Things To Know Before You Begin Searching

What are you really searching?

Finding the Web documents (a.k.a. Web "pages" or "sites") you want can be easy or seem impossibly difficult. This is in part due to the sheer size of the WWW, currently estimated to contain 5 billion documents. It is also because the WWW is not indexed in any standard vocabulary. Unlike a library's catalogs, in which can use standardized Library of Congress subject headings to find books in most large, general libraries in the U.S., in Web searching you are always guessing what words will be in the pages you want to find or guessing what subject terms were chosen by someone to organize a web page or site covering some topic.

The WWW is bigger than most people thought, according to *Bright Planet* (http://www.brightplanet.com/technology/deepweb.asp). The 41-page paper details the company's discovery of the "deep" Web, a huge reservoir of information filled with billions of documents that search engines like Google and indexes like Yahoo barely scratch the surface. The study's estimates that the Web is actually 500 times larger than any search engines have measured.

When you do what is called "searching the Web," you are NOT searching it directly. It is not possible to search the WWW directly. The Web is the totality of the many web pages that reside on computers (called "servers") all over the world. Your computer cannot find or go to them all directly. What you are able to do through your computer is access one or more of many intermediate search tools available now. You search a search tool's database or collection of sites -- a relatively small subset of the entire World Wide Web. The search tool provides you with hypertext links with URLs to other pages. You click on these links, and retrieve documents, images, sound, and more from individual servers around the world.

There is no way for anyone to search the entire Web, and any search tool that claims that it offers it all to you is distorting the truth.

Categories of search tools available now

At present, we find it useful to describe the kinds of intermediate search tools available to you in four categories. You use different strategies to find and exploit the potential of the tools in each class:

Types of Search Tools	Characteristics	Examples
Search Engines (& Meta-Search Engines)	• Full-text of selected Web pages • Search by keyword, trying to match exactly the words in the pages • No browsing, no subject categories • Databases compiled by "spiders" (computer-robot programs) with minimal human oversight • Search-Engine size: from small and specialized to 90+ percent of the indexable Web • Meta-Search Engines quickly and superficially search several individual search engines at once and return results compiled into a sometimes convenient format. Caveat: They only catch about 10% of search results in any of the search engines they visit.	• Search Engines recommended and described in this tutorial: Google, Alta Vista Advanced Search, Northern Light Power Search, Alltheweb • Meta-Search Engines: Metacrawler, Ixquick, Copernic, Vivisimo , etc.
Subject Directories	• Human-selected sites picked by editors (sometimes experts in a subject) • Often carefully evaluated and kept up to date, but not always -- frequently not if large and general • Usually organized into hierarchical subject categories • Often annotated with descriptions (not in Yahoo!) • Can browse subject categories or search using broad, general terms • **NO full-text of documents**. Searches need to be less specific than in search engines, because you are not matching on the words in the pages you eventually want. In Directories you are searching only the subject categories and descriptions you see in its pages.	• Recommended and described in this tutorial: Librarians' Index, Infomine, Yahoo!, IPL, Digital Librarian • There are thousand more of Subject Directories on practically every topic you can think of.
Specialized Databases (The Invisible Web)	• The Web provides access through a search box into the contents of a database in a computer somewhere • Can be on any topic, can be trivial, commercial, task-specific, or a rich treasure devoted to your topic	• Locate specialized databases by looking for them in good Subject Directories like the Librarian's Index, Yahoo!, or AcademicInfo; in special guides to searchable databases; and sometimes by keyword searching in general search engines

Search Tools and Methods

A *search tool* is a computer program that performs searches. A *search method* is the way a search tool requests and retrieves information from its Web site.

A search begins at a selected search tool's Web site, reached by means of its address or URL. Each tool's Web site comprises a store of information called a database. This database has links to other databases at other Web sites, and the other Web sites have links to still other Web sites, and so on and so on. Thus, each search tool has extended search capabilities by means of a worldwide system of links.

Types of Search Tools

There are essentially four types of search tools, each of which has its own search method. The following describe these search tools and then suggests exercises for achieving a familiarity with their use.

1. A **directory search tool** searches for information by subject matter. It is a hierarchical search that starts with a general subject heading and follows with a succession of increasingly more specific sub-headings. The search method it employs is known as a *subject search*.

- *Tips:* Choose a subject search when you want general information on a subject or topic. Often, you can find links in the references provided that will lead to specific information you want.
- *Advantage:* It is easy to use. Also, information placed in its database is reviewed and indexed first by skilled persons to ensure its value.
- *Disadvantage:* Because directory reviews and indexing is so time consuming, the number of reviews are limited. Thus, directory databases are comparatively small and their updating frequency is relatively low. Also, descriptive information about each site is limited and general.

2. A **search engine tool** searches for information through use of keywords and responds with a list of references or hits. The search method it employs is known as a *keyword search*.

- *Tip:* Choose a keyword search to obtain specific information, since its extensive database is likely to contain the information sought.
- *Advantage:* Its information content or database is substantially larger and more current than that of a directory search tool.
- *Disadvantage:* Not very exacting in the way it indexes and retrieves information in its database, which makes finding relevant documents more difficult.

Keyword searches require far more explanation than subject searches, because of their broader scope and greater complexity.

3. A **directory with search engine** uses both the subject and keyword search methods interactively as described above. In the directory search part, the search follows the directory path through increasingly more specific subject matter. At each stop along the path, a search engine option is provided to enable the searcher to convert to a keyword search. The subject and keyword search is thus said to be *coordinated*. The further down the path the keyword search is made, the narrower is the search field and the fewer and more relevant the hits.

- *Tip:* Use when you are uncertain whether a subject or keyword search will provide the best results.
- *Advantages:* Ability to narrow the search field to obtain better results.
- *Disadvantages:* This search method may not succeed for difficult searches.

Some search tools use search engine and directory searches independently. They are said to be *non-coordinated*.

4. A **multi-engine search tool** (sometimes called a meta-search)utilizes a number of search engines in parallel. The search is conducted via keywords employing commonly used operators or plain language. It then lists the hits either by search engine employed or by integrating the results into a single listing. The search method it employs is known as a meta search.

- *Tip:* Use to speed up the search process and to avoid redundant hits.
- *Advantage:* Tolerant of imprecise search questions and provides fewer hits of likely greater relevance.
- *Disadvantage:* Not as effective as a search engine for difficult searches.

5. Ready Reference refers to the reference materials used most often in answering such questions, shelved for convenience near the reference desk, rather than in the reference stacks (*Books in Print*, *Encyclopedia of Associations*, *Statistical Abstract of the U.S.*, world almanacs, city directories, *Ulrich's Periodicals Directory*, etc

Search Tools

A search tool employs a computer program to access Web sites and retrieve information. Each search tool is owned by a single entity, such as person, company or organization, which operates it from a master computer. When you use a search tool, your request travels to the tool's Web site. There, it conducts a search of its database and directs the response back to your computer.

Of the hundred's of search tools available, we have selected 15 that we believe are best, both singly for their performance and as a group for the diversity they provide. Table 1 lists these as Preferred Search Tools by the primary search method each use. In practice, most subject search tools provide an auxiliary keyword search, and correspondingly, keyword search tools usually provide subject searches.

When should I use a subject directory?

- When you have a broad topic or idea to research
- When you want to see a list of sites on your topic often recommended and annotated by experts
- When you want to retrieve a list of sites relevant to your topic, rather than numerous individual pages contained within these sites
- When you want to search for the site title, annotation and (if available) assigned keywords to retrieve relevant material
- When you want to avoid viewing low-content documents that often turn up on search engines

When should I use a search engine?

- When you have a narrow or obscure topic or idea to research
- When you are looking for a specific site
- When you want to search the full text of millions of pages
- When you want to retrieve a large number of documents on your topic
- When you want to search for particular types of documents, file types, source locations, languages, date last modified, etc.

When you want to take advantage of newer retrieval technologies such as concept clustering, ranking by popularity, link ranking, and so on

When should I use the invisible Web?

- When you want dynamically changing content such as the latest news, job postings, available airline flights, etc
- When you want to find information that is normally stored in a database, such as a phone book listing, listings of lawyers, doctors, etc. in a particular location, searchable collections of laws, and so on.

When should I use ready reference tools?

- When you would normally use these types of tools; dictionaries, encyclopedias, almanacs, etc.

Search Exercises

For those just starting to learn the search process, this segment is recommended to help you understand how the process works. The following is the general procedure:

- Connect to the Internet via your browser [e.g. Netscape or MS Explorer]
- In the browser's ***location box***, type the address [i.e. URL] of your search tool choice. Press Enter. The Home Page of the search tool appears on your screen.
- Type your query in the address box at the top of the screen. Press ***Enter***.
- Your search request travels via phone lines and the electronic ***backbone*** of the Internet to the search tool's Web site. There, your query terms are matched against the index terms in the site's database. The matching references are returned to your computer by the reverse process and displayed on your screen.
- The references returned are called "hits" and are ranked according to how well they match your query.

Now, conduct the following searches to become familiar with each of the four types search tools described above:

1. Directory [Subject Search]

Type ***http://www.yahoo.com*** in the location box of your Internet Browser [e.g. Netscape Navigator or MS Explorer]. Press ***Enter***. The Yahoo! Home Page is displayed. From the subject list provided, choose and click a category of your interest to follow. Choose titles that are increasingly more specific until there are no more options of interest offered. Scroll through the references or hits, and click a hit that interests you to get an abstract or title of the reference.

2. Search Engine [Keyword Search]

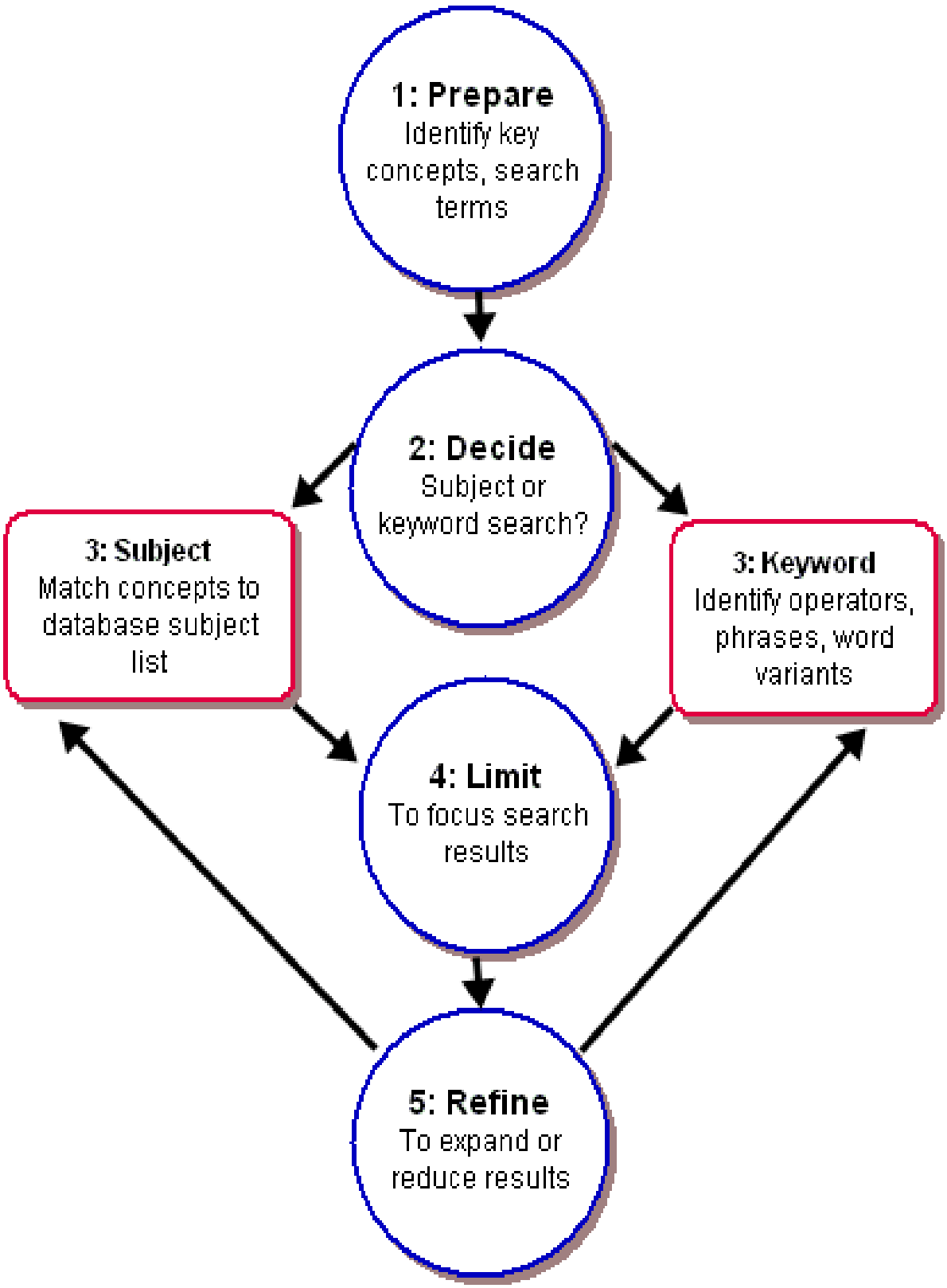
Type *http://www.google.com/*in the location box of your Internet Browser and press ***Enter*** to access the Home Page. Using keywords, type your question or query into the location box. Click ***Google Search***. (Notice the "I'm Feeling Lucky" button, what happens if you click on it?) Examine the hits of interest and click one to access the reference.

3. Directory with Search Engine [Subject with Keyword Search]

Follow the same procedure as in [1] above, except at one of the stops along the path switch to a keyword search. Type a simple query in the location box, and examine the hits of most interest.

4. Multi-Engine Search Tool [Keyword Search]

Type ***http://www.dogpile.com*** in the location box of your Internet Browser and press ***Enter***. Type the same keyword query as used in [2] above. Compare the hits with those obtained in [2].


1: Prepare
Identify key concepts, search terms
2: Decide
Subject or keyword search?
3: Subject
Match concepts to database subject list
3: Keyword
Identify operators, phrases, word variants
4: Limit
To focus search results
5: Refine
To expand or reduce results

s

Search Engine Tutorial Tip Sheet

Be natural:

Type in what you want to know, rather than a list of synonyms. Websites are written in flowing language, and search engines are being taught to understand the same.

> *If you would've asked a fellow human "Is alphabet soup nutritious?" Then ask the search engine "alphabet soup" AND nutritious rather than alphabet soup nutrition food health.*

Capitalize:

In general always use lowercase.

> *If you enter star, you will receive hits for Star, STAR, staR and so on.*
>
> *However if you are seeking info on China, the country, start it with a capital C. This will exclude a lot of sites, which focus on china tableware.*

Use Rare words:

The more unusual or uncommon the keywords you use are, the more specific the results will be. Taking a moment to think of a valid yet uncommon word is a valuable technique.

> *alcohol returns 912,620 hits (AltaVista)*
> *vodka fetched 120,740*
> *and it narrows down to 2754 hits when you enter Stolichnaya.*
>
> *Note: For a few engines the word order is important, so always enter the rare word first.*

Require words:

Adding a "+" before a search word will require it to be in every resulting hit.

> *Say you were after information on an actor from the X-Files, Forbes Angus.*
>
> *Entering those words "Forbes Angus" x-files cast fox brings out 2.7 million sites, with nothing about occasional cast member Forbes Angus on the first page.*
>
> *But if you add "+" to the least common word +"Forbes Angus" x-files cast fox then the top results of the 80 sites found will give you the info you require.*

Exclude words:

By using a "-".

> *Say you sought the homepage of Bruce Willis, a plumber in Arkansas.*
>
> *To avoid all the millions (actually 134,928) of pages dedicated to the film star, use this: "Bruce Willis" plumber Arkansas -"Die Hard" -movie -superstar –Demi*

Spell correctly:

Goto.com allows you to see how many times keywords or phrases have been searched for. Here is an example of a hard to spell place name I looked up, Machu Picchu

> *Machu Picchu 4402 (correct spelling)*
> *Machu Pichu 600*
> *Macchu Picchu 720*
>
> *A quarter of the people failed!*
>
> *Also be aware of the differences between English and American spellings, such as colour & color. In such cases use (colour OR color).*
>
> *Note: Some search engines, like Ask Jeeves, will check your spelling for you*

Recognize "stop words":

Search engines ignore the most common words, in an effort to speed things up. Several hundred of these are deemed to be "stop words". They vary from engine to engine, but always contain words like the of web a to in & is. It doesn't matter whether they are embedded in a phrase or if they have a + before them, they will not be included in the search. Usually this doesn't matter, but it is smart to be aware of the process.

> *"searching the web" contains two stop words: the & web. Consequently the search engine will only look for "searching". If you are aware of this, you can add a more relevant keyword to narrow your search, like: "people search"*
>
> *Note: Google will let you know of any words it has excluded. FAST does not use stop words.*

Use "Wild cards":

The asterisk covers all possible extensions.

> *funk* will find funk, funky, funkiest and funkadelic*
> *str*ed will find stretched, straightened and strapped*
> *"God * toga" will find "God of toga", "God on toga" and "God sold toga"*

Note: Lycos uses a "$". Not supported by Excite, Google or Infoseek

Reverse questions:

Search engines look for pieces of text that match your query. Web pages are more likely to contain answers than questions - so search for the answer. Phrase your query how you would expect the answer to read - the difference appears slight, but it makes a huge difference

> *"IRS stands for" rather than "What does IRS stand for?"*
> *"man first landed on the moon in" rather than "When did man first land on the moon?"*
> *"sky is blue because" instead of "Why is the sky blue?"*

Solve "dead links":

Try shortening the URL to the next subheading. Keep doing so until you get to the point that works. Then browse from there to see if you can track down the file that you want.

> *If http://www.spock.com/jim/life/not_as_we_know_it.html returns an error, try http://www.spock.com/jim/life/ and if you still get an error, try http://www.spock.com/jim/ and so on down to the root domain http://www.spock.com*
>
> *Note: Google has most of the web cached. If a link is dead, clicking on the Cached link will bring up how it looked when it was indexed.*

Deal with "huge" pages:

Sometimes the reason a page appears in the results is because it is one very long page of text, briefly mentioning hundreds of subjects. Sometimes these are useful, such as in genealogical searches. Often they are not...

> *In general, the most useful pages will be between 10k and 80k. To find that which you seek within a huge page, use the "Find in Page" option of your browser: for Explorer & Netscape it is in the Edit menu*
>
> *Google results list a maximum size of 101k. Many of these will be much, much larger and take forever to download.*

Use “Boolean” phrases:

Named after George Boole, Boolean phrases are a system of logical combinations, using words like AND, OR & NOT. ***It is best to always capitalize them.***

AND or "+"
Larry AND Curly AND Moe
Larry +Curly +Moe
AND requires the word to be present

OR
Chico OR Zeppo
OR allows either word to be present

NOT or "-"
Marx NOT Brothers
Marx -brothers
NOT excludes words. In this example results should display sites about communism and not comedy.

NEAR
"Salman Rushdie" NEAR teatowel
Finds keywords within 25 (Lycos) or 10 (Alta Vista) words of each other. Not supported by the other engines.

NEST THEM!
Marx NOT (Brothers OR Moscow)
("Jesus Christ" NOT Humor) AND (Mary OR Magdalene)

BUT DON'T GO TOO FAR!!
((alphabet AND Soup) NOT (twinkies OR "KFC")) AND nutritious

... is too confusing. Use "alphabet soup" AND nutritious

... and if you get a lot of KFC hits, refine the results to exclude them (most top engines support this).

Use the “right” engines:

Don't just use any old search engine - they are all very different.

Use the advanced engines These are never on the front page but they should be - I'm certain that 95% of users would have no difficulty understanding them. Ditch the standard engines and bookmark the advanced ones.

Play the field You will naturally make one engine your favorite, but when you have time you should play with the others. All the engines on this site have a unique advantage, and if you can learn what they are, your searches will become easier.

Don't flog a dead horse If you are having difficulty finding the site you want, try the same keywords on another engine (before resorting to Boolean or meditating on more appropriate keywords). Or use a metasearch engine.

According to Nature magazine (1999) the web contained roughly 800 million individual pages and 180 million images. (Amazingly, only 1.5% of the web is pornographic). The best engine listed below only covered 16% of the web.

How relevance ranking works……

What do search engines take into account?

By necessity, this section is going to be vague, since I don't know exactly what one search engine will regard as an important weighting when deciding relevance. However, I can tell you some of the things that they take into account, and I hope you will find this useful for two reasons. Firstly, if you are a searcher, it should help you to construct a better search strategy, and to understand why you retrieve what you retrieve. Secondly if you're writing a web site for yourself, an understanding of relevance ranking may come in handy when you're designing your pages.

1. Words in the title.

By title, I don't mean what you see in the main body of the screen, I mean what you see at the very top of the browser screen, or when you bookmark/favourites a site. Some search engines will pay very great attention to the words that they find in the title element of a web page, and to once again use my slightly bizarre search term, the first page retrieved by AltaVista is for 'DryNites disposable absorbent underpants', while the first entry for Northern Light is entitled 'It's all about the underpants' (I'm honestly not making any of this up, incidently!). In fact, at Northern Lights at least the first 80 entries contained the word in the title element (and probably more besides, but I couldn't cope with looking at more results!) while back at Alta Vista the same can only be said of 2 of the top ten entries. Interestingly, although Northern Light also indexes the DryNites site, it doesn't appear in it's own top ten at all. Consequently, it's clear to see that Northern Lights pays more attention to words in the title than does AltaVista.

2. The number and position of search terms input.

Most of us are aware that if we put several terms into a search the engines will usually do an OR search, which will result in a large number of references, and they will rank web pages that contain all of those terms higher than those which contain a smaller number of them. It therefore makes sense to give search engines more to work with, rather than less. However, what many people don't realise is that search engines will often (but not always) pay attention to the position of words as you have input them.
For example, the search: cars car automobile jaguar and jaguar cars car automobile look as though they are the same search. However, if you run these two searches on Alta Vista and Northern Light you will get totally different sites in the top ten, and paradoxically it seems that the first search returns rather better references than the second. I'm not going to pretend that I follow the logic behind this because I don't - all that I can assume is that both engines seem to think that most people will put their preferred term last in the list, rather than at the beginning. I would however strongly suggest that you run a couple of test searches on your own preferred search engine(s) to see if the same holds true of those as well.

3. Words in the < H1 > ... < /H1> header

If you don't author pages yourself, the above will seem like gibberish. However, < H1 > and < /H1 > are the opening and closing tags used by authors to give greater prominence; rather like chapter headings. It is therefore logical that if a search engines finds two pages that include the same words, but in one they are in the < H1 > ... < /H > tags it will give a higher ranking over and above the one which doesn't. Consequently, if you're an author yourself, it's worth while working out what your keywords are, and giving them extra prominence on your pages.

4. Repetition.

This used to be a key way in which search engines ranked pages, because if a page mentions 'widgets' seven times, it has to be more relevant than a page that only mentions 'widgets' twice, surely? Logically that is the case, but unfortunately, this idea was picked up very early on by some web site authors, and a favourite trick was to chose a background colour for a page and at the end of each page, in a small font size and with text the same colour as the background repeat keywords over and over again. The search engines would see the words but the casual viewer would simply see half a screen of so of apparently blank screen. Search engines caught onto this trick very early on however, and so they tend to either ignore repetitions such as widget widget widget widget widget widget (etc) or they will downgrade or even remove such pages from their indexes. However, some engines may well still pay attention to repetition throughout a page if it's done correctly.

5. Proximity

Pages that contain the words that you have asked for in your search where the words are close together are generally ranked higher than those pages where the words are spread throughout the page. Consequently, even if you don't do a phrase search (which is always a good idea), pages with the words 'white' and 'underpants' will rate higher when the terms are found next to each other.

6. Unusual or rare terms

If you search for the two words 'bloomers' and 'underpants' because 'bloomers' is a more unusual term (though I have to admit this is an assumption on my part; I've not actually checked this) web pages that contain 'bloomers' will rank higher than those which contain 'underpants' and pages that contain both will rank higher again. Therefore, when running your searches, simply be aware that usual unusual terms will affect the ranking you get.

7. Meta tags

These are tags that can be added to the web page by the author and which will give extra emphasis to certain terms. Some (but not all) engines look for the meta tag element (which is not visible on the page, but you should be able to see them if you View/Source). Once again however, you are at the mercy of the author here; if she remembers to put them in it will greatly affect the ranking, but if they are left out, the ranking may be lower.

8. Links

Some engines pay particular attention to the number of other pages that link to the pages that it retrieves, and they work on the assumption that if lots of people link to a page it will in some way be a 'better' page than if very few people link to it. (The flaw here of course is that new pages will have fewer links than older pages in many cases). Google [5] is a good example of a search engine that utilises this technique, and many people find that it gives particularly good results.

9. Density.

If a web page is say, 100 words long and repeats 'underpants' 5 times, that's a 5% density. If another web page is 1000 words long and contains the same word 10 times (giving a density of 1%) although the second page has more occurrences of the word the first page may well rank higher since the word is, relatively speaking, more common and therefore the page has a higher ranking.

10. Paying for placement.

Some search engines have tried using this in the past, but because people don't particularly like this (since it distorts the results they retrieve) the majority of engines have dropped this idea, preferring instead to make their money by linking appropriate advertisements to the searches the user is running.

Conclusion.

As you can see, it's a very confusing area, so if you've ever been puzzled as to why one page ranks higher than another, you're not alone! Ranking is a complicated process, not helped by the fact that engines all do it differently. That's the polite way of looking at it. The more blunt view is that it's a total mess and is compounded by the secrecy surrounding ranking. However, for those engines that use this process (rather than the Index/Directory based approach favoured by Yahoo! among others) it's the best that we can hope for. Ideally a controlled thesaurus of some sort (along the lines of the Dublin Core for example) would help bring some order to chaos, at least in the area of meta tags. However, I'm not going to hold my breath on this one since agreeing to any sort of standards on the Internet is akin to trying to herd cats.

All that I can suggest is that if you're trying a search and it doesn't work on one engine, it's not necessarily because you've done a bad search, it might just be that the ranking process just doesn't reflect your own ideas of relevance, so try another engine and run the same search again, and you may be lucky!

The "Deep" or "Invisible" Web

The content of databases will not show up in a search engine result. This is because search engine spiders cannot or will not go inside database tables and extract the data. When we refer to databases, we are talking about tables stored in such programs as Access, Oracle, SQL Server and DB2. A significant amount of valuable information on the Web is generated from databases. In fact, it has been estimated that content on the invisible Web may be 500 times larger than the content provided by search engines.

Other content not gathered by spiders includes non-textual files such as multimedia files and documents in Portable Document Format (PDF).

This aspect of the Web is sometimes referred to as the "invisible Web" because database content is "invisible" to search engine spiders. Another term used for this phenomenon is "deep Web," a much better term since database content is visible with the appropriate technology.

The phenomenon of databases on the Web has been talked about in recent years, before the terms "invisible Web" or "deep Web" were coined. People sometimes referred to them as specialty databases, subject-specific databases, virtual libraries, and other similar terms. As Web technology develops and greater amounts of information are mounted on the Web, these databases take on primary importance as information finding tools.

When dealing with the deep Web, keep these points in mind:

- **A good directory will link to database sites on the Web.** This is becausemany of these databases have Web sites of their own. For example,you can go directly to the Moreoversite and browse its datatabase of current news stories. A directory,therefore, can link to the Moreover site from its news section. A gooddirectory may be the most reliable way in which to access content on thedeep Web.

- **Many search engine sites and commercial portals feature searchable databases as part of their package of services.** This phenomenon falls under the heading of converging content, mentioned earlier in this tutorial. For example, you can visit AltaVista and look up news, maps, jobs, auctions, items for purchase, etc., all things outside the purview of a spider- gathered index.

- **Topical coverage on the deep Web is extremely varied.** This presents a challenge, since it is impossible to anticipate what might turn up in a database. In addition, this coverage will be fluid as databases proliferate on the Web.

- **Information that is new and dynamically changing in content will appear on the deep Web.** Look to the deep Web for for late breaking items. Examples include news, job postings, available airline flights, etc.

- **Information that is likely to be stored in a database is a part of the deep Web.**. This can include large listings of things witha common theme. All directories are part of the deep Web. Examples include phone books, lists of

professionals such as doctors or lawyers, patents, laws, dictionary definitions, geographical locations, items for sale in a Web store, etc.

Not surprisingly, there are Web sites that specialize in collecting links to databases available on the Web. Here are a few examples.

- CompletePlanet (http://www.completeplanet.com/) - offers searchable access to thousands of databases on the deep Web for results that include summaries from the retrieved site
- Profusion (http://www.profusion.com/) - directory of over 10,000 databases, offering the option to search for the database you need. (Was The Invisible Web)
- Subject Directory of Search Engines (http://www.searchiq.com/subjects/)- topical listing of searchable databases on the Web from the SearchIQ http://www.searchiq.com) search engine review site
- Invisible-web.net http://www.invisible-web.net/
  Directory of high quality databases on the Web especially useful to researchers
- **Price's List of Lists** http://www.specialissues.com/lol/ the Internet contains numerous lists of information. Many of these lists present information in the form of rankings of different people, organizations, companies, etc. This collection is designed to be a clearinghouse for these types of resources.

Academic & Professional Directories

- Academic Info - gateway to college and research level Internet resources http://www.academicinfo.net/
- AllLearn: Academic Directories - guides to high quality resources on the Internet in the academic disciplines maintained for distance learners by Oxford, Stanford, and Yale Universities http://www.allianceforlifelonglearning.org/er/directories.cgi
- The Best Information on the Net (BIOTN) - collection of academic resources maintained at St. Ambrose University in Iowa http://library.sau.edu/bestinfo/Default.htm
- Britannica.com - classified, rated and reviewed Internet sites from the Encyclopedia Britannica http://www.britannica.com/
- BUBL Link - UK funded project of selective resources from the University of Strathclyde Library in Glasgow, Scotland http://bubl.ac.uk/link/

- CyberStacks(sm) - A centrailized, integrated and unified collection of significant WWW and other Internet resources categorized using the Library of Congress classification scheme http://www.public.iastate.edu/~CYBERSTACKS/

- INFOMINE - large collection of scholarly Internet resources collectively maintained by libraries of the University of California http://infomine.ucr.edu/

- InfoSurf - academic subject resources from the University of California at Santa Barbara http://www.library.ucsb.edu/subjects/

- The Internet Public Library - large, selective collection from the University of Michigan http://www.ipl.org/

- Librarians' Index to the Internet - carefully chosen, organized, and annotated directory maintained by a large group of librarians in California http://www.lii.org/

- Resource Discovery Network - searchable interface to major meta-sites in academic disciplines http://www.rdn.ac.uk/

- The Scout Report Archives - searchable database of 10,000+ critical summaries of Internet resources for the academic and research community included in the Scout Reports http://scout.wisc.edu/Archives/

- Subject Guides A to Z - extensive collection of subject pages from the University of Delaware Library http://www2.lib.udel.edu/subj/

- The WWW Virtual Library - highly respected guides to many disciplines sponsored by the W3 Consortium http://www.vlib.org/

Commercial Directories & Portals

- About - large collection of topical collections gathered by company-certified subject specialists http://home.about.com/index.htm

- Google Web Directory - version of the Open Directory Project http://dmoz.org/ using the Google link ranking technology; Google search results are also included with directory results http://directory.google.com/

- JoeAnt - guide compiled by volunteers; listings include information about each site including multimedia features, chat, e-commerce, access limitations, etc. http://www.joeant.com/

- JumpCity - collection that offers a signed review of each item and a link to any Usenet newsgroup related to the topic http://www.jumpcity.com/start.shtml

- LookSmart - large collection of links to reviewed sites in thousands of categories http://search.looksmart.com/

- Open Directory Project - significant resource collection compiled by thousands of volunteer editors owned by America Online and promising extensive expansion. http://dmoz.org/

- Search Beat - large, selective dirctory organized into numerous subtopics with a friendly approach http://www.searchbeat.com/

- Top9.com - directory that ranks content based on popularity using a methodology from PC Data Online; shows the top 18 listings in each category every month http://www.top9.com/

- Web Search - well-organized, annotated collection of Web sites organized by topic and maintained by Chris Sherman of The Mining Co.

- Yahoo! - portal with one of the largest directories on the Internet but lacks reliable site evaluation, so deficient content is mixed in with the good; not an appropriate site for academic research; offers news, stock quotes, maps, free e-mail and many other services Sites that collect links to portals (and search engines) may be found under Search Engine Collections. http://www.yahoo.com/

Individual Search Engines

- Alexa Web Search - returns results powered by Google with additional information including traffic ranking, number of links to page, ownership and links to related pages of interest http://info.alexa.com/

- AllTheWeb - returns results quickly from an extremely large database; offers multimedia, Flash and FTP searches; also returns categorized topics to further focus a search http://alltheweb.com/

- AltaVista - searches Web sites and Usenet newsgroups with advanced Boolean and field search options. http://www.altavista.com/ See also:

    - Babel Fish, the AltaVista translation service. http://world.altavista.com/

- AOL Search - engine that defaults to AND logic and offers an Options template for easy search construction; has an option to view results by popularity; offers a directory based on the Open Directory http://search.aol.com/aolcom/index.jsp

- Ask Jeeves - submit questions in plain English and view suggested relevant sites (natural language engine); also offers the Open Directory ranked in order of popularity http://www.ask.com/

- Google - ranks pages by the number of links from pages ranked high by the service, including results from the Open Directory Project. Google offers a number of Services & Tools that are worth exploring, including: http://www.google.com/

    - Google Groups, a Usenet newsgroup archive http://groups.google.com/

    - Google Language Tools, for locating pages written in particular languages and located in specific countries and a translation service http://www.google.com/language_tools

    - Google Search: Unclesam, a searchable database of U.S. government Web sites (.gov and .mil) ranked by link popularity http://www.google.com/unclesam

    - Google Web Directory, which uses Google ranking technology with the

    - Google Services/Tools
      http://www.google.com/intl/en/options/

    - Google Scholar
      http://scholar.google.com/

- Guidebeam - organizes results into levels of increasingly narrow concept clusters http://guidebeam.com/

- HotBot - offers easy form-based Boolean, field, and media search options; includes its channel content with the results for searches on broad or popular terms; clusters results by presenting one hit per site http://www.hotbot.com/

- Lycos - emphasizes search results from the Open Directory and offers Web sites from the FAST Search index
  http://www.hotbot.com/

- MSN Search - derives results from a variety of sources, including the LookSmart directory, Ask Jeeves and others; applies concept matching to search statements; and offers an advanced search option with form-based field searching similar to HotBot
  http://search.msn.com/

- SearchEdu.com - service that limits results to the .edu, domain; also offers to search well-known dictionaries, encyclopedias, almanacs, etc. See also: http://www.searchedu.com/

    - SearchGov.com - .gov domain http://www.searchgov.com/

    - SearchMil.com - .mil domain http://www.searchmil.com/

- Thunderstone Website Index - search thousands of sites (vs. Web pages) from a continuously updated database
  http://search.thunderstone.com/texis/websearch/

- Teoma - returns results in three sections: popularity-ranked Web pages based on the number of same-subject pages that reference them; suggested terms to refine a search; and link collections created by topic experts http://www.teoma.com/

- WISEnut - offers a large database and a companion directory of topics related to a search http://www.wisenut.com/

Meta Search Engines

Retrieve Collated Results (limited number with duplicates removed)

- Chubba - search the Web, a dictionary/thesaurus and encyclopedia http://chubba.whatuseek.com/

- Copernic - client software that searches multiple engines and directories, removes duplicates and dead links, highlights search terms in your results, and offers a variety of search and retrieval options http://www.copernic.com/en/index.html

- Dogpile - search 20+ search engines and retrieve results by relevance or separate source engine; also presents concept clusters for viewing results organized by keywords or topics http://www.dogpile.com/

- Fazzle - searches several search services on the Web, and also offers specialty searches of downloads, images, video and other topics on the deep Web http://www.fazzle.com/

- Fossick Meta Search - searches a dynamically changing group of engines based on speed of availability, and ranks results based on search terms, popularity and other measures; offers the option to search sites from individual countries. Offers page translation using Babelfish, the AltaVista translation service. http://fossick.com/Search.htm

- GenieKnows - searches 25+ engines and directories and ranks results based on the number of sources listing a page; chooses sources to search based on consumer input; also offers searches of multimedia types http://www.genieknows.com/

- iBoogie - offers searches of the Web and multimedia, and supplies real-time concept clustering of results http://www.iboogie.tv/

- InfoGrid - offers meta and news searching; portal interface also features the Open Directory, topical InfoGrids, with additional customization of topics available in a free download http://www.infogrid.com/

- Infonetware - categorizes search results into component subtopics with options to select multiple topics for a new set of topics and a filtered results list http://www.infonetware.com/

- Ithaki - searches engines, directories, and also numerous deep Web sources http://ithaki.net/indexu.htm

- Ixquick - search engines, directories, news and MP3 files; ranks results based on top ten rankings from the source sites; allows any type of search syntax and will translate and direct your search accordingly http://ixquick.com/

- Kartoo - categorizes content into relevant concepts and sites and displays results on a graphical map; requires Flash or offers an HTML version http://www.kartoo.com/en/kartoo.html

- KillerInfo - retrieves information from the general Web and various topic-specific sources and organizes results into concept clusters based on Vivisimo's technology http://www.killerinfo.com/m/

- Mamma - retrieve results in relevancy ranked order; power search offers a user-friendly template for building a query http://www.mamma.com/

- MetaCrawler - retrieve results in relevancy ranked order; useful power search available with a template of search options http://www.metacrawler.com/

- METAEUREKA - barebones interface for an engine that sorts results by relevancy and offers a "Site info" link that returns information on the server, date last modified, size, and descriptive information if available http://www.metaeureka.com/

- Pandia Search Central - searches multiple engines and directories and also offers searches of news, books, music, videos and other specialty databases; includes a searchable version of the Open Directory http://www.pandia.com/

- ProFusion - meta engine that allows results to be sorted by relevance, title and URL; offers topic-based vertical search groups for targeted searching http://www.pandia.com/

- Query Server - offers queries of the general Web, or health, money or government sites and organizes results by concept, by site, or by both http://www.queryserver.com/

- SurfWax - offers options to see a quick view of sites in the search results list to determine relevancy and choose alternative search terms for a subsequent search from a thesaurus; offers personalization options http://www.surfwax.com/

- Turbo10 - retrieves results from multiple sources, including the deep Web, and offers sorting by speed and relevance; also offers concept clusters to organize results into keywords and subtopics http://turbo10.com/

- Verio Metasearch - retrieves the top ten hits from each source engine and directory in combined ranked order http://www.verio.com/

- Virtual Learning Resources Center - searches several high quality directories; also offers its own directory http://www.virtuallrc.com/

- Vivisimo - searches multiple engines and directories and organizes results into topical categories http://vivisimo.com/

- ZapMeta - organizes results by relevance, popularity, title, source or domain, allows users to set preferences, and features a useful advanced search interface http://www.zapmeta.com/

Retrieve Results From Each Site Separately (more comprehensive results)

- Researchville - opens multiple browser windows with search results from the originating search service; searchable categories includes news, Web guides, reference, opinion and multimedia http://www.researchville.com/

- Webtaxi - use a helpful template to build search types and access your results separately at each site; click on "Search" button for numerous meta-search options http://www.webtaxi.com/

Search Engine Collections

[Many also include directories]

- Beaucoup - compilation of 2000+ engines and directories organized by category; also offers a meta search capability http://www.beaucoup.com/

- EZ-Find at The River - search 10 search engines individually from the same site http://info.theriver.com/TheRiver/ezfind.htm

- General Internet Search Programs - compiled by BUBL Link http://bubl.ac.uk/

- MetaIQ - multi-purpose site that serves as a meta engine, a directory of searchable databases on the Web, a news source and more http://www.metaiq.com/

- Search Engine Colossus - directory of hundreds of search tools from more than 200 countries around the world http://www.searchenginecolossus.com/

- Search Engine Guide - topical collection of 3000+ engines, directories and portals http://www.searchengineguide.com/

- SearchIQ - categorized collection of engines and directories from the SearchIQ search engine informational site http://www.search.com/subjects/

Subject Based Search Engines/Directories

News Search Engines – Search Engine Watch

If you are still looking for news using "normal" search engines, stop doing it! You'll find the services below to be a much better way to search for the latest news stories from hundreds of sources on the web. These services provide exceptionally good results for current event searching, because they crawl only news sites and revisit these sites several times per day. Thus, the results are usually focused and timely.
http://searchenginewatch.com/links/article.php/2156261

News and Media Web Directories

Yahoo news and media web directories.
http://dir.yahoo.com/News_and_Media/Web_Directories/

Specialty Search Engines Deep Web

The so-called "deep" or "invisible" Web consists of content stored in searchable databases mounted on the Web. These databases usually cover a targeted topic or aspect of a topic. Search engine spiders cannot or will not index this information. There is a huge number of searchable databases on the Web. The following is a selection of sites that collect searchable databases on the Web.

- CompletePlanet - offers searchable access to thousands of databases on the "deep Web" for results that include summaries from the retrieved site http://www.completeplanet.com/

- Direct Search - large compilation of links to the search interfaces of a wide variety of research resources on the Web compiled by Gary Price http://www.freepint.com/gary/direct.htm

- Invisible-web.net - searchable directory of quality databases on the Web compiled by Chris Sherman and Gary Price as a companion to their book *The Invisible Web: Uncovering Information Sources Search Engines Can't See* http://www.invisible-web.net/

- Lycos Directory: Searchable Databases - large collection of invisible Web databases organized by topic; almost identical to The InvisibleWeb above http://dir.lycos.com/Reference/Searchable_Databases/

- ProFusion - meta engine that also offers searches of multiple "vertical search sources" on the deep Web organized into topical categories http://www.profusion.com/

- Search.Com - dozens of topic-based databases from CNet http://www.profusion.com/

FTP Search Tools

- FileWatcher.org - FTP search engine with useful advanced options for constructing searches and ordering results

Ready Reference

Internet Public Library:
http://www.ipl.org/div/subject/browse/ref00.00.00

Lakewood Public Library of Congress
http://www.lkwdpl.org/readyref/

refdesk.com
http://www.refdesk.com/

Bartleby's Reference Tools http://www.bartleby.com/reference/

Preferred Search Tools

Directory [Subject Search]	Search Engine [Keyword Search]	Multi-Engine [Meta Search]
WWW Virtual Library http://www.vlib.org/ **Refdesk.com** http://www.refdesk.com/	**AltaVista** http://www.altavista.com/ **Google** http://www.google.com	**Dogpile** http://www.dogpile.com/ **ProFusion** http://www.profusion.com/
DMOZ http://dmoz.org/ **Power Reporting** http://PowerReporting.com/	**Lasoo** http://www.lasoo.com/ **WISEnut** http://www.wisenut.com/	**InferenceFind** http://www.gocee.com/eureka/inference.htm
Internet Public Library http://www.ipl.org/ **Inomics - EconDirectory** http://www.inomics.com/cgi/econdir	**Excite** http://www.excite.com/ **Hotbot** http://hotbot.lycos.com/	**Mamma** http://www.mamma.com/ **SearchSpaniel** http://searchspaniel.com/cgi-bin/spaniel.pl
Yahoo http://www.yahoo.com/ **Librarians' Index to the Internet** http://www.lii.org/	**Teoma** http://www.teoma.com/ **Snap** http://home.nbci.com/	**Beaucoup** http://www.beaucoup.com/ **Metacrawler** http://www.metacrawler.com/index.html
Infomine http://infomine.ucr.edu/ **Specialty Search Engines** http://searchenginewatch.com/links/specialty.html	**MSN Search** http://search.msn.com/ **alltheweb** **http://www.alltheweb.com/**	**SavvySearch** http://www.search.com/ **Search Engine Colossus** http://www.searchenginecolossus.com/
Price's List of Lists http://www.specialissues.com/lol/ **Digital Librarian** http://www.digital-librarian.com/	**Ask Jeeves** http://www.ask.com/ **SurfWax** http://www.surfwax.com/	**Ixquick** http://ixquick.com/ **Vivisimo** http://vivisimo.com/
iTools! http://www.itools.com/ **The Library** http://www.geocities.com/Athens/5805/	**SearchEdu.com** http://www.searchedu.com/ **Go.com** http://go.com/	**Highbeam** http://www.highbeam.com/library/index.asp **Turbo10** http://turbo10.com/

Help in finding it!

INTERNET SEARCHER'S TOOLBOX
http://www.library.umass.edu/toolbox/tools.shtml
A Selected List of Tools to Help You Find and Use Internet Resources

LC Global Gateway
http://international.loc.gov/intldl/intldlhome.html
The Library of Congress collects materials from all over the globe. Its collections of foreign-language materials are stunning in their scope and quality. For many areas of the world, such as China, Russia, and Latin America, its collections are the finest and most comprehensive research collections outside the country of origin. "

University of Delaware Subject Guides
http://www2.lib.udel.edu/subj/

The Internet Learning Tree
http://members.tripod.com/fatih_high_school/InternetLearningTree/
The Internet Learning Center is my effort to provide a place to learn about the web and the Internet and to provide resources for both learners and trainers. It will continue to grow and add new resources.

BARE BONES 101:
http://www.sc.edu/beaufort/library/pages/bones/bones.shtml
The information contained in the following lessons is truly "bare bones," designed to get you started in the right direction with a minimum of time and effort. For more comprehensive and detailed help on searching the Web, consult our recommended list of sites in Lesson 20 at the end of this tutorial.

Best Information on the Net
http://library.sau.edu/bestinfo/alpha.htm

Search Engines and Directories from Yahoo
http://dir.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Searching_the_Web/

Eureka
http://www.gocee.com/eureka/
Eureka! is a simple and easy to use internet search engine that still provides all of the powerful and complex features of all the major search engines on the web

Reference Collection
http://library.albany.edu/reference/

The Reference Collection is maintained by the University at Albany Libraries