

This is an author produced version of a paper published in Tidskrift för Dokumentation (Nordic Journal of Documentation).

This paper does not include the final publisher journal pagination.

Citation for the published paper:

Våge, L., "Search engines of today and beyond",

Tidskrift för Dokumentation, 2002, volume 57, issue 3, pp. 79-85.

Published with permission from: Svensk förening för Informationspecialister

Search engines of today and beyond

L. VÅGE
Mid Sweden University

Abstract

The major Internet search engines nowadays makes it possible to search for keywords in vast quantities of web pages. In this article the most important crawler- based search engines are reviewed along with newer ones which are starting to make an impact. Web directories, meta-search engines and the use of sponsored links are considered as well as the future of web searching. The need for faster reindexing of the databases of search engines for the purpose of giving the users a more up-to-date view of the web is emphasized.

Keywords

Search engines, information retrieval.

Searching for information using search engines is the second most popular activity among the users of the Internet. For some it might seem that there is an abundance of search engines to choose from but that picture has become increasingly untrue as we approach the end of the second year of the new millenium. In this article I will attempt to review the most important of the few that currently dominates Internet searching, with emphasis on crawler-based search engines which probably can be said to be the most comprehensive tools at the disposal of the web searcher. Anyone reading this should of course take into consideration the volatile character of any information concerning search engines. In the past rapid and unexpected changes have been the norm rather than the exception and there are no indications that this state of affairs will not continue.

To Google or not to Google?

By far the most influential search engine of today is without a doubt Google. In fact for many people "to Google" has practically become synonymous with searching the net. This phenomenal success story started as a research project at Stanford University and Google was first launched publicly in 1998. It took a while before it's simple but effective interface caught on, but by the year 2000 everyone realized that not only did Google supply it's users with a huge index, it also had the ability to present extremely relevant results. By avoiding the pitfalls of portalisation, into which most of the older search engines had fallen during the final years of the last millennium, Google again made Internet searching popular. It was the first search engine to announce a database consisting of more than 1 billion web pages on June 26 2000 (1) and also broke the 2 billion document limit just one and a half year later on December 11 2001 (2).

More important than the sheer size of Google was the relevance ranking system called PageRank (3) which was based on link analysis. While not all might agree upon the virtues of this particular algorithm it has been generally perceived to work far better

than methods primarily built on the analysis of the frequencies and placements of keywords in the actual documents. An important break-through and a first excursion into the so called invisible web made by a major spidering search engine was the inclusion of indexed PDF-documents in the main Google database in the early spring last year. In the autumn this was followed by the addition of several other document formats like PostScript, Word, Excel and PowerPoint.

Another splendid achievement by Google in 2001 was the Google Groups search interface to the Usenet Newsgroup discussions. After acquiring the DejaNews archive in February (4) they succeeded in tracing and merging older archives and in December launched a new version of Google Groups covering nearly the entire history of Usenet all the way back to 1981. Somewhat resting on it's laurels during this year Google has not evolved as dramatically lately but have recently upgraded the Google News Search which now extracts news stories from over 4000 sources every few minutes.

In early October it became known that Google had again won a contract with the most popular web directory Yahoo (5). Many was deeply surprised however by the fact that not only did Google's results complement the links in the Yahoo directory but the two result categories were integrated and presented as one list of hits. The effect of this change has practically led to the eradication of the Yahoo concept of being a directory of selected and reviewed links which has since 1994 been the winning formula of this the first major web directory. In May this year Google also took the place of rival company Inktomi as chief supplier of web search to the hugely popular America Online Search portal (6). Likewise Google licenses web searching to New York Times, BBC, Netscape to name but a few.

Life in the FAST lane

If Google can be considered to be the leading search engine it is also clear that FAST is the only real competition at the moment. The Norwegian company FAST Search & Transfer intended Alltheweb.com to be a technology show-case and the biggest search engine in the world when it was announced on August 2 1999 (7). Indeed on a couple of occasions it has occupied the throne. Initially having a size of 200 million documents it grew to outnumber Google in June this summer with 2.1 billion searchable documents (8). While Google quickly managed to expand it's size to approximately 2.5 billion, FAST might be just as big if you take into account the fact that a certain percentage of the URLs in the Google database has been only partially indexed. This problem has been described in detail and repeatedly stressed by search engine expert Greg Notess (9).

Leaving the size wars behind it must be pointed out that the search technology and functions made available to the searcher by FAST in AlltheWeb are more than impressive. For advanced searchers there is a lot of possibilities that Google cannot provide at the moment. One thing that FAST has pioneered is the extensive customisation possibilities of both the search results and the graphical presentation. In September they became the first to index the text residing in files of the popular Flash format (10) while adding indexing of PDF documents earlier this year. FAST claims to crawl and to reindex it's database every 9-12 days (11) whereas the others normally have a spidering cycle of a month at best.

FAST has succeeded in licencing its database to a growing number of search destinations and portals in Scandinavia and through-out Europe. In 2000 it had replaced the database produced by the original Lycos T-Rex spider at all the international Lycos portals, including the American lycos.com. A successful collaboration with the leading scientific publisher Elsevier resulted in the introduction in April 2001 of Scirus - the scientific search engine (12). In Scirus you can search for documents in the visible web and to some extent the invisible web, as well as for articles in journals published by Elsevier. It should also be noted that while the AlltheWeb News Search covers only 3000 news sources the scope is far more international than that of the Google News search.

The fate of the old dinosaurs

One of the most important search engine companies of the past, Inktomi (13), has been forced to make successive reductions in their work force (14) because of their dwindling luck in the face of the competition during the last two years from Google and FAST. Inktomi has from the start adopted the business model of not maintaining their own search engine but rather providing their technology and database to portals wanting to give their users access to web searching. The first display of the Inktomi search technology was the HotBot search engine which made its debut in 1996. While still delivering to the American HotBot version (15) now owned by Terra Lycos, it has lost some of its most lucrative contracts like the ones with Yahoo, America Online and Iwon. Refocusing on search again they made a remarkable improvement this summer by quadrupling the size of its index to reach the 2 billion mark in the process also starting to index PDF documents and intensifying its spidering frequency. It remains to be seen if Inktomi will manage to recover from the setbacks it has suffered. For the time being its most prestigious customer is of course Microsoft which uses Inktomi web results on all their international search portals although the primary results at these locations come from the LookSmart web directory.

If Inktomi struggles for survival so does the oldest spidering search engine still in the business - Altavista. Starting in late 1995 Altavista seemed to be the ultimate search tool, providing a staggering 30 million document index and power searching facilities no one could match. But as time went by Altavista became a portal offering lots of other services apart from their web searching. This was not a wise move and people started to abandon Altavista's cluttered interface for cleaner and more efficient ones after 1998-99. The company has changed owners a few times and is now the property of not so prosperous Internet company CMGI. Like Inktomi, Altavista has announced that they have again turned to web searching as their first concern and has made a few improvements during 2002, one example being the Altavista Prisma function (16). Even if database size is only around 1 billion web pages and the relevance ranking algorithm still seem to be the same that has been so successfully spammed in the past, Altavista is not out yet. No other major search engine today can offer anything like Altavista's array of advanced search tools like searching for truncated words, Boolean searching with nested parentheses, case sensitivity and proximity searching.

As a result of the so called Dot-Com Crash quite a number of search engines have disappeared from the scene or have survived in name only. Among the most familiar names apart from the original Lycos, InfoSeek went extinct in early 2001 and Excite followed suit in the fall and went bankrupt like the largest European search engine,

Euroseek. Today Lycos uses FAST as mentioned earlier, Excite and WebCrawler has become meta search engines owned by InfoSpace, Euroseek is powered by Google while Infoseek is gone altogether as a search destination. In January this year the renowned search engine Northern Light was bought by Divine Inc. and it was announced that they would no longer provide free web searching (17). In the case of web directories we should remember that both InfoSeek and Excite had their own comprehensive directories of reviewed links. One of the biggest of them all, the directory originally known as Snap and later as a part of the NBCi portal was lost when parts of that portal was discarded in the first half of 2001. Although NBCi still is in operation search results are now provided by another meta-search engine product from InfoSpace, namely Dogpile.

The new kids on the block

In the midst of this dismal period for search engines two newcomers appeared in 2001. The first of these, Teoma, was the result of a research project called DiscoWeb at Rutgers University. In April 2001 the beta version of the Teoma search engine was launched (18) and quickly got a lot of attention because of the advanced link analysis techniques which were intended to produce more relevant results. Some of the ideas had already been suggested by the IBM Clever project (19) but Teoma was the first to put them to the test. Later last year Teoma was acquired by natural-language search engine AskJeeves (20) and has helped them to regain some of their former popularity. Ultimately it is up to us to decide if Teoma produces more relevant results than Google by using link analysis which identifies authoritative sites and link hubs. The down side for now is that the size of the Teoma database is just a few hundred million web pages and that the index is not updated as often as Google, FAST or Inktomi.

The second major search engine to surface in 2001 was the Korean WiseNut (21) which was launched in September. It was quite a sensation when it appeared out of nowhere claiming a database size of 1.5 billion web pages, equalling that of Google at the time. Unfortunately, WiseNut turned out to be a big disappointment as people realized that the number of URLs was obviously overstated and that the index was never updated but rather left as it was. Still there was some impressive technology at work like automatic document categorisation and inline document previews and WiseNut was soon bought by one of the leading web directories, LookSmart (22). It is only after this summer that WiseNut has again started to crawl the web and reindex it's database and we have yet to find out what the future intentions of the present owners are.

Arguably the most surprising occurrence in the search engine world during 2001 was the emergence of the Internet Archive (23) and it's search interface called The Wayback Machine. Suddenly it was possible for the casual web searcher to study copies of web pages dating as far back as 1996. In the Wayback Machine you cannot search for keywords, you must know the URL of a page you want to look at older versions of. For some web documents you might be able to find more than a hundred incarnations, for others only a few might be available. Recently a new tool was tested which made comparisons between different document versions easy to perform in the Wayback Machine. Needless to say it is an incredible resource for anyone interested in the history of the web or just wanting to access a deleted document. The Internet Archive was presented to the world on October 23 last year at the University of

California (24). This truly vast project covering well over 10 billion documents is the brainchild of Brewster Kahle of Alexa Internet, working together with among others the National Science Foundation and the Library of Congress. The Internet Archive in April this year donated a copy of its entire repository of web pages from 1996 and through 2001 to the New Library of Alexandria (25).

This year has produced another two interesting search engines which are presently undergoing public beta-testing. One of them is a project called Gigablast (26) by former InfoSeek employee Matt Wells. While the database size is just around 150 million web pages there are promising features like cached documents and precise reporting of the date any given page was last spidered. The creators of the Taiwanese OpenFind Beta (27) on the other hand has the audacity to proclaim that it has a database consisting of no less than 3.5 billion web pages, which would make it the largest search engine ever. Search result numbers in OpenFind are indeed huge but a lot of the links seem not to have been crawled for a very long time according to the dates in the result list. The interface is also a bit primitive and unstable and some of the results are duplicates or obviously erroneous, making OpenFind very difficult to form an opinion of yet.

Web directories, meta-search engines and sponsored links

The web directories which are created by human editors have clearly become more and more secondary to the spidering search engines as the size of the web continues to grow. The largest directories like Yahoo, LookSmart and the OpenDirectory (28) all have 3-4 million links making them outsized by Google and the others by a factor of at the very least 50-1. Maybe the new integration of the Yahoo reviewed links and the results from Google forming the new default Yahoo result presentation is a sign of the times. The lesser known LookSmart still delivers directory results to the Microsoft Search Portals and secondary results to Altavista. LookSmart has also acquired and incorporated the Zeal directory (29) which is produced by voluntary editors. The web directory with the largest distribution across search portals around the net, the Open Directory Project, is a co-operative effort by more than 50,000 editors and the entire database can be freely downloaded if you want to create your own search engine. Regrettably, the problems, experienced by many web site owners wanting to submit URLs and people wanting to contribute, in getting the attention of the editors of the ODP are by now well documented and diminishes the value of this directory to some extent. It is nonetheless used by AOL, Google, Lycos, Netscape, Excite, HotBot and many more.

Any article on web searching would be incomplete without mentioning the meta-search engines. Rather than adding anything new, as they don't maintain their own databases, these tools aggregate results from "real" search engines and present them after their own fashion. The best of the meta-search engines has the ability to remove duplicates, take advantage of the differing relevance ranking systems being used and even put the results in convenient folders that can be browsed. The latter of these practices is commonly referred to as automatic document categorisation and has been put to brilliant use by above all Vivisimo (30). The categories created in realtime by the Vivisimo clustering engine produces not only one layer of conceptual folders, but a hierarchical tree. This is certainly impressive and at times the effectiveness of their algorithms can be simply stunning. The American Institute of Physics (31) apparently thought so and Stanford University's HighWire Press (32) as well. Both

has signed deals to use Vivisimos technology to enhance online access to their vast repositories of research articles. Apart from the success of specialized tools like Vivisimo the general feeling about meta-search engines seem to be that we use used to like them but since Google is so good, we don't need them anymore. Maybe this is true in the respect that the relevance ranking principles adopted by the major search engines of today are much better than they were a few years back.

One of the principal problems involved in web searching so far in the 21st century is the abundance of sponsored links present in the result pages of most search engines. Furthermore these links tend to displayed very prominently, sometimes making it necessary to scroll down to see the actual search engine hits. The success of above else Overture (33) but also Google and Espotting selling their lists of paid links accounts for a lot of this problem. The Federal Trade Commission of the United States this summer demanded of the search engines that they make it more clearly visible to the users which links in fact are nothing less than plain ads from paying companies (34). That these sponsored links, as the preferred term suggested by the FTC is, are here to stay is without question. Many of the owners of the search engines badly need the extra incomes generated by these, therefore, the serious searcher will just have to learn to identify and ignore what is really pseudo- results.

Is the future here yet?

So what will the future of web searching have in store for us? Unfortunately it seems conceivable that the emerging trend of prominently featuring paid links from services like Overture, will be a crucial part of the economy of search engines even of Yahoo's magnitude. Obviously just running ads didn't work and neither did the portalisation craze, so naturally the search engines must go where the money is. Another dim perspective of how the future for some web searches might be, has been eloquently demonstrated this autumn by the Chinese government in their blocking of search engines like Google and Altavista. After at first simply refusing the Chinese Internet users access to Google at all, the firewalling procedures involved turned more subtle and starting punishing people searching for forbidden words by denying Internet access for a certain time period (35). The Human Rights Watch organization has written to the CEOs of the two search engines urging them to resist the censorship any way they can (36).

To strike a more positive note, advanced technology like lightning-fast conceptual clustering of results and the use of visualisation in result presentations have started to bear interesting fruit. It is too early to say if the novel graphical "hit lists" of experimenting search engines like the French Kartoo (37) and Mapstan (38) or the American VisIT (39) will be a significant contribution to the way we search the web in the future. However, these pictorial representations that present web pages like cities and villages on a map with roads between them, symbolising kinship through inter-linking or common keywords, can in fact convey more information than the standard linear lists of results that are used by all major search engines today.

A development of truly seminal importance would be if the biggest search engines became able to refresh their databases on-the-fly while spidering the web. Generally it takes a number of days or even weeks before the harvesting of new web pages and new versions of already existing web pages will be visible in the search engine indexes. What we really need is a scalable technology capable of instantaneously

updating databases the size of several hundreds of millions of web documents. Realtime indexing has already been implemented on the smaller scale of a few thousand news web sources by the AlltheWeb News Search and Google News among others. In the beginning of August the Nippon Telegraph & Telephone company in Japan publicly stated that their researchers had recently solved the problem of indexing without delay in databases of a massive size while crawling the web (40). So maybe some day soon search engines available to us all will harness technology that will do this. Then perhaps we will be able to get rid of the feeling that when we search the web we just find out about how it was a few weeks ago. Therefore I will conclude by saying that only when search engines can tell us what the web is like today can the future be said to be here.

References and links:

1. Google Launches World's Largest Search Engine, <http://www.google.com/press/pressrel/pressrelease26.html>, Accessed: 2002-10-20
2. Google Offers Immediate Access to 3 Billion Web Documents, <http://www.google.com/press/pressrel/3billion.html>, 2002-12-20
3. Google Technology, <http://www.google.com/technology/index.html>, Accessed: 2002-10-20
4. Google Acquires Usenet Service and Assets from Deja.com, <http://www.google.com/press/pressrel/pressrelease48.html>, Accessed: 2002-10-20
5. Yahoo Renews With Google, Changes Results, <http://searchenginewatch.com/sereport/02/10-yahoo.html>, Accessed: 2002-10-20
6. Google to Power Search Functions Across AOL Brand, <http://www.google.com/press/pressrel/aol.html>, Accessed: 2002-10-20
7. World's Biggest Internet Search Engine Goes Online, http://www.fastsearch.com/press/press_display.asp?pr_rel=77, Accessed: 2002-10-20
8. FAST's Alltheweb.com Dethrones Google As The World's Largest Search Engine, http://www.fastsearch.com/press/press_display.asp?pr_rel=137, Accessed: 2002-10-20
9. Google's Unindexed URLs, <http://www.searchengineshowdown.com/features/google/unindexed.shtml>, Accessed: 2002-10-20
10. FAST unveils Macromedia Flash searching capability on Alltheweb and to its portal customers, http://www.fastsearch.com/press/press_display.asp?pr_rel=152, Accessed: 2002-10-20
11. FAST launches world's freshest Internet search engine, http://www.fastsearch.com/press/press_display.asp?pr_rel=93, Accessed: 2002-10-20
12. Elsevier Science Launches Scirus.com, http://www.scirus.com/press/html/Scirus_Press_Release_Launch.htm, Accessed: 2002-10-20
13. Inktomi, <http://www.inktomi.com/>, Accessed: 2002-12-20
14. Inktomi Announces Staff Reductions, <http://www.inktomi.com/company/news/press/2002/staff.html>, Accessed: 2002-10-20
15. HotBot, <http://www.hotbot.lycos.com/>, Accessed: 2002-10-20

16. Altavista Launches Altavista Prisma, <http://about.altavista.com/prelease?yr=2002&dt=070202>, Accessed: 2002-10-20
17. Northern Light Technology Refines Its Business Focus, <http://library.northernlight.com/FB20020108420000192.html>, Accessed: 2002-10-20
18. Advanced Search Engine Teoma.com Launches, Bringing Authoritative Results to the Web, http://www.irconnect.com/askj/pages/news_releases.mhtml?d=25648, Accessed: 2002-10-20
19. The Clever Project, <http://www.almaden.ibm.com/cs/k53/clever.html>, Accessed: 2002-10-20
20. Ask Jeeves Acquires Teoma Technologies, Inc., http://www.irconnect.com/askj/pages/news_releases.mhtml?d=25648, Accessed: 2002-10-20
21. WiseNut, <http://www.wisenut.com/>, Accessed: 2002-10-20
22. LookSmart Strengthens Leadership Position in Search Targeted Marketing With Acquisition of WiseNut, Inc., <http://www.shareholder.com/looksmart/releaseDetail.cfm?ReleaseID=74579>, Accessed: 2002-10-20
23. The Internet Archive, <http://www.archive.org/>, Accessed: 2002-10-20
24. New Internet library holds 10 billion Web pages, <http://news.cnet.com/investor/news/newsitem/0-9900-1028-7661681-0.html>, Accessed: 2002-10-20
25. Donation to the new Library of Alexandria in Egypt, <http://www.archive.org/about/bibalex.php>, Accessed: 2002-10-20
26. Gigablast, <http://www.gigablast.com/>, Accessed: 2002-10-20
27. Openfind - GAIS30 Project, Test Site, <http://www.openfind.com/en.web.php>, Accessed: 2002-10-20
28. ODP - Open Directory Project, <http://dmoz.org/>, Accessed: 2002-10-20
29. Zeal.com, <http://www.zeal.com/>, Accessed: 2002-10-20
30. Vivísimo Document Clustering - automatic categorization and content integration software, <http://vivisimo.com/>, Accessed: 2002-10-20
31. Institute of Physics Publishing Selects Vivísimo to Enhance Online Journal Access, http://vivisimo.com/press/Press_Releases/iop.html, Accessed: 2002-10-20
32. Stanford University's HighWire Press Licenses Vivísimo Clustering Products, http://vivisimo.com/press/Press_Releases/highwire.html, Accessed: 2002-10-20
33. Overture - Search Performance, <http://www.overture.com/d/home/>, Accessed: 2002-10-20
34. Commercial Alert Complaint Letter Attachment, <http://www.ftc.gov/os/closings/staff/commercialalertattatch.htm>, Accessed: 2002-10-20
35. Google keywords knock Chinese surfers offline, <http://www.newscientist.com/news/news.jsp?id=ns99992797>, Accessed: 2002-10-20
36. Google, Alta Vista: Resist Chinese Censorship, <http://hrw.org/press/2002/09/china0907.htm>, Accessed: 2002-10-20
37. Kartoo, <http://www.kartoo.com>, Accessed: 2002-10-20
38. MapStan Search: the metaengine with knowledge capitalization, <http://search.mapstan.net/>, Accessed: 2002-10-20

39. VisIT, <http://www.visit.uiuc.edu/>, Accessed: 2002-10-20

40. Building a Better Search Engine, <http://www.pcworld.com/news/article/0,aid,103676,00.asp>, Accessed: 2002-10-20