VOLUME 9 (2005): ISSUE 1. PAPER 2

Expectations versus reality – Search engine features needed for Web research at mid 2005

Judit Bar-Ilan
Department of Information Science, Bar-Ilan University, Ramat Gan 52900, Israel
E-mail: barilaj@mail.biu.ac.il

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 fulfill the requirements of 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.

Keywords

Search engines, Webometrics, Yahoo Search, Google, MSN Search

Introduction

Even though the Web has existed 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 the 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 (Statistical Cybermetrics Research Group, 2004a) and results of specific crawls are also available for download (Statistical Cybermetrics Research Group, 2004b). Others (e.g. Spink & Jansen, 2004 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 crawl 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, http://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 (http://www.google.com), Yahoo (http://www.yahoo.com), AskJeeves/Teoma (http://www.teoma.com) and the MSN (http://search.msn.com). In this paper we will consider Google, Yahoo and MSN (Currently Askjeeves/Teoma lacks a number of major features for Web research, like links to a page). Exalead (http://beta.exalead.com/search) is a newcomer in the search scenery and it remains to be seen how it develops in the future. 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 comScore (2005), Google, Yahoo and MSN together provide 82.5% of the search results on the Web as of July 2005 (Google 36.5%, Yahoo 30.5% and MSN 15.5%).

The outline of the paper is as follows: first we list the desirable features and explain their importance for Web research. Next we examine whether the three, above-mentioned search engines fulfil these requirements and finally we discuss the implications of our findings. Search engines are changing constantly, thus we want to emphasize that points made about these tools is based on our findings as of mid-August 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.

Expectations - the list of features for an ideal search engine

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 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 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) \((a,ß)\)-currency measure.

5) indexing the whole document

When our intention is to create an exhaustive list of documents, having a certain text-based feature, then even if the 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

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 all searches for a given research 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.

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. The ability to retrieve the whole set of results, and not only the first 250 or 1000, is essential for successful Web research.

9) ranking, different sorting options

Depending on the research topic, we do not always want to look at all the search results, and often we cannot (currently, the search engines do not display the whole set of 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**

The ability to view cached results is another useful feature. It 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**

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, originally developed for retrieval in English, 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 and prefixes the basic machinery for retrieval in English is very far from sufficient. Web research does not necessarily concentrate on English, does information retrieval should work properly in other languages as well.

13) **accessible API**

An 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**

There is no need to expand on the importance of this feature. It 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, i.e., any logical proposition can be expressed as a combination of propositions that are connected by these operators only. These combinations necessitate the use of parentheses. 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). We also need to be able to run standalone NOT searches, in order to estimate sizes.

15) **advanced techniques for retrieving data for link analysis**

The Web is made up of links and nodes; 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 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 & 2004; Bar-Ilan, 2005). Such studies rely on data about links. The most basic feature is links to a specific page. Links to a site or a subsite are often of interest, and more generally we want to be able to define both the set of anchor pages and the set target pages. Sometimes we are interested in the number of pages linking to a given page, site or set of pages, but sometimes the actual number of links are of interest (i.e. if there are several links from the source to the target set, we want the tool to count that page with multiplicity). 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 to get information on the placement of the links.

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), and 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

The additional features listed above help the Web researcher in phrasing the queries even more accurately. Truncation is very important for languages with compound words and/or prefixes and suffixes. These features are important for studying linguistic aspects of the Web or for co-word analyses.

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

These features are “nice to have”, they help the Web researcher the same way as they are supposed to assist the general public.

19) ability to combine all the features in a single query (including unlimited number of search terms) or to build sets, based on previous results (“building blocks” fashion)

These are very important features; they enable the researcher to create more accurate queries.

20) non-textual retrieval capabilities

The paper concentrates on text-based retrieval, but additional media has to be taken into account as well. The specific capabilities needed for multimedia searches for Web research are beyond the scope of the current work.

Reality - search engines features in the past and in the present

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 researches published by Bharat and Broder (1998) and Lawrence and Giles (1998, 1999), such claims
were not made any more, and we simply cannot expect the search tools to be exhaustive. However, “search engine size wars” (Sullivan, 2005) continue to rage, with Yahoo claiming that it indexes 19.2 billion pages and Google disputing this claim (Price, 2005). Cheney and Perry (2005a, 2005b) compared the number of results returned by Google and Yahoo on about 10,000 queries with less than 1,000 results, they ran the test twice: in the first study Google returned 270% more results than Yahoo, while in the second experiment they found that Google produces 65% more results than Yahoo (the second test was conducted in order to overcome some unspecified methodological concerns).

Another problem is the non-uniformity of coverage. Snyder and Rosenbaum (1999) have shown that the even the relative coverage of major domain of different search engines is 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:.hu” (without the word kxht and in the Hungarian domain) in Yahoo and Google, and “site:.hu -(kxht)” in MSN on January 13, 2005 and again on July 17, 2005. We had to exclude a very rare word, because Google did not support standalone searches for sites at those dates. A third round of searches were carried out on September 3, 2005, when all three engines supported queries of the type “site:.hu” (or “site:hu” for MSN). The results for a number of domains appear in Table 1. It is easy to spot the relative differences in coverage of domains between the search engines and also between the rankings of the search engines for coverage, when considering the domains one-by-one. An additional issue is the huge changes in the number of reported pages between the two data collection points, Google and MSN report much fewer Hungarian pages the second time, while Yahoo reports more than twice as many pages and this number doubled again when we searched for the third time. Google and MSN report more pages from Djibouti than from Suriname in the first two rounds, while Yahoo’s results show the opposite. Google and Yahoo seem to have caught up with MSN on the number of Canadian pages indexed by them, and by the third round Yahoo exceeds by far MSN’s coverage of Canadian pages (at least this is what is reported).

Table 1: Domain coverage

<table>
<thead>
<tr>
<th></th>
<th>January 13, 2005</th>
<th>July 17, 2005</th>
<th>September 3, 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungary (hu)</td>
<td>13,300,000</td>
<td>5,740,000</td>
<td>5,100,000</td>
</tr>
<tr>
<td>Canada (ca)</td>
<td>32,400,000</td>
<td>63,500,000</td>
<td>58,400,000</td>
</tr>
<tr>
<td>Djibouti (dj)</td>
<td>154,000</td>
<td>141,000</td>
<td>62,400</td>
</tr>
<tr>
<td>Suriname (sr)</td>
<td>79,300</td>
<td>104,000</td>
<td>104,000</td>
</tr>
<tr>
<td><strong>Yahoo</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungary (hu)</td>
<td>12,400,000</td>
<td>26,000,000</td>
<td>15,366,935</td>
</tr>
<tr>
<td>Canada (ca)</td>
<td>34,600,000</td>
<td>79,900,000</td>
<td>66,897,959</td>
</tr>
<tr>
<td>Djibouti (dj)</td>
<td>50,300</td>
<td>51,900</td>
<td>65,838</td>
</tr>
<tr>
<td>Suriname (sr)</td>
<td>83,000</td>
<td>101,000</td>
<td>38,714</td>
</tr>
<tr>
<td><strong>MSN beta</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungary (hu)</td>
<td>22,798,200</td>
<td>15,366,932</td>
<td></td>
</tr>
<tr>
<td>Canada (ca)</td>
<td>65,151,122</td>
<td>64,853,877</td>
<td></td>
</tr>
<tr>
<td>Djibouti (dj)</td>
<td>107,487</td>
<td>45,927</td>
<td></td>
</tr>
<tr>
<td>Suriname (sr)</td>
<td>32,205</td>
<td>29,711</td>
<td></td>
</tr>
</tbody>
</table>

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 continuously 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).

Often the number of reported hits changes as we move further down the results list. For example we searched Google for “digifeed” on August 29, 2005 six times between 13:51 and 14:00 (GMT+3). At first, it reported 697 results, and then we moved to the full set of results (including the initially omitted ones). This time the number of recorded hits was 744, but by the time we moved to the last page, this number decreased to 571. After capturing all the screens, we rerun the query, this time the number of results were 1030, 1070 and 856 respectively (see Figure 1). The fluctuations are probably being caused by different servers handling the different requests, but still this is a serious reliability problem when carrying out Web research.

![Figure 1: Various Google results reported for the same query less than 10 minutes apart](image)

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) has showed that Google does not report the actual number of link pages to a given site that are indexed by it. Only recently has 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). Sometimes pressure placed on the search engines through search engine/webmasters forums result in getting more explicit answers. Search engines do not report why they do not retrieve documents indexed by them for some queries where these documents should definitely appear (Mettrop & Nieuwenhuysen, 2001). Consider, for example the query link:www.umu.se/infsrk/ISSI2005/ at Google. It produced 11 results (10 displayed) (see Figure 2) on August 29, 2005. The corresponding query at Yahoo, link:http://www.umu.se/infsrk/ISSI2005/, produced 96 results (95 shown). The second and third results (we wanted to pickup at least two URLs to cover for the undisplayed result) on Yahoo’s page (see Figure 3) are both indexed by Google (see Figures 4 and 5, for the cached versions of these pages by Google, that clearly show the link to the ISSI 2005 website).
Figure 2: Google’s results for the link query

Figure 3: Yahoo’s results on the link query

Figure 4: Part of the cached copy of Chaomei Chen’s homepage with a link to the ISSI2005 Conference (cached on 26 August, 2005)

Figure 5: Part of the cached copy of the ISSI2003 homepage with a link to the...
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), see Figure 6.

Recently Yahoo's link and linkdomain options do not operate correctly (at least one of them - see discussion at the Searchenginewatch Forum (2005)). As an example compare the number of results Yahoo reports for the query link:http://www.huji.ac.il (536,000 results) vs. the query linkdomain:huji.ac.il (309,000 results) as of August 29, 2005 (see Figures 7 and 8). Something is obviously wrong here, since all links pointing to the Hebrew University's homepage (http://www.huji.ac.il) should definitely be returned for the linkdomain query as well. However we are unable to check what is really going on, since the search engine displays only 1000 results. We also tried a number of sites having a small number of inlinks (in order to try to understand what is going on), but we were unable to reproduce the questionable results in these cases.

Figure 7: Yahoo's results on link:http://www.huji.ac.il

Figure 8: Yahoo's results on linkdomain:huji.ac.il

Sometimes Google seems to be a little weak on "search engine math". For example, on August 29, 2005 within 5 minutes we carried out the following queries:

- Peres – 2,930,000 results
- Rabin – 1,230,000 results
- Rabin Peres – 138,000 results
Rabin OR Peres – 2,280,000 results (this should be around 2,930,000+1,230,000-138,000=4,022,000 assuming the previous numbers are correct, but in any case the number of results should be greater or equal to 2,930,000)
· Peres – Rabin - 1,540,000 results (this was expected to be 2,930,000-138,000=2,792,000)
· Rabin – Peres – 948,000 results (this is quite near to the expected number of results 1,230,000-138,000=1,092,000)

The results (with the clock showing the time each search was carried out) are displayed in Figure 9.

One reason for the problematic numbers reported by Yahoo or Google could be that since the number of results is clearly higher than the number of documents displayed, the numbers are only rough approximations. As we said before, we were not able to reproduce the link vs linkdomain problem with Yahoo for small domains. For Google, however, we were able to find a "small" example as well. For the queries "digifeed" and "transnova" Google reported 792 and 998 results respectively on August 30, 2005. The AND query returned only 4 results, however for the OR query 3650 results were reported, which is much higher than the sum of the two single word queries (1790) – again something is wrong here (see Figure 10). Veronis (2005a) provides a possible explanation for Google's strange responses to Boolean queries, however he seems to suggest (Veronis, 2005b) that these problems were already solved by the end of March 2005. We carried out the problematic searches at the end of August 2005.

MSN also has problems with inclusion-exclusion. Consider the following example, again from August 30, 2005: USA (207,967,566 results), Israel (33,424,669 results), USA OR Israel (180,406,246 results) as can be seen in Figure 11. It seems that search engines always had difficulties with simple math, Ingwersen (1998) reported that the number of results AltaVista reported for the query a AND b is not identical to the number reported for the
query b and a.

Figure 11: MSN results on USA-Israel queries

Sometimes the documentation is only partial, for example, there are no details of the ranking algorithm, but for this Google provides a reasonable explanation (Google, 2005a). 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, 2005b). When is stemming appropriate? Seemingly, not for singular versus plural as can be seen in Figure 12.

Figure 12: succulent vs succulents

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 at that time. Now for a very recent example, Hurricane Katrina (currently hitting New Orleans) – there’s an active Wikipedia page for it from August 26, 2005 - http://en.wikipedia.org/wiki/Hurricane_Katrina. The current version at
the time we checked the search engines’ copies was from August 29, 2005, 18:48. Google had not yet indexed the page (but when we rechecked on August 30, 2005, the page was already indexed by Google as well) MSN and Yahoo both had a cached copy from August 27, 21:04.

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 (Sullivan, 2004d) – we were unable to find this information on the Google site (again a transparency issue). Our small experiment supported this claim in January 2005 when we looked at page http://www.gutlesspacifist.com/gp/archives/2004_04.html, its size is about 154K, and it was 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. When we rerun the test in August 2005, the 154K long page did appear as a search result, which seems to indicate that Google indexes more text from a page than before. This time we tried the Project Gutenberg EBook of Ethics by Aristotle (http://www.gutenberg.org/dirs/etext05/8ethc10.txt) – a page and several of its mirror sites are indexed by Google and its size is about 662K, however all cache version stop at 514K (abruptly at the middle of a sentence), and if we search for a phrase below this point, we are not able to find it, while for phrases above this point, the searches are successful. Interesting to note, that for the query "Greek: hae peri ta aethae pragmateia aen dikaion esti prosagoreuein politika" (very near the end of the book), pdf versions of Aristotle’s Ethics were retrieved by Google, even though these pdf files are 730K long. Thus it seems that as of mid 2005, Google indexes around 500K of text and html pages, but for pdf files the limit is higher. Duffez (2005) already reported in April 2005, that Google might be indexing more than the first 101K of a page. MSN indexes the 7-bit version of the ebook (http://www.gutenberg.org/dirs/etext05/7ethc10.txt) which is 662K long, and retrieves the Greek phrase from the end of the book. Even though Yahoo indexes both the 7 and the 8-bit version, none of these were retrieved by it when searching for the above-mentioned Greek phrase. It only indexes the first 525K of these files. The search engines do not report the maximum length of the text they index from a page, thus we can only speculate about this point.

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, 2004), when this limit is exceeded one gets a "server busy, try again later" note instead of actual results. An additional problem can be that the search engine does not report /display all the search results because of timeouts that are not reported or noticeable to the user (see AltaVista’s email answer to Alastair Smith’s query (1999)).

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

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. The users try to “please” the search engines, thus instead of being unobtrusive tools, the search engines become influential players on the Web.
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, that are more advertised. Note that the major advertising programs (the “sponsored links”) are owned by search engines (Adwords by Google, and Overture by Yahoo). Van Couvering (2004) questions whether “the current system of search provision online serves the ‘public good’”. 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. Bar-Ilan (n.d.) characterized the competing pages, the Wikipedia entry on Jew and the homepage of jewwatch.com. Googling, googlebombing, together with the flourishing search engine optimization and marketing industry (SEMPPO, 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 250. This problem can be partially overcome by using different portioning techniques (e.g. by domain or date). Date searches could be easily submitted through the “Ultimate Google Interface” (http://www.faganfinder.com/google.html) for Google, however the date limit option of Google (daterange:, using Julian date format) stopped working some time ago (again not reported). It is still documented in the Google API reference (Google, 2005e), thus it is possible that it operates through the API. Figure 13 illustrates this, the same number of results are returned with and without the daterange: operator for the query informetrics; daterange was limited to June 1-September 1, 2005 and the searches were carried out on September 1. Google’s advanced help only lists the date: modifier (allows one to search for “references” in the last 3, 6 or 12 months (Google, 2005c). This option limits the results in some way, but not correctly, as can be seen in Figure 14: the number of results in the last 6 and 12 months is considerably less than the number of results in the last 3 months (!). All the searches were carried out within three minutes, and were rerun to verify the numbers. A similar search at Yahoo provided non-contradictory answers. Link searches cannot be combined with any other option at Google, thus even the daterange work-around (had it worked) would not help to break the results into smaller chunks in order to get more than 1000 results. AltaVista’s advanced search form (http://www.altavista.com/web/adv) can be used for date limited searches for Yahoo (Yahoo powers AltaVista).

Figure 13: Informetrics without and with the daterange: operator
9) ranking, different sorting options

Only MSN 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 “Katrina” and setting freshness to maximum (see Figure 14) and all the other parameters to minimum, the first result is an item from August 30, 2005, while the next few items are from August 31, 2005 (the searches were carried out on September 1, 2005 at 9:00PM GMT+3). The top results of the “freshness” search are slightly different form the results of the regular search for Katrina.

Different search engines employ considerably different ranking algorithms. Dogpile has a very nice tool (http://comparerecherchesinines.dogpile.com/index.aspx) that allows users to compare rankings of the top ten results of Google, Yahoo, MSN and AskJeeves. They recently published a study on the overlap based on more than 22,000 searches (Dogpile, 2005). The study was carried out by a team lead by Amanda Spink and Jim Jansen. 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; n.d.).

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 MSN and Yahoo (through its AltaVista search engine).
interface). The search engines allow the user to set the number of results per page. None of them employ clustering techniques (like the ones implemented at Vivisimo, http://vivisimo.com) and the users cannot change the output format for individual results. They allow the users to turn filtering of sexual content on/off.

11) cached results

Google, Yahoo and MSN offer access to the cached version of the pages. Google and MSN also provide the date on which the page was cached, which is an additional useful feature.

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 Google (http://www.google.com/apis) and Yahoo (http://developer.yahoo.net) offer APIs.

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. Even results using “simple search engine math” (+/-) do not seem to work properly at Google (see Figure 16). It is expected that the number of results for the query “Karolinska” equal the sum of the number of results for the query “Karolinka Stockholm” and Karolinka – Stockholm”, however the results are very different (823,000 vs 1,370,000).

Figure 16: An example of Google’s “search engine math”

Yahoo and MSN 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). Google has a * operator (2005c), red * blue means that red and blue are separated by one or more words. Interestingly, less than a year ago (Google, 2004) the * used to mean separated by exactly one word, thus it was meaningful to search for “informetrics * bibliometrics” vs. “informetrics * * bibliometrics”. The new meaning of the * operator, as far as we understand, is that it only excludes “bibliometrics informetrics” and “informetrics bibliometrics” (as phrases) from the set of results for informetrics bibliometrics (not a phrase search this time), however this was not the case, and we couldn’t figure out the meaning of this operator.
NOT (or a "-" before the search term) is an operator, that is useful as a standalone operator. Currently both MSN and Google support it, but Yahoo does not.

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. Then the link: operator retrieved pages with links to the specified page and in case the target was a directory, all pages linking to any page (indexed by the search engine) in this directory, although this was not clearly stated in their help pages (AltaVista, 2002; AlltheWeb, 2003). Google enables to retrieve some of the link pages (some, not all, see Searchenginewatch Forum, 2004; Bar-Ilan, 2002b) 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 (Callishain, 2004) 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/, and not just the homepage. Recently, MSN (2005b) also added the linkdomain: operator.

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 32nd word in a query (used to be ten words only). MSN apparently has a 10 word limit in queries (not documented), for the query "1 2 ... 11" no results, but the page http://www.timeanddate.com/calendar/?year=2004&country=1 is indexed by it, and the cached page also contains all the numbers between 1 and 11 (see Figure 18). Yahoo does not seem to have a query length limit. All search engines limit word length as well (see Notess, 2005b). There is no easy way to run date limited searches from Yahoo, AltaVista’s advanced interface is much better for that. Currently MSN does not allow date limited searches.

It is very useful to be able to search within specific sites or domain, with or without additional search terms (see for example Table 1). Such searches combined with link searches often appear in webometric research. For example the query "linkdomain:issi-society.info -site:issi-society.info" returns all the pages that link to pages of the ISSI society, excluding self-links (which are often created for navigational purposes). Such query is not possible at Google for two reasons, it does not have a linkdomain operator, and even if we substitute linkdomain:issi-society.info for link:www.issi-society.info/ (the homepage of the society’s site), we are unable to exclude self-linking pages, since the link operator cannot be combined. This is a serious limitation for conducting link based research.
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, they are all case insensitive, stemming cannot be influenced (there is some stemming at Google, Google, 2005b), 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 (see Figure 18). There does not seem to be any stemming at MSN. From the researcher’s point of view it is best that the search engine lets her decide whether to turn stemming, case sensitivity and truncation on or off. Currently one can choose whether site collapse is to be on or off at MSN and at Yahoo, if the AltaVista interface is used (maximum two pages from a site).

Figure 18: Both “dog” and “dogs” are highlighted when searching for “dog” at Yahoo
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, but is seemingly improving, compared to our previous experience with this feature. On September 3, 2005, it gave us reasonably related pages to Ronald Rousseau’s homepage, but when looking for related pages to Wikipedia’s definition of “chunnel” (the nickname of the English channel), the results were very disappointing (See Figures 19 and 20).

Yahoo offers something along the lines of related searches, for wide topics they have a list of queries under “Also try”. For “information” the full list of related searches included more than one hundred suggestions, including “cook county jail mates information”, but not including “information retrieval” (probably people search more for “cook county jail mates information” than for “information retrieval” (See Figure 21, the search was conducted on September 3, 2005). MSN allows to “play around” with ranking (the effectiveness of this option will have to be further investigated).

Personalization is a hot topic. Google recently rolled out its “Google Personalized Search” service, where personalization is based on the user’s search history. This feature may be useful for Web research as well (as a way to get “better” pages among the first 1000 retrieved), however its usefulness will have to be explored in the future.

Figure 19: Pages related to Ronald Rousseau’s homepage
19) ability to combine all the features in a single query (including unlimited number of search terms) or to build sets, based on previous results ("building blocks" fashion)

Google and MSN limit the number of search terms in a query (although Google recently raised the limit from 10 to 32. This can be a shortcoming for Web researchers who try to create accurate queries. Google does not allow complex Boolean searches (use of parenthesis) and does not allow combining the link: operator with any other search term. Google allows to "search within" the previous results, but this is only another way of adding additional search terms to the query, and not to build upon previous result sets, as available in some search systems (e.g., Dialog or Web of Science).

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. The non-textual search capabilities that are of interest to Web researchers are outside the scope of the current paper.
Conclusions

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. Lennart Bjorneborn 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!

Some words about the reliability of this study. We saved all the searches, and were careful to state the date (and often the time) the searches were carried out. The searches are extensively captured and documented in this paper. This was necessary, since almost certainly it will be impossible to reproduce the exact results we received for our searches at the reported time. We are ready to provide the interested reader with the raw data for this study.

The majority of the examples use Google. The reason for that was that since Google is currently the most popular search engine (comScore, 2005), most of our attention was concentrated on its performance. In the follow-up studies we plan to put the other search engines "under the microscope" as well. Even though the search engines have shortcomings, we use them both in our everyday lives and for research, because they are the only large-scale information retrieval tools from the Web that are currently freely available to all.

We intend to revisit the available features on a periodical basis (say once a year) and to publish follow-ups. A preliminary version of this paper appears in the Proceedings of ISSI 2005 (Bar-Ilan, 2005b), where we surveyed search engine capabilities as of the beginning of 2005. Comparing the two versions we can already see differences and developments, indicating that periodical feature evaluation is indeed needed. The Web and the search engine scenery change constantly, thus in the future it is quite plausible that new features, useful for Web research will become available.

We recommend establishing a set of tests with clear methodological guidelines about running and documenting these tests, for evaluating search engine performance on a periodical basis.

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


Dogpile (2005). Different engines, different results. 


<http://www.google.com/webmasters/4.html> (accessed August 29,
2005)


<http://www.ischool.washington.edu/jwj/google/> (accessed September 1, 2005)


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).


Yahoo (2005). Help: Using meta search words to find specific URLs, sub-pages, link popularity and more.