Search Engine

By

S. Sudesh

Faculty – IT & Systems

ICFAI Business School

Nr. GNFC InfoTower, Sarkhej – Gandhinagar Highway

Bodakdev, Ahmedabad –54

E-mail: **ssudeesh2000@yahoo.com**

SEARCH ENGINE

**"*Knowledge is of two kinds. We know a subject ourselves, or we know where we can*
find information upon it."**

--**Samuel Johnson, 1744**

ABSTRACT

Nearly 85% of people are using only E-mail service on Internet. Nothing more than that. Some other categories of people do only chatting with unknown people. But finally people do not realize that E-mail is one of the 10 major services available on Internet. " Effectively utilizing the available resources" – is the key to success. But among us how many are really know how to utilize the Internet to its fullest efficiency. Here, in this script, we will look into some of the aspects of search engine, it's working and some tips to use search engine.

KEYWORDS: Search Engine, Comparison of Search Engines, Search Strategies

0. INTERNET

The Internet is an umbrella term used to describe a collection of many separate networks, worldwide, connected together using a standardized set of communication protocols. In other words, the Internet is a public, cooperative, and self-sustaining facility accessible to hundreds of millions of people worldwide.

In addition to being a gigantic commercial and entertainment medium, the Internet is the World's biggest library, where information is available on every conceivable topic, making research projects a breeze.

1. SEARCH ENGINE

There are an estimated 800 million documents publicly available on the Internet. Document growth is, at minimum, doubling each year. Two-thirds to three-quarters of all users cite finding information as one of their primary uses of the Internet. Two-thirds to three-quarters of all users cite the inability to find the information they seek as one of their primary frustrations (second only in frustration to slowness of response).

Search engines are programs that searche documents for specified keywords and return a list of the documents where the keywords are found. They are huge databases of web page files that have been assembled automatically by machines. There are two types of search engines:

1. Individual. Individual search engines compile their own searchable databases on the web.

2. Meta.   Metasearchers do not compile databases. Instead, they search the databases of multiple sets of individual engines simultaneously.

Specialty Search Engines are those used for product searches, Competitor Intelligence, Market research, finding places, documents, News group and recent news, Search for people, Business E-mails & addresses, etc.

Search engines compile their databases by employing "spiders" or "robots" ("bots") to crawl through web space from link to link, identifying and perusing pages. Sites with no links to other pages may be missed by spiders altogether. Once the spiders get to a web site, they typically index most of the words on the publicly available pages at the site. Web page owners may submit their URLs to search engines for "crawling" and eventual inclusion in their databases. Whenever somebody searches the web using a search engine, they're asking the engine to scan its index of sites and match their keywords and phrases with those in the texts of documents within the engine's database. *It is important to remember that when you are using a search engine, you are NOT searching the entire web as it exists at this moment. You are actually searching a portion of the web, captured in a fixed index created at an earlier date.* Spiders regularly return to the web

pages they index to look for changes. When changes occur, the index is updated to reflect the new information. However, the process of updating can take a while, depending upon how often the spiders make their rounds and then, how promptly the information they gather is added to the index. Until a page has been both "spidered" AND "indexed," you won't be able to access the new information.

2. HOW DO THEY RANK WEB PAGES

In ranking web pages, search engines follow a certain set of rules. These may vary from one engine to another. Their goal, of course, is to return the most relevant pages at the top of their lists. To do this, they look for the location and frequency of keywords and phrases in the web page document and, sometimes, in the HTML META tags. They check out the title field and scan the headers and text near the top of the document. Some of them assess popularity by the number of links that are pointing to sites; the more links, the greater the popularity, i.e., value of the page.

A Web page, or document, can contain various kinds of content (as opposed to display or presentation options like sound, animation or frames), some of which is not shown when you view the document in your browser:

Ø   Title — an embedded description provided by the document designer; viewable in the titlebar (it is also used as the description of a newly created bookmark by most browsers)

Ø   Description — a type of metatag which provides a short, summary description provided by the document designer; not viewable on the actual page; this is frequently the description of the document shown on the documents listings by the search engines that use metatags.

Ø   Keywords — another type of metatag consisting of a listing of keywords that the document designer wants search engines to use to identify the document. These too, are not viewable on the actual page.

Ø   Body — the actual, viewable content of the document.

Search engines may index all or some of these content fields when storing a document on their databases. (Over time, engines have tended to index fewer words and fields.) Then, using proprietary algorithms that differ substantially from engine to engine, when a search query is evaluated by that engine, its listing of document results is presented in order of 'relevance.' Because of these differences in degree of indexing and algorithms used, the same document listed on different search engines can appear at a much higher or lower ranking (order of presentation) than on other engines.

3. SEARCH SERVICES

Effective searching requires understanding how best to utilize the features of your search services. But, Internet searching is a highly competitive, dynamic area. New search engines are cropping up continually, others are folding or being acquired, and feature sets change almost daily in order to keep pace. The major search services on the Internet are essential starting points for users seeking information. As such, they routinely are some of the most visited locations on the Web. Search services can be divided into two groups; commercial and non-commercial. Commercial search services go to the effort to catalog information on the Internet to attract attention and advertising revenues. Non-commercial services exist for many different reasons.

There are more than 2,500 search services presently on the Web. There are a dozen or more big, major Internet search services. There are also 'metasearch' services that provide a central access point to multiple of these services.

Search services on the Internet come in two main flavors: 1) 'search engines' that index words or terms in Internet documents; and 2) 'directories' that classify Web documents or locations into an arbitrary subject classification scheme or taxonomy.

4.   CREATING SEARCH STRATEGY

It's always a good idea to THINK about your search before you begin. Create a search strategy in your head by asking yourself this question:

What do I want to do?

1)   Browse?

2)   Locate a specific piece of information?

3)   Retrieve everything I can on the subject?

Your answer will determine how you conduct your search and what tools you will use.

1.   If you're browsing and trying to determine what's available in your subject area, start out by selecting a subject directory like *Yahoo!* Then, enter your search keyword(s) into one of the mega-search engines, such as *Ixquick,* just to see what's out there.

2.   If you're looking for a specific piece of information, go to a major search engine such as *Fast Search* or *Google,* or to a specialized database such as *Voice of the Shuttle* (for humanities research) or the *Bureau of the*

Census **(for statistics).**

3. If you want to retrieve everything you can on a subject, try the same search on several search engines. Also, don't forget to check resources off the web, such as books, newspapers, journals and other print reference sources.

5. CREATING A SEARCH STATEMENT

When structuring your query, keep the following tips in mind:

Ø Be specific e.g.: Tourism Australia Agents

Ø Whenever possible, use nouns and objects as keywords e.g. : Tourism India

Ø Put most important terms first in your keyword list; to ensure that they will be searched, put a + sign in front of each one e.g.: +hybrid +electric +gas +vehicles

Ø Use at least three keywords in your query e.g.: vitamins drugs interaction

Ø Combine keywords, whenever possible, into phrases e.g.: "search engine tutorial"

Ø Avoid common words, e.g., green, unless they're part of a phrase e.g.: "green tea"

Ø Think about words you'd expect to find in the body of the page, and use them as keywords e.g.: Boarding Tourism India

Ø Write down your search statement and revise it before you type it into a search engine query box e.g.: "college savings plans" "Section 529" +state +sponsored

6. SEARCHING MATH

Quick Tips for Boolean Searches:

In Boolean searches, always enclose OR statements in parentheses.e.g.: "financial aid" AND (college OR university)

Always use CAPS when typing Boolean operators in your search statements. Most engines require that the operators (AND, OR, AND NOT/NOT) be capitalized.

Standard Syntax	Meaning	Alternative Syntax	If Not Supported

AND	both required	+	ignored
OR	either required	blank	all support
AND NOT	exclude following	-, BUT NOT, NOT	ignored
NEAR	required within set word distance	ADJ	ignored
BEFORE	first required before within distance		ignored
AFTER	first required after within distance		ignored
( )	Nesting		ignored
" "	treat as phrase	checkbox option	

7. QUERY CONCEPT

One of the bigger mistakes you can make in preparing a query is not providing enough keywords. On average, most users submit 1.5 keywords per query. This number is insufficient to accurately find the information you are seeking. Thus, a central task in query formulation is for you to identify sufficient number of appropriate keywords.

If you are new to searching, the first task when formulating a search is to write down what information you are seeking. Before doing a search, it is important to bind your topic as completely yet succinctly as possible. Formulating a query is akin to solving a mystery. Some pieces of information are available, but if sufficient information were available the answer would be known and there would be no need to seek more. This is the essence of a query: missing information. It is up to you, the searcher, to define your snare — the query (quarry? pun intended) — sufficiently to trap that missing information and solve the mystery.

It is useful to begin by listing what you do know according to these standard categories. Jan lists these for the mystery bird:

Ø **WHO / WHAT?** — gray and white bird, about the size of a crow; yellow and black beak

Ø **WHERE?** — downtown office buildings in the City of Minneapolis

Ø **WHEN?** — daylight in the Spring

Ø **HOW?** — fast flyer, hunting pigeons (?) as prey

Ø **WHY?** — hunting bird; why never seen before? blown off course? is it migrating?

Of course, not all of these five categories will apply to a given query, and the specifics will obviously vary for your desired topic. There are many common words in these responses that are prepositions, conjunctions or common verbs. These include: and, about, the, of, a, in, as, if, not, why, never, before, is and it. These common words are referred to as "stoplist" words: they are essential to the connecting tissue in language, but they are filler in any search request. All search engines ignore them because they have minimal information value and are found so commonly in all languages. Search services include of the order of 600 of these

common words in their "stoplists"; if you use them in a query, they are ignored. Therefore, you should ignore them as well.

Almost without exception, the central keywords in your queries will be nouns. Though sometimes adverbs and adjectives can help refine your search, the key pivot point is a noun, or series of nouns. Why is this?

The most precise terms we have in language are for tangible, concrete "things" or objects. Actions and modifiers are very diverse, easily substitutable, and generally not universally applied in any given description.

8.   KEYWORDS – THE BEST WAY TO OPTIMZE THE SEARCH

The focus on the correct way of finding out the keywords is very important functions in searching. Single word keywords tend to be hyper-competitive. Suppose your company organizes packaged tours to Australia. A search for "tourism" or "travelling" in any search engine will probably generate hundreds of thousands of pages. Due to sheer number of pages that single word searches can throw up, most search engine users have realized that they can get more relevant pages if they search for phrases rather than individual words. Statistical research has shown that most people are now searching for 2 or 3 word phrases rather than for single words. Single word keywords won't get you targeted traffic. When people search for "tourism", they are not necessarily looking for tourist destinations in Australia - they may be interested in any other country of the world. Try to think of keywords, which apply to the geographic area, that your product or service is designed to serve. Like the following:

1)   Tourism in Australia

2)   travel to Australia

3)   traveling in Australia

4)   travel agencies in Australia

5)   traveling agencies in Australia

6)   Australian travel agencies

9.   GUIDE TO EFFECTIVE SEARCHING ON THE INERNET

1.   Always keep in mind the who, what, where, how and why in formulating the query.

2.   Never use articles, pronouns, conjunctions or propositions – the connecting tissue in language – in queries.

3.   The keywords in your queries will most often be nouns – and then likely no more than 6 or 8 of them.

4.   Truncation or word stemming keeps your keyword count down and makes for simpler queries.

5.   You can use synonyms both to find the right "level" for your query subject and to ensure proper coverage.

6.   Always look for natural phrases in your query concepts – they are one of the most powerful weapons available.

7.   AND can be used more frequently as Boolean operator. Use OR to string together synonyms; be careful about mixing it in with AND. Use NEAR as an alternative to phrases and an improvement to AND, but only when you know the concepts are closely linked. AND NOT is a powerful operator, and proper care must be taken. A single instance will cause a document to be excluded.

8.   Don't assume evaluation order. Specify the order you want by using parentheses.

9.   Avoid complicated nesting with too many parentheses; they can sometimes give results you did not intend.

10.   Use multiple search services for an important query.

11.   Use search engines with full-text indexing and Boolean support for the most demanding queries.

12.   Avoid misspelling, redundant terms, ignored terms and special characters, improper Boolean and complicated nesting.

13.   Use the plus (+) and minus (-) signs in front of words to force their inclusion and/or exclusion in searches.

14.   Use double quotation marks (" ") around phrases to ensure they are searched exactly as is, with the words side by side in the same order.

15.   Type keywords and phrases in lower case to find both lower and upper case versions. Typing capital letters will usually return only an exact match.

10.   TROUBLE SHOOTING. WHAT TO DO IF….

10.1   Your Search returns a "ZILLION" Documents

You probably typed only one term, and it was pretty common. Think of some synonyms. Try adding at least two more specific terms to your string.

10.2   Your Search returns too few Documents

You're probably searching in the wrong place or your search is too narrow. Maybe you didn't configure your search correctly. Maybe the information you seek isn't on the web. Try omitting some of your search terms. Try your search on another engine, metasearcher, directory, people search, or specialty resource. Ask for help.

10.3   Your Search returns a "404 -- File Not Found" Message

This message tells you that the file you seek has been moved, removed, or renamed. Go back to the search engine and do a phrase search or a field search on the title. Try shortening the URL to see if the file might still be on the same server. Try your search on *Google,* which maintains cached copies of pages.

10.4   Your Search returns a "Server does not have a DNS entry" Message

This message tells you that your browser can't locate the server (i.e. the computer that hosts the web page). It could mean that the network is busy or that the server has been removed or taken down for maintenance. Check your spelling and try again later.

10.5   Your Search returns a "Server Error" Or "Server Is Busy" Message

The server you are attempting to contact may be offline, may have crashed, or may be very busy. Try again later.

10.6   You can't find the Home Page for a Well-Known Product or Organization

Try guessing, experimenting with different top-level domain names by using the name, brief name, or the acronym. Many organizations use one of these in their URL.

EXAMPLE: www.dell.com (for Dell Computers); www.redcross.org (for the American Red Cross)

Note: This doesn't always work. You won't find the American Medical Association this way. The "ama" acronym was grabbed first by the American Marketing Association, so their URL is: www.ama.org, while the American Medical Association had to settle for www.ama-assn.org. However, it is worth a try.

11.   CONCLUSION - SEARCH ENGINES OUTPERFORM

Your ability to find the information you seek on the Internet is a function of how precise your queries are and how effectively you use search services. Poor queries return poor results; good queries return great results. Contrary to the hype surrounding "intelligent agents" and "artificial intelligence," the fact remains that search

results are only as good as the query you pose and how you search. There is no silver bullet.

In the words of Clifford Lynch in a recent Scientific American article:

"One sometimes hears the Internet characterized as the world's library for the digital age. This description does not stand up under even casual examination. The Internet - and particularly its collection of multimedia resources known as the World Wide Web - was not designed to support the organized publication and retrieval of information, as libraries are. But if the Internet is to grow and thrive as a new means of communication, something very much like traditional library services will be needed to organize, access and preserve networked information. Even then, the Net will not resemble a traditional library, because its contents are more widely dispersed than a standard collection. Consequently, the librarian's classification and selection skills must be complemented by the computer scientists' ability to automate the task of indexing and storing information. Only a synthesis of the differing perspectives brought forth by both professions will allow this new medium to remain viable."

Personally I do not see the demise or "death" of search engines. Major search engines will continue to be one of the most important first access points to the Internet. The sheer growth and chaos of the Internet assures this. But there will also be twin, divergent forces toward consolidation on the one hand and specialization on the other.

The first-generation of Internet search services are facing an untenable impasse. Sheer document volumes - approaching 1 billion in the near future - doesn't bode well for either traditional search engines or search directories. The largest search engines today approach nearly 150 million documents. Results are overwhelming and force users to study tutorials such as this one in order to query for meaningful results. This problem will only get worse. Directories, however, with their (supposedly) higher-quality content selected by human reviewers are also falling more rapidly behind. For example, perhaps 1.5 million new documents are being added daily to the Web, surely a volume that can not be hand-screened by humans at acceptable cost. And, spot checks of Yahoo! already indicate that 20-25% of existing links are no longer in existence or out of date.

REFERENCES

1.   An E-Book on "7 Secrets to Unlimited Traffic"- by Allan Rayne published By Internet Warriors Inc.

2.   A guide for E-commerce and EDI – published by American Petroleum Institute

3.   http://www.whatis.com

4.   http://www.webopedia.com

5.   http://www.sc.edu/beaufort/library/

6.   http://wcp.oclc.org/

7.   http://www.1st searchranking.com – an E- Newsletter titled " How should you optimize your site for "? – The frames and Search Engines " – article by Sumantra Roy

8.   http://webreference.com

9.   An E-book on " The Internet World" – by Kim Eyer – Eyer Station Web Development

10. http://www.searchenginewatch.com

11. http://www.brightplanet.com

12. http://www.thefrontpage.com

13. http://www.internetstats.com

14. http://www.cyberstats.internet.com

15. http://www.statmarket.com

16. http://www.eyescream.com/yahootop200.html

BRIEF BIOGRAPHY OF AUTHOR

S Sudeesh holds B.E in Mechanical Engineering from Manonmaniam Sundaranar University and M B A with specialisation in IT & Systems from Madurai Kamaraj University. He has worked as a Web Designer for about one year. Currently he is working as an IT Faculty in ICFAI Business School, Ahmedabad. His main interest lies in Internet, Web Designing and System Analysis and Design