsable, categorized, and selective collection) than as a general Web search engine, but because of the degree of integration of the two functions it has always deserved a seat in both camps. Yahoo! integrates browsing

particularly well because, when “searching” in Yahoo!, Yahoo!’s classification headings are searched and when “browsing” at any of the levels within the classification scheme, the searcher can choose to “search” just within that category. Yahoo! further integrated resources by providing the option of automatically searching not just its own database but also a larger Web database (first AltaVista and now Google). With the “portalization” of Web search services, the majority of services have moved toward this kind of integration of tools. As we will see, the integration applies not just to the integration of search and Web directory resources, but to other tools as well.

A final major point to consider when examining the benefits of a portal is the ability of the user to personalize the home page. Most Web search services that provide portal features also allow you to customize your page. (The same is true for other kinds of portals than Web search portals. News sites, such as MSNBC and CNN, also provide more than just their own news and make their sites personalizable.) If you haven’t personalized at least one search engine home page, put down this book and do it now! By doing so, when you log on you will see your own selection of categories of news headlines, your local weather, and your own stock portfolio. With only a little more effort, you can personalize such things as your own list of upcoming meetings, sports scores only for the teams you follow, and your local TV listings.

In the chapters on the individual services, the portal features will be identified and discussed to varying degrees, depending on how integrated they are with the searching, or how unique, useful, and interesting the feature is. Attempts are made in both Table 1.1 and the index at the end of this book to provide ways for you to easily identify which engines have a particular portal feature or type of feature.

Table 1.1 lists the more common portal features and identifies which are available within the Web search services. A check mark indicates that the feature is available either on the site’s regular home page or on the personalized home page. Be aware that these change constantly, so periodically take a close look at search service home pages to see if some new useful features have appeared.

	Alta-Vista	Excite	Fast Search	Google	HotBot	Lycos	Northern Light	Yahoo!
Personalizable Page		✓			✓-	✓		✓
Web Directory	✓	✓		✓	✓	✓	✓	✓
Yellow Pages	✓	✓			✓	✓		✓
White Pages	✓	✓			✓	✓		✓
Image Search	✓	✓	✓		✓	✓		
Audio/Video Search	✓	✓	✓		✓	✓		
News	✓	✓			✓	✓	✓	✓
Weather		✓				✓		✓
Sports		✓				✓		✓
Stocks		✓			✓	✓	✓	✓
Maps/Directions	✓	✓			✓	✓		✓
Shopping	✓	✓			✓	✓		✓
Horoscope		✓				✓		✓
TV Listings		✓				✓		✓

Table 1.1 Inclusion of typical portal features by the major search engines (*cont.*)

	Alta-Vista	Excite	Fast Search	Google	HotBot	Lycos	Northern Light	Yahoo!
Calendar		✓			✓	✓		✓
Address Book		✓						✓
Family Filter	✓		✓	✓		✓		✓
International Versions	✓	✓		✓		✓		✓
Translation	✓					✓		
Alerts		✓					✓	✓
COMMUNICATION SERVICES								
Free Home Pages						✓		✓
Free ISP		✓				✓		
Free E-mail	✓	✓				✓		✓
Free Voice Mail		✓				✓		✓
Discussion Groups/ Message Boards		✓			✓	✓		✓
Chat		✓				✓		✓

The fact that the portal aspect of these services is treated secondarily to the search function is not to say that the former is less important than the latter. Portals are treated that way because the aim of this book is to address effective Web *searching*, and what is said about portals will be in that context of searching rather than vice versa. Indeed, every searcher should consider and take advantage of what the portal concept offers. We don't just go on the Web to search. For many people, the selection, customization, and use of a portal is what, one day soon, will make accessing the Web a more frequent occurrence than picking up the telephone.

COMPONENTS OF A TYPICAL SEARCH ENGINE HOME PAGE

Whether a Web search service is primarily portal- or search-oriented, the visual appearance of the home pages differs tremendously. This is actually somewhat beneficial to the searcher as a way of obtaining a mental image of each of the various services. However, until one has gotten fairly intimate with several of the engines, the lack of consistency can add confusion. For this reason, it will be worthwhile to look at a "typical" search engine service home page to identify the content and features that the services tend to have in common. Once the similarities are seen, it's easy to take a quick look at any search engine service and get a feel for what can be done with it.

AltaVista contains most of the elements typically found on search engine home pages. See Figure 1.1.

Database Options

Some search engines provide a choice of what collection of sources is to be searched. The options may include a search of the service's main Web database or searches of other collections (databases), such as images, audio and video, proprietary journal literature, and discussion groups.



Figure 1.1 Typical home page (AltaVista)

- | | |
|---------------------------|-------------------------|
| ① Advertisement | ⑤ Search Options |
| ② Query box | ⑥ News |
| ③ Link to Advanced Search | ⑦ Other Portal Features |
| ④ Help Links | ⑧ Site Promotion |

The default and most obvious choice is “the Web,” meaning all of the Web pages included within the database of that search service. Often a text box or radio buttons are provided for search options. When this is the case, alternatives offered are usually ones for which search features and structure are similar to that for the Web search.

Frequently, there will be links elsewhere on the page for searching other databases (such as stock information databases), but for these links the search engine used is usually different and often provided by another company. For example, Excite provides such a link for stock quotes and weather, each of which lead to very different looking interface pages.

Query Box(es)

These boxes are the heart of it all since it is here where you enter your query. Exactly what you can enter (phrases, Boolean logic, etc.) depends upon the search engine. (Boolean logic, discussed in the next chapter, is in this context the capability of using “operators” such as +, -, AND, OR, and NOT to retrieve only those records that have a particular combination of terms.)

Query Modifier Options

About half of the search engines provide some option on the home page for modifying your query. The options are most often presented either as a pull-down window, radio buttons, or check boxes. They provide options for qualifying the search by language, date, special content, applying Boolean operators, etc.

Link to the Advanced Version

For all engines that provide an advanced-version option, there will be a link somewhere on the home page that leads to the more advanced version. Often the link itself is surprisingly small, almost as if they really don't want you to see it. Keep in mind that if you prefer the advanced version, you can just bookmark the advanced version's page rather than, or in addition to, the service's main home page.

Advertising

Advertising on search engines is almost inevitable. For most companies that provide these search engines, advertising and licensing of their software provide the main revenue stream from search engine operations. (If you hadn't noticed it, the fact that the ads are related to your search topic isn't just an amazing coincidence. If you do a search that includes the word “furniture,” an ad for a furniture store pops up. The advertiser has paid for that to happen. If you look at it from a positive perspective, this very targeted advertising can

be beneficial to the consumer as well as the advertiser. I wish the junk mail that comes through the postal service were as relevant.)

Directory (Topics, Channels, Classification)

For the major engines, extensive listings of additional Web information resources usually appear in one of the following formats (or as a variation on one of the following):

- “Directory,” or a classified list of selected sites. Keep in mind that in each engine, these “selected” sites constitute only a small portion of the number of sites found in the Web database of the search engine. Some of the search engines have a directory they have created and maintain themselves, while others use a directory that’s made available to several engines. Currently, Open Directory (available in its “native” form at www.dmoz.org) and LookSmart (LookSmart.com) are the directories of choice for several search engines.

Since Open Directory is the more research-oriented directory encountered, it’s worthwhile to go into a little detail about it here, rather than repeat that detail in each of the search engines that use it. Open Directory is the largest of the Web directories, with over 2 million records. Unlike Yahoo!, with several hundred paid editors, Open Directory uses volunteer editors—over 30,000 of them. On the one hand, this could mean more variable quality in their choices of sites, but on the other it means that many of the editors are much more experienced in their specific areas than Yahoo! editors can be. On the whole, the quality of the content seems to be quite good and a good page is more likely to quickly get into Open Directory than into Yahoo!. Open Directory has 15 top-level categories, and most categories/subcategories go down four or five levels. It contains cross-references and descriptions (“scope notes”) for categories, and allows searching within each level of the

hierarchy as well as at the top level. Individual search engines implement Open Directory somewhat differently

- “Channels,” or specialized pages on particular broad topics such as business, entertainment, or sports. Each of these pages may contain directory listings for that topic, searchable sites, etc. In Excite, for example, under the Business category, you’ll find a link to the Business section of Excite’s directory, a stock quote search, company directories, a collection of online business tools, and a variety of other business-related links.

Site Promotion

This is where the search engine producer puts in its plug for how great the search service is. It usually highlights special features or content, so at least glance at this occasionally. The services use this area to point out some interesting features that might otherwise be missed.

Other Portal Features

This category covers the numerous and varied features such as those listed in Table 1.1.

Help Links

This will lead you to one or more pages that tell how the search engine allegedly works. While most of what you read in the help pages will be correct, unfortunately, some services occasionally promise things they don’t really deliver. In some cases the services provide features that aren’t documented in the help screens. Some services have been known to make major changes without taking the time to update their help screens. In general, the help screens are done conscientiously and the quality has continued to improve.



Tip: Learn two or three engines well, but use the others frequently.

WHAT TO REALISTICALLY EXPECT FROM THESE SERVICES

Especially for those who have extensively searched such online services as DIALOG and LEXIS-NEXIS, expectations for Web search engines may need to be tempered considerably. The variety of features, the sophistication and reliability of features, and, in some cases, the reliability of retrieval provided by Web search engines still are often not up to par with that provided by those established commercial services. The very nature of general Web search engines, particularly their goal of reaching tens of millions of users, at the moment precludes the level of customer support one expects from those older services.

However, the level of tolerance of such shortcomings can be significantly raised when we remind ourselves that the Web search services are FREE! There are no per-minute charges, no subscription charges, and no output charges.

The gap between traditional retrieval expectations and Web search expectations is further narrowed when a couple of other factors are considered. Recognition of both of these factors is important for the searcher who wants to get the most out of either kind of search service.

First, Web search engines are dealing with very unstructured data, or at least data with very little consistency of structure. Indeed, there is a definite structure to the HTML behind the Web pages, but for the actual intellectual content, about the only “intellectual” structure is found in the titles and metatags. The body of the pages has little consistent structure that the Web search service can use for structured searching. This situation will change as Web page builders begin to make better use of options such as XML (eXtensible Markup Language), which provides virtually unlimited identification

of the various kinds of data that might exist on a page. Some search engines are prepared to take advantage of this and are just waiting for sites to provide them with this kind of structure within pages.

Second, the sheer volume of data currently on the Web—in combination with the volume added every day—should add a degree of respect for what the Web search engines have accomplished in a very short period of time. The fact that there's at least an elementary level of access to the hundreds of millions of pages of material is a feat that should inspire much more awe than disappointment.

In a July 1999 article (“Accessibility of Information on the Web,” *Nature* 400:107-109, 1999), Steve Lawrence and C. L. Giles reported on their continuing study of the degree to which search engines cover the total content of the Web. In the article, the researchers estimated that the Web at that point contained 800 million pages of information and that the major search engines each covered well less than a quarter of that material. They estimated that of the 800 million pages, Northern Light covers only 16 percent, SNAP and AltaVista 15.5 percent, HotBot 11 percent, and for the others they studied, less than 10 percent each.

It should be pointed out that their numbers are not accepted by all observers. Some search engine producers, in particular, feel that the numbers given are greatly inflated by the fact that a very large number of the pages counted in the study are actually duplicates, with different URLs really referring to the same page (e.g., www.onstrat.com and onstrat.com), or actual duplicates of the same page on different servers, etc. Plus, a large portion is spam. If these observers are correct, Web search engines actually are covering a much larger proportion of the Web than indicated by the Lawrence and Giles study.

Whichever is correct, to add some perspective to those numbers, keep in mind that covering even a fourth or so of the published Web pages may actually be pretty good. Though there are of course the big issues of selectivity and quality to consider, in regard to extent of coverage consider that the more traditional indexing services have never covered anywhere near those percentages of “published”

material. Respected services such as *Chemical Abstracts*, *Psychological Abstracts*, and others don't even make an attempt to cover everything published that makes mention of, respectively, chemistry or psychology. In a nutshell, take advantage of what the Web search engines do cover, and search more than one engine when you want to retrieve as much on your topic as possible.

Even when several engines are searched, be aware that there is one very large portion of the Web that search engines at present cannot cover: the so-called "invisible Web." These are primarily Web pages that lie behind password-protected sites and/or pages that are part of databases that require user input in order to be searched. To access the content of these databases, you must either register and enter a password and/or enter a query on a search page found at the site. If you need access to the pages contained in these sites, you need to go directly to the site, rather than attempt to search them using a general Web search engine. For an excellent collection of links to this type of site, take a look at the Direct Search site compiled by Gary Price of George Washington University (gwis2.circ.gwu.edu/~gprice/direct.htm).

For a reasonable set of expectations regarding *searchability*, there is one overreaching aspect that needs to be considered. In general, most Web search engines are not designed for the *serious* searcher. For the most part, they are designed for the casual user, not the person who needs to apply what they retrieve in the business and research environment. When a search engine's documentation uses *Baywatch* stars in its search examples, we get a sense of their assumed audience. Facing this fact while at the same time making the best use of what is offered can prove to be the prudent approach. If serious users take advantage of the more sophisticated features offered, more sophisticated features may follow. With the number of competing search engines catering to the casual searcher, some may break away and target those who need heavier-duty retrieval power. Indeed, we've already seen this happen in the case of Northern Light. Other search

engines have also begun to at least take greater note of the needs of the “extreme searcher.”

There are some other things *not* to expect:

- Consistency from one search engine to another. This can be seen as more of a positive than a negative as it’s too early in the game to come to definitive conclusions about what are the best ways to provide Web searching.
- The traditional tools you’re used to with the older online vendors (such as controlled vocabulary, full range of Boolean and proximity connectors, tailored output formats, etc.)
- Comprehensive bibliographic searching—For listings of what has been published in journals, books, technical reports, dissertations, etc., the Web search engines will still not provide even moderately definitive results, especially for retrospective searches. For many subject areas, the best bet for bibliographic searching is to either use one of the commercial services or find a database on the Web, such as ERIC (the database for the education literature), that covers your area of interest.
- To know what’s happening during the search. Experienced online researchers often like to know all the finer details of what’s happening behind the scenes so that they can get a good sense of whether they’re really accomplishing their retrieval goals. Exactly what’s happening behind the scenes is considered very proprietary by the Web search services (for competitive purposes) and this, in combination with some obvious inconsistencies, means that extensive knowledge of the details is usually not achievable. (In terms of my own desire to know every last little detail of what is happening, my own advice to myself is, “Get over it.”)

Finally, *don’t* expect all the specifics you learn about any particular search engine today to be true tomorrow. Rather, learn what factors are

involved in the searching process so you can interpret what you are seeing and so you can make the next move in a reasoned manner.

SUBJECTS/AREAS COVERED BY SEARCH ENGINES

For none of the search engines profiled in detail here is there any documented or noticeable intent to focus on one type of Web page content over another. This is of course at least partly due to the fact that the engines covered here are the “general” Web search engines and we’re not addressing the specialized search engines, of which there are an increasing number (see Chapter Twelve).

UPDATE FREQUENCY

The “currentness” of the contents of a Web search service’s database is primarily dependent upon how frequently crawlers crawl known sites, how quickly the new and changed pages they find are added to the database, and how quickly “submitted URLs” are visited and added to the database.

Even within a single Web search service, these factors can change frequently. Sites currently within a search engine’s database may be revisited every few weeks, but more popular sites may be visited more frequently and less popular sites less frequently.

The timespan from when a new page was submitted or crawled until it gets fully indexed ranges from a day (maybe less) to a matter of months. Various engines make various claims, with varying levels of credibility. You may be able to find a page that was added yesterday. However, be aware that it may also take weeks or months in some engines. Pages that are linked from high-profile sites have a good chance of being found more quickly than those from obscure sites.


Some services promise to get submitted sites added within a day or two, while others let you know it may be a matter of weeks. Also, just because a page has been added to the database doesn’t mean

that it's fully indexed—this may be done in stages, with the URL itself indexed first, then the title, and, sometimes even months later, the text of the page.

TYPICAL RETRIEVAL AND RANKING FACTORS

Once the user has entered a query, that input goes to the program that searches the engine's database to determine (1) which records should be considered as having matched the query, and (2) in what order those records should be displayed. These two functions can work rather independently or they can be essentially a single function.

The first function, the identification of records, is most typically done based on either (a) using a default approach in which the user has entered terms, phrases, or sentences without any required syntax, or (b) using input from the user that conforms to a syntax involving criteria such as Boolean operators, proximity operators, field specifiers, etc.



Tip: Bookmark your favorite search engine for direct access, rather than using the search links offered as the default when you first loaded your browser (for instance, the Netscape Search link on Netscape's Netcenter or the Microsoft default page on Internet Explorer). On any site where you see a single query box for which you can choose from a list of search engines, remember that you are most often using a dumbed-down version of some of those engines.

When the user has not used a structured syntax, the most simplistic approach for identifying the records is for the retrieval program to take all or some of the words the user entered, connect them with either a Boolean AND or OR, and search the database using that Boolean expression. With only a small degree of marketing license, this can be referred to as “natural language searching,” which in a rudimentary sense it is. Those who have spent a major portion of their lives working with the tremendously sophisticated and complicated aspects of natural language processing (NLP) may be understandably annoyed when natural language terminology is used so loosely. Most search engines go beyond that rudimentary form and indeed make use of more sophisticated approaches and techniques. In most of the major engines, however, whether explicitly or otherwise, the Boolean matching is an integral part of the whole process. There are alternatives that bypass the Boolean and identify the records to be retrieved on the basis of popularity factors and sophisticated linguistic analysis involving such factors as co-occurrence of terms.

When the user makes use of a specified syntax, such as Boolean, that may even override an engine's default algorithm. By choosing to go with a syntax, the user is saying, “Thanks anyway, but I know what I'm doing and I'd prefer to take more control of the process.” Some might think of the two approaches as the difference between a TV dinner and a meal prepared from scratch. The relative merits of the product depend on how good a cook one is. A single engine



Tip: If it's not documented, guess but don't assume.

If it is documented, don't necessarily assume it always works—i.e., don't assume that it was you who made the mistake if it doesn't work.

may provide all of these alternatives: a default algorithm based on implicit Boolean and other criteria, user-applied syntax, and sophisticated linguistic analysis.

With the first function of the program being the identification of “qualifying” records, the second major function of the search engine’s retrieval/ranking program is to determine the relative relevance of each record. This is often expressed as a “score” or “ranking”—i.e., the program’s estimate as to how well a particular record meets the intent of the query. As stated above, this can be integrated into the first function, with a record’s “ranking” determining whether or not the record is retrieved (only those meeting some threshold score will be displayed in the results).

Because of the competitive nature of the search engine industry, details of the retrieval and ranking algorithms are often closely guarded. For effective use of search engines, it’s useful to go into a little more detail about the factors that are involved—the things the search engine looks for in a record to determine if it should be retrieved and how it should be ranked in terms of relevance. The latter usually determines the order in which records are presented to the user. In the profiles later in this book, the “known” factors for each engine will be discussed briefly. Those interested in knowing more should examine whatever details are provided in the engine’s online documentation.

The factors that go into determining whether or not the record is retrieved and the record’s ranking (score) usually incorporate some combination of the following:

- Popularity of the page—How “popular” a page is has become a factor for most engines. In some engines (like Google) it’s the primary factor. Popularity is usually measured in one of two ways. “Link” popularity assigns a value to a record based on how many pages link to it. “Click” popularity assigns a score to a record based on how often people have clicked on that record at other times when the user’s particular query was searched.

- Frequency of terms—If a query term occurs more than once in the record, points are accrued. Greater numbers of occurrences may add additional points, but most search engines put a limit on how far this goes, in order to defeat programmers' attempts to unjustifiably increase rankings by simply repeating a word numerous (even hundreds of) times. The length of the document is sometimes also factored in, with two occurrences in a short document providing more points than two occurrences in a long document.
- Number of query terms that are matched—If your query consists of three words, those records having all three words will get more points than a record having only one or two.
- Rarity of terms—If your query has one term that's very common and a second that occurs only a few times in the search engine's database, a record containing the rare term may get a higher score than one with the common term.
- Weighting by field—If a query term occurs in the record's title, that counts for more than if it only appears later in the text.
- Proximity of terms—If two of your query terms are close together that counts for more than if they are far apart.
- Weighting according to the order in which the searcher entered terms—A record containing your first term may get more points than one containing the word you entered second.
- Word variants (and/or truncation)—Some engines can identify words that have the same root as your query term (for example, plurals). The engine may then retrieve records containing those variants as well as records containing your exact term.
- Case-sensitivity—Some engines distinguish uppercase from lowercase. In these situations, the engine can refine your search by returning only those records with an exact case match. If in your query you enter "AIDS," those engines can return only those records that have that word in all uppercase, and prevent you from having to look at lots of records about instructional aids, breathing aids, etc.

- Analysis of documents in the database—Term association, associative networks, cluster analysis, co-occurrence, and a variety of other linguistic-based approaches may be applied.
- Relevance feedback applied to retrieved records—As a second step on the user’s part, in some engines you can identify a record you like and ask for “more like this one.” The engine then examines records that have similar content to the record you liked.
- Date—More recent records are given more points than older records.

BENCHMARKS

To understand the differences between the search engines, it makes sense to do some specific head-on comparisons as to how much is actually retrieved by one engine versus another. In interpreting results of such comparisons, considerable caution should be applied because of the numerous variables involved, such as presence of duplicates among the results in any engine, reliability of numbers reported by the engines, constant changes in sizes of the databases and so on. The best benchmarking for search engines is probably that done by an individual comparing results for words, phrases, etc. in subject areas relevant to the individual’s particular area of research. The following “benchmarks” however, which come from a variety of subject fields, should give some idea of the relative performance of the engines.

Before examining the chart that follows (Table 1.2) the reader should acknowledge some caveats. First, the numbers shown are those *reported by the service* for each search. It was not feasible to check if the numbers are actually “correct” in terms of whether each of the reported numbers represents a valid, unduplicated, still-available page. For a good analysis of these factors see Greg Notess’ Search Engine Showdown at searchengineshowdown.com.

Perhaps most importantly, there is one conclusion that the reader must not draw from the chart: that one can pick the engine with the highest numbers and stick with that one engine. Each of the major engines, because of the low degree of overlap (which is discussed in greater detail in the next section), can produce a significant number

Table 1.2 Benchmarking Results

	AltaVista	Excite	Fast Search	Google	HotBot	Lycos	Northern Light (1)
aberystwyth	73,795	20,320	70,468	158,000	36,100	37,439	61,138
chrodegang	422	90	356	255	98	232	201
"alvin toffler"	6,810	635	12,667	12,500	10,600	9,849	9,628
"sidereal messenger"	247	246	383	312	157	257	256
+"red wine" +cancer +resveratrol	428	353	1,021	968	1,100	616	766
+crumpton +maryland +auction	27	26	53	45	40	52	86
(trilobite OR trilobites) AND morphology	802	639	1,262	--	1,200	--	1,170

(1) Web results only (not from Special Collection)

Note that the "winner" for each benchmark is indicated in bold. What can be concluded is that there's a wide variation in the retrieval of the various engines and no particular engine always comes up with the largest retrieval. The primary reasons for the differences in numbers are the size of the database, the quality of the retrieval algorithm, and the depth of indexing of the pages contained in the database. Each of these factors also contributes to the fact that, for a typical question, each of the larger engines will retrieve records missed by the other large engines.

of results not found by its “competitors.” Even the smaller engines often retrieve some records not retrieved by the larger engines. Using only one search engine in most cases will deprive the searcher of these relevant records.

OVERLAP OF RETRIEVAL BETWEEN ENGINES

One of the most important points that can be made about using Web search engines effectively is the following: **If you’re interested in good recall (finding most of the sites that match your needs) you MUST consider searching more than one search engine.**

This is not to say that you *always* need to search more than one engine. If you’re looking for a specific page, or a specific piece of information and you find it in the first engine you search, wonderful!

However, if you’re looking for background material, if you’re not sure exactly what it is you’re after, if you look at the results from one engine and aren’t sure you have found the best answer or the full answer, **you MUST consider searching more than one search engine.**

This can be brought home by an example. Five search engines were searched for the phrase “erris head.” The following were the numbers of distinct records that were retrieved by each:

Fast Search	45
Northern Light	36
AltaVista	31
Excite	16
HotBot	9

At first glance there may seem to be a clear “winner.” However, an analysis of the individual records showed that there were a total of 64 unique records. Among these 64 records:

- The highest-retrieving engine found only 70 percent
- The second-highest-retrieving engine found 12 that weren’t in the first.

- The top two together still missed 7 records (over 10 percent)
- Excite and HotBot, the two with the lowest numbers, together found 7 records that the top three missed.

This is just one example, but similar testing using other words produces approximately comparable results.