SEARCH ENGINES IN-DEPTH

Jadranka Stojanovski
Ruder Boskovic Institute
jadranka.stojanovski@irb.hr
http://library.irb.hr

Introduction

"Which of the following has helped you find your way to the Web sites you use?" was the question given to the unknown number of Internet users. According this study provided by Forrester Research search engines remain the leading way users in the United Kingdom find web sites. The "*UK Internet User Monitor*" survey found 81 percent of users said that search engines helped them find the web sites they use, up from 67 percent in 1999. The next most popular source was by following links, a method used by 59 percent of those surveyed (Table 1).

Source	1999	2000	Change
Search engine	67%	81%	14%
Link from another site	39%	59%	20%
Viral marketing	28%	56%	28%
Television	16%	48%	32%
Guessed the address	22%	41%	19%
Online advertising	10%	20%	10%
Radio	6%	19%	13%
Direct mail	5%	10%	5%

Table 1. UK Internet User Monitor, May 2000

Search engines seem to be everywhere. The majority of the web public use search engines to find information at least weekly, if not daily. The availability of free search engines that index words from millions of web pages has been one of the driving forces of the web. They are changing and growing rapidly, and no one knows for sure in which direction they are going.

This article will first provide an overview of the web space and search engine features in general. This will be followed by a detailed survey of the seven main search engines.

Web space

It is possible to search only a part of the web space called "visible" web that contains mainly of the static web pages. Static web pages are manually produced, they offer a generic information and most of them are indexable. On the other side dynamic web pages are computer generated, offer customised information and are not indexable.

The “invisible” web contains pages with authorisation requirements, pages excluded from indexing using the robot exclusion meta-tag, badly designed pages with frames, non-HTML pages and dynamically generated web pages. The “visible” web contains static web pages, “publicly indexable” pages (Lawrence and Gilles, 1999).

The approximate size of the “visible” web is growing very fast:

December 1997	320 M pages
February 1999	800 M pages
February 2000	>1,2 G pages
July 10, 2000	2.1 G (7 M pages per day growth)

Search engines

When using a search engine, the user is searching a database of indexed web sites. All search engines have three primary components:

- **“spider”** - programmes that examines web site;
- **index**/database (title, URL, metatags, whole page is indexed) - LookSmart, Inktomi, etc.
- **retrieval software** (first step is matching which is similar, what makes a great difference is second step - ranking - different search engines are using different algorithms).

Relevancy ranking and the way how it is calculated is the "top secret" by most of the search engines. Most search engines use the location and frequency of keywords on a web page as the basis of ranking it in response to a query. The exact mechanism is slightly different for each engine. In addition to location and frequency, some search engines base relevancy ranking algorithm on the popularity by number of links or by number of "clicks" on the user side. Some search engines that support the meta description and keywords tag will also give pages and extra boost if search terms appear in these areas. All approaches have problems with cyberspamming. The sites that attempt to do “a simple spam” (as “stacking” or “stuffing” words on a page) are penalised by all major search engines.

According Search engine watch at http://www.searchenginewatch.com the biggest search engine is Google (Figure 2).

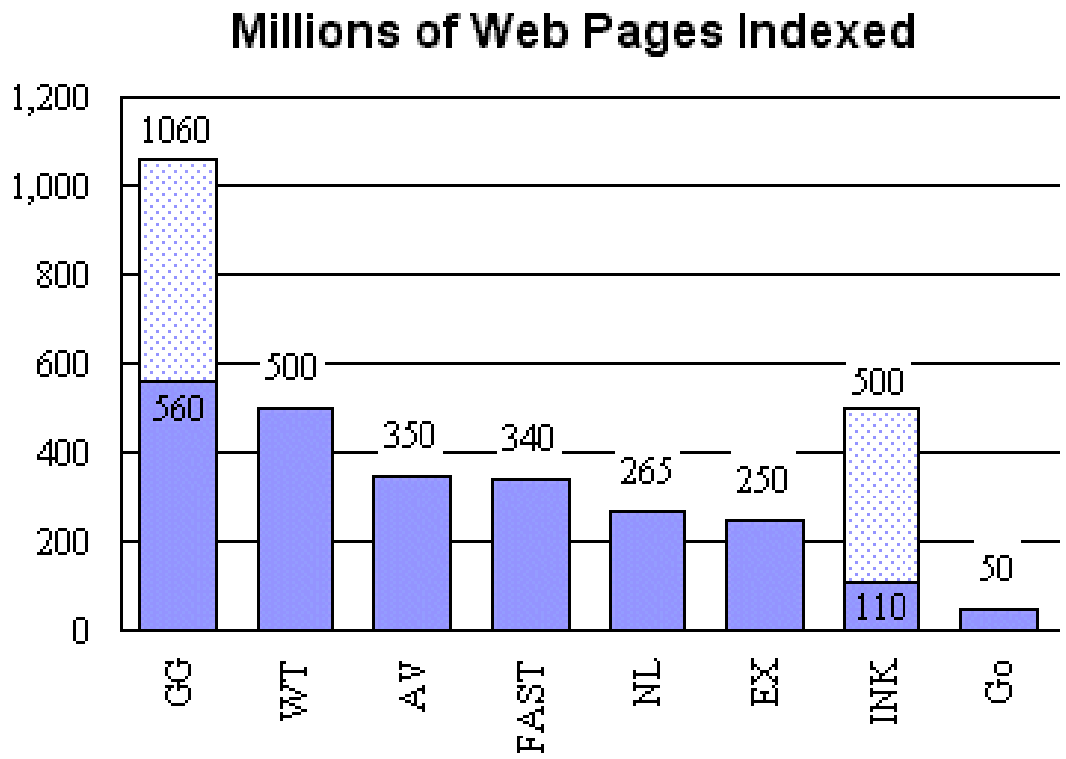

Figure 2. Sizes are as reported by each search engine and as of June 6, 2000.
GG=Google, WT=WebTop.com, AV=AltaVista, FAST=FAST, NL=Northern Light, EX=Excite, INK=Inktomi, Go=Go (Infoseek)

Simple research provided by the author of this paper on September 25, 2000 (query was Croatian word "korisnici" = “users” in English) shows different data (Table 1):

SEARCH ENGINE	SIMPLE SEARCH	ADVANCED SEARCH
Lycos	12,988	12,837
Fast	12,982	12,989
Google	8,060	-
Northern Light	7,172	7,172
Alta Vista	7,086	7,488
HotBot	5,100	5,100
Excite	about 145	about 145
Snap	66	66
Go (Infoseek)	1,079	-
MSN Search	1,921	-
Web Crawler	22	-

Table 1. Hits by simple query

Lycos and Fast are obviously using the same (Fast) database that is the biggest one at the moment.

The general search engine features are:

1. options
2. size / number of results
3. speed
4. percentage of relevant hits
5. are the search results sorted by relevancy
6. freshness
7. low percentage of *dead link*s
8. display (summary, date, URL…)
9. logic of the simple and advanced search
10. help
11. added value

Alta Vista

A very good comprehensive, fast and powerful search engine. AltaVista provides a lot of search construction options for sophisticated searchers, and therefore has long enjoyed favour by information professionals. Besides traditional Boolean search options, AltaVista has many field limits, and also has an interesting forced phrase searching feature.

AltaVista translates its results into several different languages, which can come in handy when you run across a page in a language you don't understand. The translation is not precise, but it can be good enough to get the clue of what is on the page. AltaVista also has a special image and media finder. The image finder returns a result list complete with image thumbnails.

PROS	**CONS**
powerful search features	inconsistent results
size	only 10 hits per page
translation service (only 5k)	no sorting options
image search	
high quality index	
international approach	
intuitive interface	

Search features and results:

? ? OPERATORS: + - “”, Boolean operators: AND, OR, AND NOT, proximity operator NEAR

? ? LIMITS: language, time period

? ? **FIELD SEARCHING:** anchor: applet: domain: host: image: link: text: title: URL: like:

? ? TRUNCATION: right and internal (* for 0-5 characters)

? ? CASE SENSITIVE

? ? natural language search statement (Ask Jeeves)

? ? no stopwords in Advanced Search!

? ? DISPLAY: + language, date and size (bytes)
options: Translate, Company Fact Sheet, More pages from this site, Related Pages

? ? SORTING: Advanced Search - “ranking keywords”

? ? LIMITS: no refine option, link to “related pages”

Search inconsistencies:

? ? **time-outs**: May stop processing a search and provide partial results to expedite. Repeating the exact same search several times in a row may find additional results

? ? **Counting**: Can't count accurately. The numbers present are often estimates and can change greatly especially with complex search statements that use multiple fields.

? ? different number of hits by Simple and Advanced Search (identical search statement!)

? ? inconsistency with diacriticals
+éléphant -elephant

? ? putting limits on language can rise number of hits!

? ? field search by title don’t find all documents

Excite

Excite has a popular, medium size database, also used by Webcrawler. Excites features, its personalised page with news and portfolio tracking are tops. Unfortunately its search engine relevancy could be improved.

Excite provides a few more options than Google for creating a detailed search, but the beauty of Excite is its concept searching. If you type a word in the search box, not only do you search for that word itself, but also forms of the word, synonyms of the word, and other words that are related to that word. User doing more complex search should be very careful: if he/she use a Boolean operator in the search string, it will turn the search into an exact keyword search, eliminating any concept searching for word variations, synonyms, etc.

PROS	**CONS**
offers sophisticated personalization My Excite	middle size database
“more like this” feature	no truncation controlled by user
choosing “site”	no international approach

very relevant results for very popular queries	Boolean must be typed in all CAPS
News Search for web newspapers access	towards user-consumer
numerous reference databases (dictionaries, almanac, encyclopaedia)	

Search features and results:

? ? OPERATORS: + - “”, Boolean operators: AND, OR, NOT (default is OR) - **must be in uppercase**!

? ? Advanced search: CAN contain, MUST contain, MUST NOT contain, SHOULD contain

? ? LIMITS: Advanced search: language, country and domain limits

? ? FIELD SEARCH: Title, URL, Image, Java applets, Links

? ? DISPLAY: + relevancy level (per cent)
(Titles only, View by web site)

? ? SORTING: relevancy - first ten web pages, first ten directories

? ? REFINE:
- additional keywords (“add to your search”)
- link “Search for **more documents like this** on

Fast

Accordingly his name Fast is very fast search engine having in mind one of the largest database. Results seem to be listed in order of how many keywords that match. Seems to index all the keywords in a document. The main problem with this search engine is a redundancy: if a site has multiple pages on the same topic, it returns them all requiring the searcher to sift through many redundant listings.

PROS	**CONS**
size	not so fresh index
speed	lack of command Boolean searching, truncation, and many field searches
don’t have stopwords	
100 hits per page in advanced search	
relevance ranking algorithm may be significantly more effective than others	

Search features and results:

? ? default is “all of the words”, “any of the words” and “exact phrase” options

? ? **FIELD SEARCH:** title, link name, URL, link to URL, with options Should Include, Must Include and Must Not Include

? ? no limits!!

? ? DISPLAY: +

? ? SORTING: relevancy (efficient relevancy algorithm!)

? ? REFINE: no “refine” option

Google

Google is a popularity engine like Direct Hit. Despite his name that sounds like a baby babbling, Google claims to use a complicated mathematical analysis based on hyperlinks on the web, to return high-quality search results so you don't have to sift through junk. Google gives an excerpt of the text that matches the query--with the search terms in bold. Google presently has indexed more than 1 billion web pages, the most number of any search engine. They use Open Directory for directory listings.

This is the simplest search engine to use that provides very few options for searchers to construct a detailed search. In fact, just about the only options are a minus sign to exclude terms and parentheses to force terms to be searched as a phrase.

Google's results list provides a similar pages feature that allows the computer to construct automatically a new search for pages similar to a specified page - a good option if searcher find one good page and want more like it.

PROS	**CONS**
simple interface	limited searching capabilities
fast!!!	link search must be exact
often has excellent results	
stopwords search could be forced	
can go to the web, or a "cached" copy, which Google stored when it retrieved the page!!!	
Option of 10, 30, or 100 records per page of results	

Search features and results:

? ? default is AND

? ? FIELD SEARCHING: link: (find pages which contain certain link) related: (calls Google Scout which find similar linking pattern for given URL)

? ? LIMITS: not supported (even language)

? ? DISPLAY: + “cache”

? ? SORTING: relevancy by number of linked pages

? ? REFINE: Google scout will find similar (?) pages

? ? 10,30,100 results per page

Search inconsistencies

? ? link and minus combination doesn’t work, example:

? ? *link:www.whitehouse.gov -clinton*
gives the same score as

link:www.whitehouse.gov

HotBot

HotBot uses both the Open Directory Project and Inktomi's databases. Like most search engines it suffers when it comes to finding relevant listings. While the Open Directory Project and Inktomi represent two of the largest databases their size doesn't mean the algorithms used to sort the data can guarantee accurate results. Still HotBot it is one of the better individual search engines and includes powerful advanced search options.

The best way to use HotBot is to go directly to the Advanced Search page, which provides many more search options. HotBot's form interface is a favorite among many users. It allows searcher to build a sophisticated search without having to remember operators and limits.

HotBot results employ direct hit technology, which provides a list of the top ten most popular links for any given search. The theory behind direct hit is that the sites that people go to most based on a given search are also likely to be the most relevant sites for that search.

PROS	**CONS**
double system: Inktomi and Direct Hit database and Open Directory	database size shrunk
Interface is user-friendly	has stop-words
Many field searching options	consumer oriented
Text Mode - Much faster to use at www.hotbot.com/text	
New index every two weeks?	

Search features and results:

? ? "all of the words", "any of the words" , "exact phrase", "the person" and "links to this URL"

? ? Operators: +,-,"", Boolean operators: AND, OR, NOT

? ? FIELD SEARCHING: title: domain: depth: feature:(frame, image, applet…), likdomain: outgoingurlext: scriptlanguage: **after: before:** within:(3/months)

? ? LIMITS: +time period, and media type (Javascript, image, video); Adv. S. offers limit on domain (.edu, .hr), geographical location and page depth

? ? TRUNCATION: left, right and inside; **"stemming"** - Adv. S. -grammatical word variants (plural, singular, tense)

? ? CASE SENSITIVE searching could be forced!

? ? DISPLAY: **three options**: full description, short description and URL only
full description includes document title, URL, first few document lines (ONLY), relevance score, date
10, 25, 50, 75, 1000 records at a time

? ? REFINE:
- on given set
- "This site only"

Search inconsistencies

? ? Counting: same word - different number of hits (all of the words, any of the words, exact phrase, Boolean phrase)

? ? Stop Words**:** HotBot and the other Inktomi databases have an extensive, dynamic stop word list. Many common words and numbers will not be searched. The list changes as the frequency of terms in the database change. When a stop word is in a phrase, it may not be obvious that the whole phrase is not being searched. Example : “online review”
However, online was a stop word that day, so the search was actua
the single word search of world

Go.com

Go.com uses the search capabilities of the former Infoseek, and still employs many of the same features. Provides quality results thanks to its ESP search algorithm. It also has a large human compiled directory.

Go.com supports Boolean searching and allows searching with many limits. Additionally, the advanced search provides a variety of specialised search engines for finding specific information. Go.com provides the opportunity to search in levels, by performing one broad search first, then narrowing it down by searching again within the given results, rather than searching the Web all over again. Its translation option allows the translation of any page.

PROS	**CONS**
sorts by site and date	less powerful search features
rich supplemental resources	small database
rich portal content	consumer oriented
additional reference databases	
translation service - eng, fra, ger, ita, spa, por - larger documents than Alta Vista	
refine options good for diacriticals	

Search features

? ? Operators: +.-."", (Boolean operators): NO (default is OR)

? ? LIMITS: domain, geographical location, Infoseek subject category

? ? **FIELD SEARCHING**: title:, URL:, link:, site:

? ? TRUNCATION: not supported! (does automatic “intelligent pluralization”)

? ? CASE SENSITIVE!

? ? NO STOPWORDS!

? ? DISPLAY: + relevance score, date and size (bytes)
50? (25) records at a time in Adv. S.

? ? SORTING: by relevance score and by site (“ungroup results” can be activated)

? ? REFINE:
“Find similar pages” - very useful
“Search within results” - additional search statement

Lycos

Bringing together data from FAST, Direct Search and the Open Project Directory, Lycos also supports Boolean searching, and by far has the most extensive options for proximity searching as any search engine on the Web. Lycos, like Go.com, provides the option of searching by levels, where you can search within a previous set of results. It will also offer suggested searches following your initial search. Lycos' advanced search provides a variety of specialised search engines for locating specific information. An automated tracking feature at Lycos allows users to register and have their searches updated automatically. Lycos' results page offers a popular links region where the most popular links for certain searches will be distinguished from regular results.

PROS	**CONS**
long tradition conglomeration of databases, online services and Internet properties	small database in regular Lycos slow to refresh the database
popular Top 5% sites	
large database in Lycos Pro (Fast)	
advanced features in Lycos Pro	
extensive portal content	

Search features and results:

? ? default is Boolean AND, Advanced search interface has AND, OR and NOT

? ? (ADJ, NEAR, FAR, BEFORE) until 2000 - not any more!

? ? pull-down menus with “all of the words” , “any of the words” and “exact phrase”

? ? Lycos Pro does not have stop words!

? ? FIELD SEARCHING: Title, URL, Host/Domain, Your URL, Only this host, Exclude this Host

? ? DISPLAY: search results grouped by site (no option to ungroup the results)

? ? REFINE: options for refining the search or searching within hits

Northern Light

Northern Light uses its own proprietary database covering 220 million web pages and 20 million articles. When a searcher conduct a search it puts a column of folders on the left of related topics. Below each site listed, Northern Light puts a folder of additional pages from that site, eliminating the appearance of multiple and or redundant listings from the same site.

Northern Light runs of a limited database covering only a slice of net. Results are interspersed with articles from Northern Light's special collection, with a fee from $1 to $4.

PROS	CONS
parallel with their large database of web pages NL offers Special Collection	not very relevant results
large database	only 10 hits at time
reach search feature (proximity,Boolean, truncation, Power Search, Business Search)	
different Help levels - general, search, power search	
Current News (the most recent - two weeks)	

Search features and results:

? ? Operators: +, -, "", Boolean operators: default AND, (AND, OR, AND NOT)

? ? FIELD SEARCHING: title: URL: text: (pub: company: ticker: recid:> Spec.Coll.)

? ? LIMITS: + date, document type (Company information, Educational material, Press release, Product review), publication, subjects (Arts, Business, Education, Travel), domain and industry (Retail, Insurance, Telecommunications and Economics)

? ? TRUNCATION: unlimited truncation symbol * and % for single character - both for internal or end truncation

? ? CASE SENSITIVITY: mixed case - hits that make an exact case match, all uppercase will match both (lower and upper case)

? ? no stopwords!

? ? DISPLAY: language, date, site type, document type

? ? SORTING: by date in Power Search

? ? 10 records at a time can be switched to 25 (&us=25 to the end of the URL after search)

? ? REFINE: Custom Search Folders will organise the full set of search results into a subject, source, document type or language folder; options for refining the search or searching within hits

? ? DISPLAY: search results grouped by site (no option to ungroup the results)

Search inconsistencies

? ? Truncation inconsistencies have been noted

Tips for formulating searches

✍✍ become familiar with **several** search engines

✍✍ study search engine **instructions** for simple and advanced techniques

✍✍ check to see whether automatic stemming or automatic **truncation** is used

- select terms and consider unique words, phrases, and **synonyms**, don't use very common words
- most search engines permit Boolean searching in various forms:
- **"and" "or"** or **"not"**
- pull down menus - "all the words"
- plus **(+)** or minus **(-)**
- Boolean searches override relevancy ranking
- refine your search with field searching
- Time periods
- Geographic Location
- Media types
- URLs
- Review these Feature Comparison Charts for further help
- www.infopeople.org/src/chart.html
- www.curtin.edu.au/curtin/library/staffpages/gwpersonal/searchtut/index.html
- www.lib.berkeley.edu/TeachingLib/Guides/Internet
- AND DON'T FORGET YOUR LIBRARY!

Future development

With thousands of millions of web pages nowadays, it is getting harder and harder for search engines to keep up with a demand for accuracy. What are the search engines likely to do about it in the near future? The way of development in search engines is that they can adapt to user's needs. In the future a sophisticated robot will search for concept, no words. A spider will be "trained" to pick out only high-quality web pages, to make their selection more precise. Probably indexing tools will have to perform more sophisticated analysis of the web page they are indexing. They will measure how many links are in the page, internal as well as external, how much text is included, how many graphic images are animated. Implementation of standards will be crucial demand and movement from HTML to XML is taking place already. Librarians and their experiences could be of great importance in the future classification and cataloguing web pages process, according adopted metadata standards.

Conclusion

As already mentioned search engines are inconsistent, inaccurate, unreliable, sometimes inaccessible, incomplete, out of date, don't live up to their advertisements, error prone, and provide millions of irrelevant items. If we take a look at the Lycos top ten most popular searches, which are as follows:

- Pokemon
- Britney Spears
- Dragonball
- The WWF

- Eminem
- Tattoos
- Napster
- Pam Anderson
- Mother's Day
- Victoria's Secret

we can realise that all are about entertainment, games, and sex. Who wants to speak about relevancy ranking any more?

References

Beyond the hype: Dissecting AltaVistas claims. (1999, November 1). *The Search Engine Report.* http://www.searchenginewatch.com/sereport/99/11-avclaims.html

Botluk, D. (2000, Sept) Update to Search Engines compared. http://www.llrx.com/features/engine3.htm

Habib, D. P. & Balliot, R. L. (1999, September 19), How to search the World Wide Web: A tutorial for beginners and non-experts. http://204.17.98.73/midlib/tutor.htm

Hasebrook, J. P. (1999, April-June), Searching the Web without losing your mind: Traveling the knowledge space. *WebNet Journal,* 1(1).

Notess, G. (1999a), Search engine showdown. http://www.notess.com/search/stats/

Notess, G. (1999b), Multiple search engines, search engine showdown. http://www.notess.com/search/multi/

Repman, J. & Carlson, R.D. (1999, May), Surviving the storm: Using metasearch engines effectively, *Computers in Libraries,* 19(5).

Search engine shootout. http://home.cnet.com/category/topic/

TargetedListings (1999), Understanding how search engines rank Web pages. http://www.targetedlistings.com/tips6.html

WebSideStory. http://www.websidestory.com/content.cfm?Pg=3&PR=25

Wiens, B (2000, Sept 15), *Websearching*. http://www.benwiens.com/websearch.html).

Whos the biggest of them all? (1999, November 1). *The Search Engine Report.*