

Principles and History of the Web

Search Engines

Martin Vlachynsky
The University of Aberdeen
(Part of Managing Innovations class assignment)

4/2010

Introduction

We can't imagine our life without Google, or at least Yahoo!. But Web search engines did go a long way, before they reached today's performance, and their development is far from being finished. We offer you a sight into 2 decades of Web search engines...

Web Search Engines: Principles and History

Create and sort a pool of data; find the most appropriate information; deliver it - this is the basic principle of Web search. Modern web search engine consist of four basic parts [Hock, 2004, pp.62].

- A spider/crawler/ robot – computer program, running continuously on search engine's servers and entering every page it finds. If it finds a new page or content changed since the previous visit, it informs the index. The freshness and extent of results is dependant mainly on the quality of the spider.
- Index – a database of results delivered and updated by the spider. This is where the engine searches for the results. The quality of index determines the speed of delivery, but also richness of results.

- Search engine itself – an algorithm 'that identifies (retrieves) those pages in the database that match the criteria indicated by a user's query' [Hock, 2004, pp.62]. It decides which result is the most appropriate for your query.
- Interface – a graphical and functional solution of the communication with user.

Search Before The Web

World Wide Web emerged during years 1989 - 1991 in CERN, the European Organization for Nuclear Research. The very first website¹ served also as the first web navigator, since it maintained list of newly emerging sites. During the first months, Web consisted of few hundred pages, which belonged mainly to scientific institutions and universities. Users could navigate from one to another using direct hyperlink, or with help of news sites manually collecting new additions to web by e-mails, word-of-mouth or just randomly found [Schwartz, 1998]. First tool created to enable centralized search in number of these hand-made catalogues was W3Catalog.

The Web soon outgrew the possibility of self-maintained news sites and the need for general search engine was intense. There were few hundreds of sites 'providing services such as USENET news feeds, "anonymous" FTP1 archives, WHOIS directory servers, and community specific information, such as bibliographic databases for biological scientists.' [Schwartz, Emtage, Kahle, Neuman, 1992] The most prominent were WHOIS and X.500, where individual administrators could submit specific details about their sites, like domain, name or e-mail. However, the spectrum of information available was very limited and required direct (and technically non-trivial) cooperation of administrators.

It is important to mention the difference between the Internet and the Web. Mentioned tools were not web search engines; they navigated users in FTP (File Transfer Protocol) space. FTP is a way, how computers, interconnected in the Internet network communicate and share files. World Wide Web is a solution, how to turn their content into one entity, which is accessible by a Web browser. While there were hundreds of thousands computers connected on the Internet, sharing millions of files, the Web consisted of few hundred pages.

The first tool, partially comparable to modern web search engine was Archie, created in 1990. It was a simple script, programmed to fetch site listings of anonymous FTP files from world servers, index them and allow users to search it. However, Archie was indexing only files, not their content, and users had to know exact name of the file [Schwartz, Emtage, Kahle, Neuman, 1992]. Archie pioneered also the commercial side of searching the web. While initially offered free, since 1992 sites that incorporated Archie had to pay for it [Savetz, 1993]. Rapidly growing number of files caused constant problems with Archie's architecture and respond to a query could take several hours on a busy day [Bowman, Danzig, Schwartz, 1993].

¹ Still maintained at <http://info.cern.ch>

Archie served as a search tool mainly for Internet, not Web. Gopher was created in 1991 for similar reason, but instead of collecting information about all files, it collected information about text files only on its servers. Veronica and Jughead were tools to search Gopher servers [Wall, 2001].

Web 1.0 Search

Terms Web 1.0, Web 2.0 and Web 3.0 do not refer to some official standards but to a paradigm of how Web is constructed and used. The differences can be characterized as 'technological (scripting and presentation technologies used to render the site and allow one user interaction); structural (purpose and layout of the site); and sociological (notions of friends and groups)' [Cormode, Krishnamurthy, 2008]. It is more paradigm evolution than revolution; the shift has been continuous and we can only identify the prevailing paradigm. Web 1.0 is represented by static pages, created and updated by administrators, with no or little visitor's interaction².

The first robot, used to automatically and repeatedly access sites, appeared in 1993 and was called World Wide Web Wanderer. Its function was to count the size of the Web and index all pages in the first web index Wandex. However, its purpose was not to serve as a search tool [Underwood, 1994].

Its counterpart ALIWEB, which emerged around the same time and intended to serve primarily as an index, with a simple search function, used different attitude. It asked sites' administrators to submit an index file with site description. With no robots there was no bandwidth overload and the problem that 'they retrieve all documents that can be reached, even those that are not interesting or suitable to index,' disappeared.' [Koster, 1994] More information included in the description enabled users to search specific information. However, this project failed, mainly because it did not manage to convince site administrators to submit index files and its database remained very small, despite efforts to incorporate other databases in it [Sonnenreich, 1997].

Soon several other spiders started to crawl Web, most notably JumpStation and the Repository-Based Software Engineering (RBSE) engine. The former one is considered the first real web search engine. It independently crawled the Web and gathered information about titles and headers. Index was searchable using a simple web form for the first time; users could use keywords instead of exactly matching words. However, JumpStation did not rank the results, just provided a list of results. The project faded out after finance sources dried. RBSE engine offered simple ranking system, however as a NASA project it was not primarily intended for broad public.

The ranking system was further improved by Excite (first versions appeared in 1993), which used statistical analysis to discover relationships between words and deliver more accurate results [Sonnenreich, 1997]. Another major move ahead came with WebCrawler in 1994, which was the first web search engine to index full text, instead of just titles and headings [Underwood, 1994]. Later in

² For more details see [O'Reilly, 2005]

1994, Lycos search engine was the first to offer number of advanced features we can see today – it had very large index, offered not only links as results but also brief snippets from sites, simple use of operators and match score for results.

Number of similar search engines followed in 1994 and 1995 – most notably Infoseek, OpenText, Magellan, Inktomi, Northern Light, AskJeeves and AltaVista. The last mentioned had the most advanced spider [Lewis, 1995] and backed by high-end hardware it was prepared to cope with rising demand.

Year 1994 is connected with one more major milestone in web search history – the foundation of Yahoo!. The biggest problem of existing web search engines was relevancy. They depended on proper naming of documents and robots experienced increasing difficulty to recognize relevant and dependable sources among the millions of newly emerging sites. Yahoo! offered different solution – human edited hierarchical and searchable web directory. Despite not being a real search engine (it was using Inktomi's search engine and developed own one only later in 2004) it quickly gained large popularity. Visitors could search information using keywords, but they also simply could browse the directory full of descriptive information until they reached the desired source.

Yahoo! is one of the symbols of the dot.com boom. In late 90ties, it was one of the main stars in business world with share value rising from less than \$1 in 1996 to over \$118 in 2000. Yahoo! soon turned into a web portal – complex website offering news, e-mail and other services beside the basic search option. Advertising was important source of income, but Yahoo! also charged commercial sites for incorporation into its directory. Most of existing search engines followed the strategy of transformation into web portal.

Web 2.0 Search

Creating a web page became increasingly easy in late nineties and social networks like MySpace shifted interpersonal communication from privacy of e-mails into public virtual space. Sites enabled users to create profiles, blogs, or to submit reviews. The ability of visitor to interact with the site is the main feature of Web 2.0.

Main search engines were losing ability to cope with this environment. The relevancy of results depended mainly on simple statistical analysis of keyword density. Not only it was far from providing ideal results, rising commercial importance of Web caused emergence of Search Engine Optimization (SEO) techniques, which further biased the reliability of results provided by search engines.

Sergey Brin and Lawrence Page proposed a new attitude [Brin, Page, 1998] in 1997. Their search engine was supposed to count backlinks³ (which have similar meaning as citations in academic

³ Hyperlinks pointing to the web document we inspect

writings) and use them to construct PageRank, a metric used to rate sites importance. In 1998, Google was launched.

Using backlinks as a voting system gave users the power to decide, what content should rank at the top. Google has been continuously improving its algorithm and today it uses dozens of indicators to count the value of a site – number and quality⁴ of backlinks, freshness, relevancy of content, domain and many others. Google's market share today accounts for about 2/3 of US searches [Nielsenwire. 2010]. Other engines follow Google's methods.

Google also chose different marketing strategy. It did not turn into web portal, but kept the extremely simple and clean design. The main income comes from contextual and search related adverts, which were pioneered by Goto.com (later bought by Yahoo!), but it was again Google to turn it into massive commercial success with its AdWords.

Modern search engines offer broad number search functions, like image searches or local searches⁵. Google itself runs over a hundred services beside search and offers complex solutions for websites, especially in terms of IT, marketing and analytics.

Web 3.0 search

While Web 2.0 was characterized by user's activity, Web 3.0 is characterized by the Web's response to this activity. The Web tries to understand user's action and react in a desired way – to deliver tailor-made content for each different visitor. This is possible thanks to large amount of information, available about the user, especially through the history of his actions and his personal information from virtual social spaces, most notably Facebook.

Search engines actual and expected developments reflect this situation. Personalized search based on user's search history appeared in 2005. However, Google missed the boom of social networks⁶ and thus the opportunity to harvest more information about users.

Related Web 3.0 challenge is the maximal freshness of results. New information spread extremely fast inside social networks, most notably in Twitter, which already has own search. Google offers live search box since December 2009, however, are disputable in terms of quality and relevancy [Sullivan, 2009].

Growing trend is also more focus on vertical search [Sullivan, 2007]. Search engines are trying to search a specific field to deliver most appropriate results. User can do it manually (for example by choosing to search in local search engine version), but the main challenge for engines is to discover

⁴ Links coming from reliable sites like universities' sites or major newspaper have bigger value than links coming from personal blog situated on a free hosting

⁵ There are over 30 search options in Google

⁶ Social network Orkut was rather a failure, Google launched brand new social service Buzz just in 2010

the proper field of user's interest. To achieve this, they need tools to not only discover and index the Web content, but also user's characteristics.

Bibliography

Brin, S., Page, L. 1998. *Anatomy of a large-scale hypertextual web search engine*. In Proceedings of the 7th International World Wide Web Conference (Brisbane, Australia, Apr. 14 –18). pp. 107–117.

Cormode, G., Krishnamurthy B., 2008. *Key differences between Web1.0 and Web2.0*. First Monday, 13(6), June 2008.

Hock, Randolph. 2004. *The Extreme Searcher's Guide to Web Search Engines. A Handbook for the Serious Searcher*. Medford, NJ:

Internet World Stats. 2001. *World Internet Usage and Population Statistics* [online] Available at: <http://www.internetworldstats.com/stats.htm> [Accessed 10 April 2010].

Lewis, P.H. 1995. *Digital Equipment Offers Web Browsers Its 'Super Spider'*, The New York Times, December 18, Available at: <http://www.nytimes.com/1995/12/18/business/digital-equipment-offers-web-browsers-its-super-spider.html> [Accessed 19 April 2010].

Martijn Koster. 1994. *ALIWEB – Archie-like indexing in the web*. In Proceedings of the First International Conference on the World Wide Web, Geneva, Switzerland,.

Nielsenwire. 2010. *Nielsen Reports February 2010 U.S. Search Rankings*. The Nielsen Company [online] , Available at: http://blog.nielsen.com/nielsenwire/online_mobile/u-s-web-searches-top-10-2-billion-in-january [Accessed 22 April 2010].

O'Reilly, T., 2005, *Design Patterns and Business Models for the Next Generation of Software*, oreilly.com, [online] (updated Oct. 2009) Available at <http://oreilly.com/web2/archive/what-is-web-20.html> [Accessed 8 April 2010]

Savetz Kevin. 1993. *Life Before (And After) Archie*. Internet Business Journal [online] Available at: http://www.savetz.com/articles/ibj_bunyip.php?sort=date [Accessed 10 April 2010].

Schwartz, Candy. 1998. *Web search engines*. Journal of the American Society for Information Science, 49(11), 973–982.

Schwartz, M. F., Emtage, A., Kahle, B., & Neuman, B. C. (1992). *A comparison of Internet resource discovery approaches*. Computer Systems, 5, 461–493.

Slawski. Bill. 2006. *Just what was the first search engine? SEO by the Sea* [online] Available at: <http://www.seobythesea.com/?p=106> [Accessed 12 April 2010].

Sonnenreich, Wes. 1997. *A History of Search Engines*. Wiley [online] Available at: <http://www.wiley.com/legacy/compbooks/sonnenreich/history.html/> [Accessed 19April 2010].

Sullivan, Danny. 2007. *Search 3.0: The Blended & Vertical Search Revolution*. Search Engine Land [online] , (Updated on November 27, 2007) Available at: <http://searchengineland.com/search-30-the-blended-vertical-search-revolution-12775> [Accessed 22April 2010].

Sullivan, Danny. 2009. *Search & Real Time Madness*. Search Engine Land [online] , (Updated on December 10, 2009) Available at: <http://searchengineland.com/search-real-time-madness-31668> [Accessed 22April 2010].

Underwood, Lee. 1994. *A Brief History of Search Engines*. Webreference [online] (Updated: August 18, 2004) Available at: http://www.webreference.com/authoring/search_history/ [Accessed 09April 2010].

Wall, Aaron. 2001. *History of Search Engines: From 1945 to Google 2007*. Search Engine History [online] Available at: <http://www.searchenginehistory.com/> [Accessed 06April 2010].

Web Search Engine History Timeline

By author

