Journal of Novel Applied Sciences
Available online at www.jnasci.org
©2014 JNAS Journal-2014-3-7/724-728
ISSN 2322-5149 ©2014 JNAS

INTRODUCTION TO WEB SEARCH ENGINES

Mohammad Reza pourmir

Computer Engineering Department, Faculty of Engineering, University of zabol, Zabol, Iran

Corresponding author**:** Mohammad Reza Pourmir

ABSTRACTA: web search engine is a program designed to help find information stored on the World Wide Web. Search engine indexes are similar, but vastly more complex that back of the book indexes. The quality of the indexes, and how the engines use the information they contain, is what makes or breaks the quality of search results. The vast majority of users navigate the Web via search engines. Yet searching can be the most frustrating activity using a browser. Type in a keyword or phrase, and we're likely to get thousands of responses, only a handful of which are close to what we looking for. And those are located on secondary search pages, only after a set of sponsored links or paid-for-position advertisements. Still, search engines have come a long way in the past few years. Although most of us will never want to become experts on web indexing, knowing even a little bit about how they're built and used can vastly improve our searching skills. This paper gives a brief introduction to the web search engines, architecture, the work process, challenges faced by search engines, discuss various searching strategies, and the recent technologies in the web mining field.

Keywords**:** Introduction, Web Search, Engines.

INTRODUCTION

A web search engine is a program designed to help find information stored on the World Wide Web. The search engine allows one to ask for content meeting specific criteria (typically those containing a given word or phrase) and retrieves a list of references that match those criteria.

Search engines are essentially massive full-text indexes of web pages. They use regularly updated indexes to operate quickly and efficiently. The quality of the indexes, and how the engines use the information they contain, is what makes -- or breaks -- the quality of search results. Search engine indexes are similar, but vastly more complex that back-of-the-book indexes. Knowing even a little bit about how they're built and used can vastly improve the searching skills.

Other kinds of search engine are enterprise search engines, which search on intranets, personal search engines, which search individual personal computers, and mobile search engines. Some search engines also mine data available in newsgroups, large databases, or open directories like DMOZ.org. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically. Most web sites which call themselves search engines are actually front ends to search engines owned by other companies.

History:

The first Web search engine was "Wandex", a now-defunct index collected by the World Wide Web Wanderer, a web crawler developed by Matthew Gray at MIT in 1993. Another very early search engine, Aliweb, also appeared in 1993, and still runs today. The first "full text" crawler-based search engine was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any web page, which became the standard for all major search engines since. It was also the first one to be widely known by the public. Also in 1994 Lycos (which started at Carnegie Mellon University) came out, and became a major commercial endeavor.

Soon after, many search engines appeared and vied for popularity. These included Excite, Infoseek, Inktomi, Northern Light, and AltaVista. In some ways, they competed with popular directories such as Yahoo!. Later, the directories integrated or added on search engine technology for greater functionality.

Search engines were also known as some of the brightest stars in the Internet investing frenzy that occurred in the late 1990s. Several companies entered the market spectacularly, recording record gains during their initial public offerings. Some have taken down their public search engine, and are marketing enterprise-only editions, such as Northern Light.

Before the advent of the Web, there were search engines for other protocols or uses, such as the Archie search engine for anonymous FTP sites and the Veronica search engine for the Gopher protocol. More recently search engines are also coming online which utilise XML or RSS. This allows the search engine to efficiently index data about websites without requiring a complicated crawler. The websites simply provide an xml feed which the search engine indexes. XML feeds are increasingly provided automatically by weblogs or blogs. Examples of this type of search engine are feedster, with niche examples such as LjFind Search providing search services for Livejournal blogs.

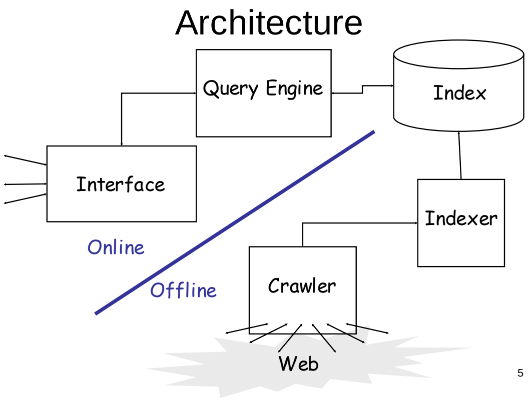
Timeline		
Year	**Engine**	**Event**
1993	Aliweb	Launch
1994	WebCrawler	Launch
	Infoseek	Launch
	Lycos	Launch
1995	AltaVista	Launch (part of DEC)
	Excite	Launch
1996	Dogpile	Launch
	Inktomi	Founded
	Ask Jeeves	Founded
1997	Northern Light	Launch
1998	Google	Launch
1999	AlltheWeb	Launch
2000	Teoma	Founded
2003	Objects Search	Launch
2004	Yahoo! Search	Final launch (first original results)
	MSN Search	Beta launch
2005	MSN Search	Final launch
	FinQoo Meta Search	
	Quaero	Final launch
2006	Kosmix	Beta launch

THE WORKING PROCESS

The Architecture:

Search Engines for the general web do not really search the World Wide Web directly. Each one searches a database of the full text of web pages selected from the billions of web pages out there residing on servers. When we search the web using a search engine, you are always searching a somewhat stale copy of the real web page. When you click on links provided in a search engine's search results, you retrieve from the server the current version of the page.

The outline process of web search engine architecture is:

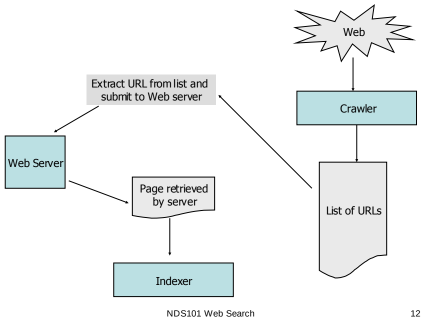


Working Process:

To find information on the hundreds of millions of Web pages that exist, a search engine employs special software robots, called spiders or Crawler, to build lists of the words found on Web sites. When a spider is building its lists, the process is called Web crawling. They find the pages for potential inclusion by following the links in the pages they already have in their database (i.e., already "know about"). They cannot think or type a URL or use judgment to "decide" to go look something up and see what's on the web about it.

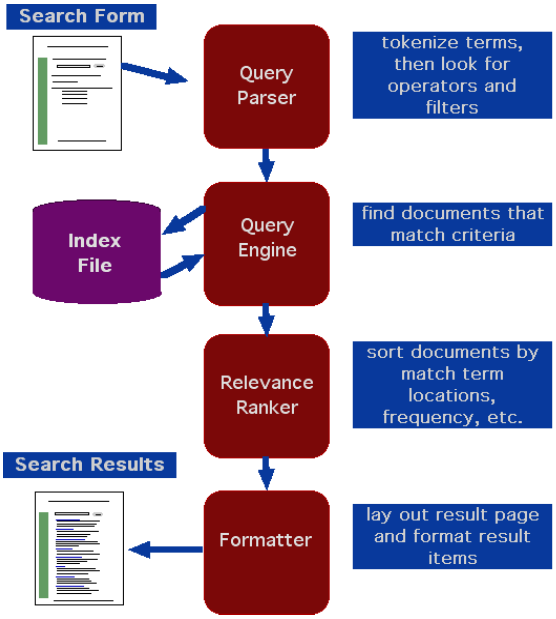
If a web page is never linked to in any other page, search engine spiders cannot find it. The only way a brand new page - one that no other page has ever linked to - can get into a search engine is for its URL to be sent by some human to the search engine companies as a request that the new page be included. All search engine companies offer ways to do this.

After spiders or crawlers find pages, they pass them on to another computer program for "indexing." This program identifies the text, links, and other content in the page and stores it in the search engine database's files so that the database can be searched by keyword and whatever more advanced approaches are offered, and the page will be found if your search matches its content. The above flow of working process can be shown in a flow chart:



The steps involved in working process of search engine are:

1.Document Gathering - done by Crawlers, spiders.
2.Document Indexing - done by Indexer
3.Searching
4.Visualisation of Results


Search Form
Query Parser
tokenize terms, then look for operators and filters
Index File
Query Engine
find documents that match criteria
Relevance Ranker
sort documents by match term locations, frequency, etc.
Search Results
Formatter
lay out result page and format result items

Challenges faced by Web search engines

- The web is growing much faster than any present-technology search engine can possibly index (see distributed web crawling).
- Many web pages are updated frequently, which forces the search engine to revisit them periodically.
- The queries one can make are currently limited to searching for key words, which may result in many false positives.
- Dynamically generated sites may be slow or difficult to index, or may result in excessive results from a single site.
- Many dynamically generated sites are not indexable by search engines; this phenomenon is known as the invisible web.
- Some search engines do not order the results by relevance, but rather according to how much money the sites have paid them.
- Some sites use tricks to manipulate the search engine to display them as the first result returned for some keywords. This can lead to some search results being polluted, with more relevant links being pushed down in the result list.

Web Search Strategies

I. Search steps:

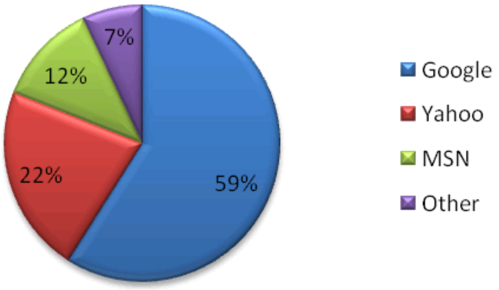
1. Analyze the search topic and identify the search terms (both inclusion and exclusion), their synonyms (if any), phrases and Boolean relations (if any)
2. Select the search tool(s) to be used (meta search engine, directory, general search engine, specialty search engine)
3. Translate the search terms into search statements of the selected search engine
4. Perform search
5. Refine the search based on results
6. Visit the actual site(s) and save the information (using File-Save option of the browser)
7.

II. Tips for effective web searching:

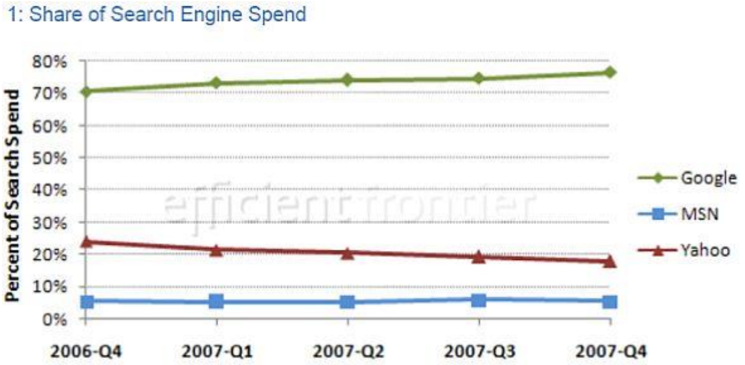
a. Broad or general concept searches: start with directory-based services (want a few highly relevant sites for a broad topic)
b. Highly specific or topics with unique terms/ many concepts: use the search tools
c. Go through the ‘help’ pages of search tools carefully
d. Gather sufficient information about the search topic before searching
    i. Spelling variations, synonyms, broader and narrower terms
e. Use specific keywords, rare/unusual words are better than common ones
f. Prefer phrase & adjacency searching to Boolean (‘stuffed animal’ than ‘stuffed’ and ‘animal’)
g. Use as many synonyms as possible - search engines use statistical retrieval methods and produce better results with more query words
h. Avoid use of very common words (e.g., ‘computer’)
i. Enter search terms in lower case. Use upper case to force exact match (e.g. ‘Light Combat Aircraft’, ‘LCA’)
j. Use ‘More like this’ option, if supported by the search engine (e.g. Excite, Google)
k. Repeat the search by varying search terms and their combinations; try this on different search tools
l. Enter most important terms first - some search tools are sensitive to word order
m. Use the NOT operator to exclude unwanted pages (e.g.: bio-data, resumes, courses)
n. Go through at least 5 pages of search results before giving up the scan
o. Select 2 or 3 search tools and master the search techniques

Emerging Technologies

A recent enhancement to search engine technology is the addition of geo coding to the processing of the ingested documents. Geo coding attempts to match any found references to locations and places to a geospatial frame of reference, such as a street address, gazetteer locations, or to an area (such as a polygonal boundary for a municipality).



Through this geo coding process, latitudes and longitudes are assigned to the found places, and these latitudes and longitudes are indexed for later spatial query and retrieval. This can enhance the search process tremendously by allowing a user to search for documents within a given map extent, or conversely, plot the location of documents matching a given keyword to analyze incidence and clustering, or any combination of the two.



For years, traditional search engines have presented results in a relevance-sorted list, while the recent trend in "vertical search" engines is to produce a limited slice of data for a single market. Meta-search engines simply combine results from other search engines, and clustering engines deliver an inconsistent experience for consumers.

In contrast, the Kosmix search engine—based on new algorithms and its own crawl and index—categorizes Web pages, producing a multi-dimensional view of the results. This approach reduces the number of clicks and queries it takes for users to find what really matters to them, and presents useful information they may not have considered otherwise.

CONCLUSION

The usefulness of a search engine depends on the relevance of the results it gives back. While there may be millions of Web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve.

REFERENCES

Best of the Web 1994. Navigators http://botw.org/1994/awards/navigators.html

Search Engine Watch http://www.searchenginewatch.com/

Web Growth Summary: http://www.mit.edu/people/mkgray/net/web-growth-summary.html

[Marchiori 97] Massimo Marchiori. *The Quest for Correct Information on the Web: Hyper Search Engines.* The Sixth International WWW Conference (WWW 97). Santa Clara, USA, April 7-11, 1997.

[McBryan 94] Oliver A. McBryan. GENVL and *WWWW: Tools for Taming the Web. First International Conference on the World Wide Web.* CERN, Geneva (Switzerland), May 25-26-27 1994.