Expectations Versus Reality – Web Search Engines at the Beginning of 2005

Judit Bar-Ilan

barilaj@mail.biu.ac.il
Department of Information Science, Bar-Ilan University, Ramat Gan 52900 (Israel) and School of Library, Archive and Information Studies, The Hebrew University of Jerusalem

Abstract

Web research is based on data from or about the Web. Often data is collected using search engines. Here we describe our "wish list" for the ideal search engine, explain the need for the specific features and examine whether the currently existing major search engines can at least partially fulfil the requirements from the ultimate search tool. The major search tools are commercial and are oriented towards the "average" user and not towards the Web researcher, and therefore are unable to meet all the requests. One possible solution is for the research community to recruit the necessary funding, resources and know-how in order to build a research-oriented search tool.

Introduction

Even though the Web exists for only fifteen years, it has become a major information source and communication medium and has an influence on our information and communication behavior both in our everyday and in our scientific lives. Web research is multidisciplinary; it is extensively researched by computer and information scientists, sociologists, economists, linguists, psychologists, communication scientists and others. In order to carry out studies on the Web, most researchers need data. Some researchers (e.g. Thelwall, 2001 or Albert, Jeong & Barabasi, 1999) are capable (both technically and financially) to build there tools to collect the necessary data from the Web. The SocSciBot developed by Thelwall's team is freely available for anyone who has the resources to run the crawler and to store its outputs (The Statistical Cybermetrics Research Group, 2004a) and results of specific crawls are also available for download (The Statistical Cybermetrics Research Group, 2004b). Others (e.g. Jansen, Spink & Saracevic, 1999 or Huberman & Adamic, 1999) are lucky enough to have access to data collected by commercial search tools. Some Web studies are conducted by researchers working at search engines (e.g. Broder et al., 2000 or Fetterly et al., 2004), thus they can relatively easily crawls the Web or access data from the search engine's database.

Other researchers can either envy the lucky and resourceful ones or can turn to freely available information retrieval tools to search engines, or to databases of Web preservation projects (e.g. the Internet Archive, www.archive.org). Currently Web preservation projects have either limited access (copyright problems) or provide minimal or no text retrieval tools, thus they are of very limited use for Web research. Thus the best free tools are the search engines. The search engine scenery has undergone major changes in the last few years, and currently there are only a few major players (Sullivan, 2004a): Google (www.google.com), Yahoo (www.yahoo.com), AskJeeves/Teoma (www.teoma.com) and the newcomer MSN beta (http://beta.search.msn.com/). In this paper we will consider Google, Yahoo and MSN beta (Askjeeves/Teoma lacks a number of major features for Web research, like links to a page). There could be additional search engines having special capabilities relevant for Web research, but in our opinion the coverage of the search tool is a central feature (i.e., in this case, size matters) and thus the examination of the search tools is limited to these search engines only. According to Sullivan (2004b), Google, Yahoo and AskJeeves together provide 97% of the search results on the Web as of May 2004 (Google 54%, Yahoo 41% and AskJeeves 2%).

The list of features for an ideal search engine

The list of features is inevitably subjective**.**

1) coverage

If we are interested in producing any quantitative measures, then the coverage of the search tool should be uniform (over all the existing domains) and exhaustive. This is essential for measures based on the size of the site or domain, or on its visibility (e.g. number of inlinks or outlinks), but is also a basic requirement for sampling Web sites or pages. This request is not 100% achievable: new Web

Proceedings of ISSI 2005, pages 87–96.
Edited by P. Ingwersen and B. Larsen.

pages are being created continuously and it is impossible to capture them instantly. Still we want the search engine to cover as many pages as possible.

2) reliability
According to the Oxford dictionary (1989), reliability is "the extent to which a measurement made repeatedly in identical circumstances will yield concordant results". In order to be able to work with search engines, the results sets have to be stable (or almost stable) for some short period of time (e.g., an hour or a day). It is acceptable that search results change over time, because of the dynamic changes occurring on the Web and because the databases of the search engines get updated. However it is not acceptable to have major fluctuations in retrieval results, when it is clear that the reasons for the major changes are not caused by the above mentioned reasons.

3) transparency, disclosure, clear documentation
Search engine companies should clearly inform the users about the search features and capabilities, procedures and operational methods and promptly announce any problems with the search tool. In order to be able to use the results of the search tools for Internet research, it is imperative that the search engines' features work according to the specifications, and if not users should be notified as soon as possible.

4) timeliness
The search engine's database should be frequently updated, so that the snapshot the search engine has of the Web is as similar to the real Web as possible. The database freshness can be measured by Brewington and Cybenko's (2000) (α,β)-currency measure.

5) indexing the whole document
When our intention is to create and exhaustive list of documents, having a certain text-based feature, then even if search phrase appears at the bottom of a document (e.g. a bibliographic reference), it should be part of the list.

6) response time, accessibility
Internet researchers, like all other users want to receive immediate results. Timeouts can cause reliability problems, because these may change the number of results retrieved (often without the search engine providing information about the sudden change in the reported results). A search tool becoming inaccessible or having frequent timeouts, interferes with the search process. Since the Web is very dynamic by itself it is imperative that the searches be carried out in the shortest possible time span.

7) objectivity – no commercial influences and no influence on the environment
We are looking for the prefect tool that has no biases at all and enables us to study the Web while utilizing the tool, and not to study the picture we receive through the "eyes of the search engine". This is an ideal request; our actual objective is to approximate this situation.
The search engine should be a tool that allows us to access information through it; it should not influence the Web itself. However the whole search engine optimization industry was founded in order to design and redesign Web pages so that they come up high on specific search terms at specific search engines. By trying to "please" the search engines, instead of being an unobtrusive tool, they become influential players on the Web.

8) all reported results are retrievable
Document and word counts are often insufficient for Web research (especially when these numbers are unreliable). In order to study the documents themselves, we have to access them. Thus knowing that there are 11,203,349 pages that the search engine marked as relevant to our search, but being able to access only 1000 is not satisfactory.

9) ranking, different sorting options

Depending on the research topic, we do not always want to look at all the search results. In these cases, ranking becomes very important. Ranking algorithms are well-kept secrets, both from the competition and from potential spammers. In an ideal search tool Web researchers should have the ability to pick the parameters that influence the ranking (e.g. dates, weighted terms, placement, inlinks, anchors).

10) flexible output display

By this we mean the ability to choose the number of results per page, what information to show (e.g. URL only, snippet, size, title, URL, language) whether the results should be clustered or not, and whether to show only a sample of the pages from each site (this option is called site collapse). An additional requirement is to be able to set these preferences. Easy browsing of the result set is also needed, i.e. the ability to quickly jump forward to see, say result number 845.

11) cached results

This feaure helps the researcher understand why the page was retrieved (often pages change between the time they were visited by the search engine and the time they are visited by the user). In addition, if and when the host of the page is down or unreachable, one can still have a look at the cached version.

12) high quality retrieval in languages other than English as well

This issue is very problematic: even though about 70% of all Web pages were estimated to be in English as of 2000 (Cyber Atlas, 2000), in 2004, two thirds of the Web users were not native speakers of English (Global Stats, 2004). Information retrieval research is very heavily geared towards English. The major search engines enable users to search in languages other than English as well, and usually because of the lack of satisfactory tools in the local language, these tools are used for searching in other languages as well (there are of course some exceptions, Russian for example). For languages where there is heavy use of compounding, inflection, prefixes the basic machinery for retrieval in English is very far from sufficient.

13) accessible API

Accessible API (application programming interface) allows customization and development of useful tools and interfaces based on the publicly available features of the search engine.

14) full Boolean searches, diversity of operators

This feature allows the researcher to tailor the searches to his/her needs. In terms of logic, AND, OR, and NOT are a complete set of operators, but only when the operations can be combined. Thus obviously we not only need AND, OR and NOT but some way (parentheses or reuse of partial results) in order to have full logical expressibility. This however is not enough, for text retrieval we need additional operators, like phrase searches, NEAR or ADJACENT (with flexible definitions of what is meant by these operators).

15) advanced techniques for retrieving data for link analysis

The Web is made up of links and nodes, and links are studied actively in a number of areas: investigations of the Web structure, its evolution, the creation of communities and social networks on the Web, ways to improve information retrieval, the use of indicators based based on linkage and characterization of the link structure, (e.g., Broder et al., 2000; Kumar et al., 2003; Kumar et al., 1999; Faba-Perez, Guerrero-Bote, & De Moya-Anegon, 2003; Kleinberg, 1999; Ingwersen, 1998; Thelwall, 2003; Bar-Ilan, 2005). Such studies rely on data about links. The most basic feature is links to a specific page. We also need links to a site or a subsite and more generally to be able to define both the set of anchor pages and the set target pages. Currently we can only count link pages; we would also like to have information on the actual number of links. We should be able to define how to handle relative links. An additional step forward would be to have at least a basic characterization of the links: navigational links, embedded content links, links in lists and placement of link.

16) wide variety of search modifiers

Our basic assumption is that the researcher knows what she/he wants, is able to understand the different features and is able to choose the correct options to solve the problem at hand. This is in contrast with what the search engine developers assume about the general public (WWW10 Panel, 2001). In order to be able to phrase the queries more accurately we need a number of ways to limit our searches, a partial list contains: dates, domains, languages, geographic area, file formats, placement in file (e.g., title, url, anchor), by metadata fields if and when they exist in the documents.

17) additional features: stemming on/off, truncation left/right, wildcards, case sensitivity on/off, spell check, site collapse on/off

These features listed help the Web researcher in phrasing the queries even more accurately.

18) search assistance: relevance feedback, similar/related pages and searches, personalization

19) ability to combine all the features in a single query (including unlimited number of search terms)

20) non-textual retrieval capabilities

Most of the paper discusses text-based retrieval, but additional media has to be taken into account as well. We are not experts on multimedia searches and our space is limited anyway, but this point will definitely have to be discussed in the future.

Search engines in the past and in the present

Search engines are changing constantly, thus we want to emphasize that every point made about these tools is based on our findings as of mid-January 2005. To support the findings, we saved and documented every example presented in the paper and the interested reader will be provided with the saved copies of the search examples and other documentation on which the paper is based.

1) coverage

In 1995 the Web world was naïve enough to accept that Lycos "claims to have indexed 91% of the Web" (Ambrogi, 1995). After the studies published by Bharat and Broder (1998) and Lawrence and Giles (1998, 1999), such claims were not made any more, we simply cannot expect the search tools to be exhaustive. Another problem is the non-uniformity of coverage. Snyder and Rosenbaum (1999) demonstrated that the even the relative coverage of major domains of different search engines was not the same. Thelwall (2000) examined the coverage of large national domains, and found that the coverage was so uneven that reasonable calculation of the Web Impact Factor (WIF, Ingwersen, 1998) was not possible based on data provided by the search engines. Current results show that coverage of search engines is still uneven. We searched for "–kxht site:.xy" (without the word kxht and in the domain "xy") in Yahoo and Google, and "site:xy -(kxht)" in MSN on January 13, 2005. We had to exclude a very rare word, because Google does not support standalone searches for sites. The results for a number of domains appear in Table 1. It is easy to spot the relative differences between the domains by the search engines and also between the rankings of the search engines for coverage, when considering the domains one-by-one.

Table 1: Domain coverage

	Hungary (hu)	Canada (ca)	Djibouti (dj)	Suriname (sr)
Google	13,300,000	32,400,000	154,000	79,300
Yahoo	12,400,000	34,600,000	50,300	83,000
MSN beta	22,798,200	65,151,122	107,487	32,205

2) reliability

Past examples of non-reliability of search engines include AltaVista's results count (Notess, 2000). Rousseau (2000) recorded daily fluctuations in the number of results retrieved by AltaVista; these fluctuations were compared to the monotonously growing number of results reported by Northern Light. Bar-Ilan (2000) observed huge daily fluctuations in results retrieved by Hotbot when compared with Snap, where both search tools were powered by Inktomi. Search engine stability over time can be measured by the set of measures proposed by Bar-Ilan (2002b). In a current example, submitting the query "-kxht site:.sr" three minutes apart produced first about 83,000 results and the second time only about 81,800 results at Yahoo.

3) transparency, disclosure, clear documentation

A very recent report (Wouters, 2004) discusses search engines' disclosure practices (mainly related to paid placement and paid inclusion).

Unfortunately search engines do not always report problems. For example, Bar-Ilan (2002) showed that Google did not report the actual number of link pages to a given site that are indexed by it. Only recently Google admitted this (Searchenginewatch forum, 2004). Even when the search engines receive explicit questions they do not always bother to give satisfactory answers, as was the case with Hotbot (Bar-Ilan, 2000). Often pressure placed on the search engines through search engine/webmasters forums result in getting more explicit answers. There are cases when search engines do not retrieve documents indexed by them for some queries even though these documents should definitely appear (Mettrop & Nieuwenhuysen, 2001).

Often the documentation the search engines provide does not reflect the search engines total capabilities. Features that exist are not mentioned, while features that are advertised do not work correctly or are non-existent. For example, Yahoo's linkdomain: retrieves pages linking to a given site (Notess, 2004), however the linkdomain meta-word (an extremely useful feature for link analysis) is not documented at Yahoo (2005).

One of Google's featured operators is OR (Google, 2005). However OR has never worked properly on Google, for example the search for Gretel produced 884,000 results, the search for Hansel 694,000 results, but the search for Hansel or Gretel produced only 607,000 results (by the way Gretel OR Hansel produced 615,000 results).

Sometimes the information is only partial, for example details of the ranking algorithm, but for this Google provides a reasonable explanation (Google, 2004b). On the other hand it is rather unclear what stemming algorithm Google applies: "when appropriate, it will search not only for your search terms, but also for words that are similar to some or all of those terms" (Google, 2003). When is stemming appropriate? Not for singular versus plural: succulent –963, 000 results; succulents – 360,000 results.

It seems that not much importance is attached to the help pages, these pages are often difficult to locate, for example there is no link to help or documentation from the Yahoo search page (http://search.yahoo.com/)

4) timeliness

Sometimes search engines fail to update their indexes often enough, in the past such problems were reported for AltaVista and for Northern Light among others (see for example Olsen, 2001 or Sullivan, 1998). Thelwall (2001b) checked how long it took for some search engines to discover previously unindexed pages that have links to them from pages submitted to the search engines. As for how fast and often pages are being reindexed, consider the Wikipedia entry for Prince Harry, http://en.wikipedia.org/wiki/Prince_Harry_of_Wales. On January 12, 2005 the Prince appeared in a Nazi costume at a dress party. The Wikipedia entry has already been updated twenty five times discussing the controversy by early January 14, 2005. Google had a cached copy from December 21, 2004 (revisited next on January 18, 2005) and MSN beta had a cached copy from January 9, 2005. Yahoo did not cache the specific page.

5) indexing the whole document

Lycos used to index only the titles, header text, and an excerpt of the first 20 lines, or 10% of the document together with a set of keywords extracted from the document. Some sources claim that currently Google only indexes the first 101K of a Web page (French, 2004) – we were unable to find this information on the Google site (again a transparency issue). Our small experiment supports this claim, consider the page http://www.gutlesspacifist.com/gp/archives/2004_04.html, its size is about 154K, it is indexed by Google, and near the bottom of the following text appears: "war should result in a response that includes repentance" (this text appears in the cached version as well). When searching for this phrase, Google retrieved two results, but not the above-mentioned page. Yahoo also cached the page, and retrieved it for the specific phrase; the same was true for MSN beta. For all three search engines, entering the URL of the page shows whether the search engine indexes the page or not.

6) response time, accessibility

Although Google is almost always accessible, in July 2004 it was affected by the MyDoom worm and was down for a few hours (Shim & Kanellos, 2004). Yahoo seems to be limiting the number of search requests per timeframe (French, 2004b), when this limit is exceeded one gets a "server busy, try again later" note instead of actual results.

7) objectivity – no commercial influences and no influence on the environment

Introna and Nissenbaum (2000) discuss extensively the political power large search engines have. Search engines are commercial and have to show profits, thus they will naturally choose to cover more popular topics more extensively, topics for which more advertisement can be sold. Note that the major advertising programs (the "sponsored links") are owned by search engines (Adwords by Google, and Overture by Yahoo). Paid inclusion is another controversial issue – although it does not guarantee placement, it guarantees a certain level of coverage and frequency of updates, which already give the participants in these programs some advantage. Currently only Yahoo has a paid-inclusion program among the search engines discussed here. When the program was introduced in May 2004, it became a heavily debated issue (Sullivan, 2004c).

"Googling" ("to look up someone's name [at Google] in an effort to find out more about them". Whatis.com, 2004) and "google bombing" ("an attempt to influence the ranking of a given site in results returned by Google", Wikipedia, 2005) have become accepted social activities. This together with the florishing search engine optimization and marketing industry (SEMPO, 2004) indicate the influence search engines have on their environment. At the University of Washington, there is even a course on Google ("it has become a social phenomenon," Janes, 2004).

8) all reported results are retrievable

Currently all the search engines discussed in this paper limit the number of results they are willing to display for a given query. Google and Yahoo display 1000 results, MSN beta 500. This problem can be partially overcome by using different portioning techniques (e.g. by domain or date). Date searches can be easily submitted through the "Ultimate Google Interface" (http://www.faganfinder.com/google.html) for Google, however link searches cannot be combined with any other option. AltaVista's advanced search form (http://www.altavista.com/web/adv) can be used for date limited searches for Yahoo (Yahoo powers AltaVista).

9) ranking, different sorting options

Only MSN beta has an option for influencing search result ranking (MSN, 2005) – one can use slides to set the importance of the exactness of the match, of the links pointing to the page and according to the date the page was added to the index. However when searching for "Gaza pullout" and setting freshness to maximum and all the other parameters to minimum, the first result is a news item from December 9, 2004 (cached on January 13, 2005), while the second item is from January 14, 2005 (the day of the searches). Thus it seems that the time is the last time the search engine visited the page.

Different search engines employ considerably different ranking algorithms. A new tool, Jux2 (http://www.jux2.com/stats.php) allows users to compare rankings of the top ten results of Google, Yahoo and AskJeees. Vaughan (2004) empirically compared search engine rankings with human judgment. Different measures for comparing rankings were introduced by Fagin et al. (2003) and by Bar-Ilan, Levene and Mat-Hassan (2004).

10) flexible output display

Some of the requirements are fulfilled by the search engines. One can turn the site collapse option on/off at all three engines. The search engines allow the user to set the number of results per page. None of them employs clustering techniques (like the ones implemented at Vivisimo, http://vivisimo.com/) and the users cannot change the output for individual results. They allow the users to turn filtering of sexual content on/off.

11) cached results
Google, Yahoo and MSN beta offer access to the cached version of the pages. Google and MSN also provide the date on which the page was cached.

12) high quality retrieval in languages other than English as well
The search engines discussed here usually do not employ specific techniques to improve the search results for non-English languages (Bar-Ilan & Gutman, 2005). For German, it seems that Google does employ some additional techniques (Guggenheim & Bar-Ilan, 2005).

13) accessible API
Currently only Google offers an API. There are some rumors on the Web, that MSN will also have a developer API during 2005 (xmlhub.com, 2004).

14) full Boolean searches, diversity of operators
The commercial search engines serve the "public", and the public does not want to use Boolean operators, and when they do they are often used improperly (Jansen et al., 2000). Thus full Boolean searches (allowing the use of parenthesis or some other technique to express compound propositions) are not high on the search engines' agenda. As we saw before, even the standalone OR does not work properly for Google, and parenthesis are meaningless. Yahoo and MSN beta do not say anything about supporting parentheses, but they seem to be applicable. NEAR or ADJACENT is not supported by any of these search engines (AltaVista used to have a NEAR operator). With Google (2004c) an * in a query means exactly one word, thus red ** blue means that red and blue are separated by exactly two words, this option seems to work properly only when the whole expression is within quotation marks.
15) advanced techniques for retrieving data for link analysis

The current capabilities of search engines for retrieving backlinks are worse than what they used to be, when AltaVista and AlltheWeb were still independent services. Google enables to retrieve some of the link pages (some, not all, see Searchenginewatch Forum, 2004) to a specific page only, and this search cannot be combined with anything else. At Yahoo, the undocumented feature linkdomain: works for the time being and it can be combined with other search terms, but this allows retrieving links to pages of a given host, however such a search does not work, if for example we want to study the links leading the Ronald Rousseau's site: http://users.pandora.be/ronald.rousseau/. MSN retrieves link pages targeted to a specific page only.

16) wide variety of search modifiers
Some of the limits, i.e. limiting the search to a certain domain or a language, do exist. They cannot always be combined. Google for example ignores all words after the tenth word in a query, making it difficult to create complicated queries for Google. There is no easy way to run date limited searches from Yahoo, AltaVista's advanced interface is much better for that. Currently MSN provides fewer modifiers than the other two search engines.

17) additional features: stemming on/off, truncation left/right, wildcards, case sensitivity on/off, spell check, site collapse on/off
All three search engines employ some kind of spell checking, allow to turn site collapse on/off, they are all case insensitive, stemming cannot be influenced (there is some stemming at Google, the situation with the other engines is not clear), currently they do not allow the use of wildcards or other means of truncation. Yahoo may also employ some kind of stemming, even though the number of results retrieved for "dog" and for "dogs" is different, in both cases both "dog" and "dogs" are highlighted in the snippets. There does not seem to be any stemming at MSN beta.

18) search assistance: relevance feedback, similar/related pages and searches, personalization
Google provides an option for retrieving "similar pages" to a specific page; however this feature is of limited use. For example when finding pages similar to Ronald Rousseau's English language homepage, we got rather mixed results (probably because there isn't enough text on the page); for Peter Ingwersen's homepage the results were much better, when searching for pages similar to the

Introna-Nissenbaum paper on the politics of search engines, the results were rather mixed again, but slightly better than for Ronald Rousseau's homepage.

Yahoo offers something along the lines of related searches, for wide topics they have a list of queries under "Also try". For tsunami they listed about one hundred, all of them contained the word tsunami. MSN allows to "play around" with ranking (the effectiveness of this option will have to be further investigated). "My Yahoo!" allows personalization, but is targeted towards the Yahoo portal (Yahoo, 2005b).

19) ability to combine all the features in a single query (including unlimited number of search terms)
Google limits the number of search terms in a query to ten. This is a serious shortcoming for Web researchers who try to create accurate queries.

20) non-textual retrieval capabilities
Currently, all three engines offer image search (probably mainly based on textual descriptions). Much research is going on in the area of multimedia information retrieval.

Conclusion

The currently available commercial search engines are rather far from the Web researcher's dream of an ideal search tool. What we need is a powerful, reliable and flexible tool to serve the scientific community. Most probably we have not covered the list of wishes of Web researchers, but we have started the list. One of our colleagues suggested that we call this ideal search engine "Webomet". Now that we have a name for it and a basic set of features - all we need is financing, resources and know-how!

Acknowledgments

We acknowledge the contribution of Lennart Bjorneborn to the list of features, and especially thank Ronald Rousseau who gave us the idea (quite some time ago) to write this paper.

References

Albert, R., Jeong, H., & Barabasi, A. L. (1999). The diameter of the World Wide Web. *Nature*, 401, 130-131.

Ambrogi, R. J. (1995). *Legal research on the Internet. A primer*. Retrieved January 11, 2005, from http://www.legaline.com/column10.htm

Bar-Ilan, J. (2000). Evaluating the stability of the search tools HotBot and Snap: A case study. *Online Information Review*, 24(6), 439-449.

Bar-Ilan, J. (2002). How Much Information Search Engines Disclose on the Links to a Web Page? – A Longitudinal Case Study of the 'Cybermetrics' Home Page. *Journal of Information Science*, 28(6).

Bar-Ilan, J. (2002a). Methods for Measuring Search Engine Performance over Time. *JASIST*, 54(3), 308-319, 2002.

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *IPM,* 41(4), 973-986.

Bar-Ilan, J., & Gutman, T. (2005). How do search engines respond to some non-English queries. *Journal of Information Science*, 31(1), 13-28.

Bar-Ilan, J., Levene, M., & Mat-Hassan, M. (2004). Dynamics of search engine rankings – A case study. In *Proceedings of the 3rd International Workshop on Web Dynamics*, New-York, May 2004. Retrieved January 14, 2005, from http://www.dcs.bbk.ac.uk/webDyn3/webdyn3_proceedings.pdf

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. In *Proceedings of the 7th International World Wide Web Conference*, April 1998, Retrieved January 11, 2005, from http://www.ra.ethz.ch/CDstore/www7/00/index.htm

Brewington, B. E., & Cybenko, G. (2000). Keeping up with the changing Web. *Computer*, 33(5), 52-58.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th International World Wide Web Conference, April 1998. Retrieved January 15, 2005, from http://www-db.stanford.edu/pub/papers/google.pdf

Broder, A., Kumar, R., Maghoul, F., Raghavan. P., Rajagopalan, S., Stata, R., Tomlins, A. & Wiener, J. (2000). Graph structure in the Web. In *Proceedings of the 9th International World Wide Web Conference*, May 2000. Retrieved January, 10, 2005, from http://www9.org/w9cdrom/160/160.html

Cyber Atlas (2000). *Web pages by language*. Retrieved January 15, 2005, from http://www.clickz.com/stats/sectors/demographics/article.php/408521

Faba-Perez, C., Guerrero-Bote, V. P., & De Moya-Anegon, F. (2003). Data mining in a closed Web environment. *Scientometrics*, 58(3), 623-640.

Fagin, R., Kumar, R. and Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1), 134-160.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2004). A large-scale study of the evolution of Web pages. *Software: Practice and Experience*, 34(2), 213-237.

French, G. (2004). *Google indexes document's first 101K*. Retrieved January 13, 2005, from http://www.webpronews.com/insiderreports/searchinsider/wpn-49-20040406GoogleIndexesDocumentsFirst101k.html

French, G. (2004b). *Tool compares Google and Yahoo algorithms*. Retrieved January 13, 2005, from http://www.webpronews.com/insiderreports/searchinsider/wpn-49-20040312ToolComparesGoogleAndYahooAlgorithms.html

Global Reach. (2004). *Global Internet statistics (by language)*. Retrieved January 15, 2005, from http://www.global-reach.biz/globstats/

Google (2003). *The basics of Google search*. Retrieved January 14, 2005, from http://www.google.com/intl/en/help/basics.html

Google (2004b). Information for Webmasters. Retrieved January 14, 2005, from http://www.google.com/webmasters/4.html

Google (2004c). Google Help: Cheat sheet. Retrieved January 15, 2005, from http://www.google.com/help/cheatsheet.html

Google (2005). *Advanced search made easy*. Retrieved January 14, 2005, from http://www.google.com/intl/en/help/refinesearch.html

Guggenheim, E., & Bar-Ilan, J. (to appear). Tauglichkeit von Suchmaschinen für deutschesprachige Abfragen. *Information, Wissenschaft und Praxis*

Huberman, B. A., & Adamic, L. A., (1999). Growth dynamics of the World Wide Web. *Nature*, 401, 131.

Ingwersen. P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The Information Society*, 16, 169-180.

Janes, J. W. (2004). *LIS 598. Google*. Retrieved January 14, 2005, from http://www.ischool.washington.edu/jwj/google/

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of user queries on the Web. *IPM*, 36, 207-227.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632, 1999.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). On the bursty evolution of Blogspace, In *Proceedings of the 12th International World Wide Web Conference*, (pp. 568-576). Retrieved January 15, 2005, from http://www2003.org/cdrom/papers/refereed/p477/p477-kumar/p477-kumar.htm

Kumar, S. R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling emerging cyber-communities automatically. In *Proceedings of the 8th International World Wide Web Conference*, May 1999. Retrieved July 30, 2002, from http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html

Lawrence, S., & Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280 (5360), 98-100.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. *Nature*, 400, 107-109.

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation*, 57(5), 623-651.

MSN (2005). Change your results by using results ranking. Retrieved January, 14, 2005, from http://beta.search.msn.com/docs/help.aspx?t=SEARCH_PROC_BuildCustomizedSearch.htm#sb_sliders&FORM=HEHP

Notess, G. (2000). Search engine inconsistencies. *Online* (March 2000). Retrieved January 13, 2005, from http://www.onlinemag.net/OL2000/net3.html

Notess, G. (2004). Yahoo! Review on Search Engine Showdown. Retrieved January 14, 2005, from http://www.searchengineshowdown.com/features/yahoo/review.html

Olsen, S. (2001). AltaVista serving up out-of-date listings. Retrieved January 13, 2005, from http://news.com.com/2100-1023-274839.html?legacy=cnet

Oxford Dictionary (1989). *Reliability*. Retrieved January, 13, 2005, from http://dictionary.oed.com/cgi/entry/50202002?query_type=word&queryword=reliability&first=1&max_to_show=10&single=1&sort_type=alpha (accessible through subscription).

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, 2/3(1), paper 2. Retrieved January, 13, 2005, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Shim, R., & Kanellos, M. (2004). *Google downed by latest MyDoom*. Retrieved January 13, 2005, from http://news.zdnet.co.uk/internet/0,39020369,39161678,00.htm

Searchenginewatch forum (2004). Google say not reporting all backlinks. Retrieved January 13, 2005, from http://forums.searchenginewatch.com/showthread.php?t=2423&page=2&pp=20

SEMPO (2004). Summary report: The state of search engine marketing 2004. Retrieved January 14, 2005, from http://www.sempo.org/research/SEMPO-Market-Sizing-2004-SUMMARY-v1.pdf

Snyder, H. & Rosenbaum, H. (1999). Can search engines be used as tools for web-link analysis? A critical view. *Journal of Documentation*, 55, 375-384.

Spink, A., & Jansen, B. J. (2004). *Web search: Public searching the Web*. London: Springer.

Sullivan, D. (1998). *Northern Light add search functions, freshens index*. Retrieved January 15, 2005, from http://searchenginewatch.com/sereport/article.php/2166471

Sullivan, D. (2004a). *Major search engines and directories*. Retrieved January 10, 2005, from http://searchenginewatch.com/links/article.php/2156221

Sullian, D. (2004b). *comScore Media Metrix search engine ratings*. Retrieved January 11, 2005, from http://searchenginewatch.com/reports/article.php/2156431

Sullivan, D. (2004c). *Yahoo reawakens the paid inclusion debate*. Retrieved January 14, 2005, from http://searchenginewatch.com/searchday/article.php/3355221

The Statistical Cybermetrics Research Group (2004 a). *SocSciBot3*. Retrieved January 10, 2005, from http://socscibot.wlv.ac.uk/help/tutorial3.html

The Statistical Cybermetrics Research Group (2004b). *The academic weblink database project*. Retrieved January 10, 2005, from http://cybermetrics.wlv.ac.uk/database/

Thelwall, M. (2000a). Web impact factors and search engine coverage. *Journal of Documentation*, 56, 185-189.

Thelwall, M. (2001). A web crawler design for data mining, *Journal of Information Science* 27(5), 319-325.

Thelwall, M. (2001b). The responsiveness of search engine indexes. Cybermetrics, 5(1), paper 1. Retrieved January 13, 2005, from http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3). Retrieved January 15, 2005, from http://informationr.net/ir/8-3/paper151.html

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *IPM*, 40 (4), 677-691.

Whatis.com (2002). *Googling*. Retrieved January 14, 2005, from http://searchwebservices.techtarget.com/sDefinition/0,,sid26_gci799367,00.html

Wikipedia (2005). *Google bomb*. Retrieved January 14, 2005, from http://en.wikipedia.org/wiki/Googlebomb

Wouters, J. J. (2004). *Searching for disclosure: How search engines alert consumers to the presence of advertising in search results*. Retrieved January 14, 2004, from http://www.consumerwebwatch.org/news/paidsearch/finalreport.pdf

WWW10 Panel (2001). Search: Beyond the keyword interface. At *The 10th International World Wide Web Conference*, Hong-Kong, May 2000. Outline retrieved January 15, 2005 from http://www10.org/program/w10-panel.shtml

xmlhub.com (2004). *MSN beta RSS results*. Retrieved January 15, 2005, from http://www.xmlhub.com/rssmsn.php/

Yahoo (2005). *Help: Using meta search words to find specific URLs, sub-pages, link popularity and more*. Retrieved January 14, 2004, from http://help.yahoo.com/help/us/ysearch/tips/tips-08.html

Yahoo (2005b). *What is My Yahoo!?* Retrieved January 16, 2005, from http://help.yahoo.com/help/us/my/my-01.html