Search Engines And Effective Information Retrieval

Sam Berner
May 23, 2001

1. Worms, Spiders and Other Crawlies with Faces

The Internet explosion is both good and bad news. The good one is that there is now an increasingly greater pool of information available to everyone than was, say, twenty years ago. The bad news is that it is still very hard to find it, despite the affirmations to the contrary by all those wonderful techno-geeks who keep creating "sophisticated search tools" by the score, in an attempt to produce the one that will do it all. As the World Wide Web continues to expand that very expansion threatens to overwhelm the user and make it difficult, if not impossible, to locate material of interest. "*Trying to find information on the Internet is often described as trying to get a sip of water from a fire hydrant*." (**McGuire, et al., 1997**). The "sophisticated tools" a.k.a. search engines, are supposed to help one change the hydrant into an eco-friendly garden sprinkler.

So what exactly is a search engine? For one, they are not at all as "sophisticated" as their creators would want potential buyers to believe. The enchanted spell (scripted code) for one of the oldest search engine (ALIWEB) is online, in public domain (http://aliweb.emnet.co.uk/). Other simple codes can be found in various books (e.g. **Sonnenreich & Macinta, 1998**). According to **Maze, et al. (1997**), a typical search tool is a rather simple software application.

A generic search tools is composed of a **robot** (also known as spider, crawler, or worm), an **index**, a **database**, and the user **interface**. They troll the Web, collecting information on new and updated sites (**Strom & Tittle, 1997**). The basic functions of a generic search tool are:

1. finding web pages
2. harvesting web pages and building an index
3. searching the index with a user query
4. providing the user interface

Robots are essential ingredients of all current Web search tools [http://www.its.bldrdoc.gov/projects/t1glossary2000/_knowledge_robot.html]. The robot is a software program located on a specific machine, and it uses HTTP to perform intelligent tasks such as retrieving World Wide Web documents and indexing references [http://www.kayak.ie/gaeilge/display_resour.htm]. Most search tools speed the process of discovery by employing multiple robots of different types. The more robots crawling the Web, the better their chances of keeping up.

Their more common tasks include:

1. discovering web pages for inclusion in a database
2. indexing web pages
3. measuring the size and scope of the Web
4. maintaining a database of Web pages by checking old links for updates and relocation
5. mirroring sites
6. generating visual and navigatable maps of relationships and hierarchies in collections of documents
7. providing statistical and owner information for visited sites
8. provide "push" technology for users **(Chun, 1999)**

Users never really come into direct contact with any of the parts of Web search tools that we have discussed so far. Robots, indexes, databases, and search engines all live and operate on the server. The **interface** is the part of the search tool with which users directly interact. *"In the graphical environment of the Web, interface design has to do with constructing visual meaning. The happy marriage of architecture and interface--of logical structure and visual meaning--creates a cohesive user experience. This marriage is crucial to helping users get around on the Web"* (**Flemming, 1998**). The interface interprets for the user the search engine's abilities. It provides the place for entering the query into the engine's language and potential actions. From here the engine searches the index and reports back its findings the interface presents the results to the user in one of a variety of possible formats. Yet, although this is a crucial part of the tool, it is also the most malleable and thus the most subject to change. From the perspective of the creators of a Web search tool, the user interface is the simplest component. The creation and maintenance of indexes and search-and-retrieval software consumes far more time and resources than the series of Web pages and simple code and scripts that the search tool contributes to the interface. The interface is far more complicated, however, than the server's end of things implies; it ultimately plays as large a role in the effective retrieval of information as any of the other components I have discussed.

The interface ultimately defines the set of potential user queries, in the same way that the search engine defines the set of possible searches and the index defines the available information. The search engine may be able to perform a certain type of search, but if the interface provides no way for the user to request that search, it will never happen. **Westera (2001)** provides a good comparative overview of the user interfaces of the eight topmost search engines freely available online. However, as **de Wet (2000)** found in his research, the evaluation of search engine interfaces is a very subjective exercise, as searchers tend to develop brand loyalty after getting used to a certain search engine.

The query page of the interface is a static file created beforehand and accessed over and over by thousands of users for millions of searches. Even in interactive pages to which the user can add or delete extra command lines. The interface is a standard shell – the same home page at a fixed address for every user who wanders by. Every user needs a starting point, and the starting point has to be prepared for any user. On the other hand, since the search tool cannot know the content and order of the search results before a search is submitted, there is no fixed page for the user to view the results. To display the search results, the search tool creates new temporary pages on the fly, by plugging the search results into a template of HTML codes. These pages are typically a set of citations, each composed of, at minimum, a hypertext link in the form of a Title or URL. This link is to the named page itself, not to a copy of the page in the search engine's database. Thus, if the page linked to the results has been moved or deleted, or if its home-server is unavailable or too busy, or if any of a number of other errors occurred, the user may not be able to get to that page (**Maze et al, 1997**).

Different user interfaces might require very different formats for the same search of the same database, even with the same search engine. Each interface can offer its own combination of operators, symbols, and punctuation marks (or lack thereof) independent of the index and search engine. Some interfaces let the user include traditional Boolean operators in the command line. Operators like ***AND***, ***OR***, ***NEAR*** and ***NOT*** connect words to form a search statement, or query. The browser submits the search statement to the server, where the search engine transforms it and compares it to the index. Some search tools offer symbols (e.g. ***+, -, /, &***) in place of word operators. Some interfaces allow natural language phrases instead of, or in addition to, Boolean options. Other interfaces offer more than one command box for a single search. In addition, there may be special boxes and menus for such limits as date ranges and file formats. Similar boxes may look for contents only in specific fields, such as Title and URL. Users must pay attention and routinely look for the appropriate syntax for the interface to a new search tool. Sometimes the same syntax carries different meanings in different tools. Even a familiar tool may change its interface overnight.

2. Where Agents Have IQ

As amounts of information available via networks and databases rapidly and continually increases, current search engines are limited in their ability to satisfy users needs. Intelligent Agents (IA) have proven to be the needed item in transforming passive search and retrieval engines into active, personal assistants. **Jansen (1996)** provides on overview of the use of intelligent agents and their use in search engines. According to **Yee (1991)** search engines on the Web, and on-line databases have problems with their retrieval engines. Users have a number of problems interacting with on-line databases: finding appropriate subject terms, a large number of hits along with failure to reduce the retrieval sets, zero hits and failure to increase the retrieval sets, and failure to understand the cataloging rules. In addition, lack of understanding of the indexes, types of files, and the basic database structure leads to the use of articles (i.e., the, a, etc.), stop words, placing the author's first name before the last name, and hyphenation problems. Although this is a ten-year old research, and despite the improvements in search technology, problems faced by users of online resources have not changed much.

Intelligent Agents don't work for the CIA. They are computer programs that help their users with repetitive tasks while still accommodating individual habits (**Roesler & Hawkins, 1994**), i.e. they perform tasks for users based on the users' preferences. Examples of the tasks they perform are filtering mail, scheduling appointments and locating information. Once a procedure has been specified by the user, the IA can then perform the task autonomously, as well as mimic the user's steps when there is no procedure in place. They can traverse networks and engage in complex patterns of two-way communication with other users and agents, thus accomplishing more complex tasks . IA also have decision-making capacity, making inferences, choosing among different strategies and planning tasks.

Jansen (1994) stipulates that an autonomous, intelligent agent can "rapidly" customize a search or retrieval engine query's result. The agent uses both user preferences and information content of the document and query. The end product of this research will be an autonomous, intelligent agent that resides with an existing search or retrieval engine. This combination will result in improved information retrieval performance for the user. Currently Ask Jeevs and Copernic, among others, use intelligent agents in their search engines. The newest IA on the market is the Swiss MARVIN, an IA-based search engine for medical databases (**Emmen, 2001**). Google, the 2000 winner of Wired Readers Raves Award, also uses an IA (**Google, 2000**)

How are intelligent agents different from regular search engines. The main difference is that an agent is more interactive and can perform many tasks at different locations. First of all, for example, if one searches using a search engine one may get a list of matches, which one might have to follow and possibly not get the relevant information. Secondly, using a search engine may increase the percentage of those matches that might not be relevant to the inquiry. However, if one were to use an agent, the agent could submit one's keyword(s) to many different search engines and follow those corresponding links and gather the information without any intervention from the user.

An intelligent agent uses such technology as the spider, which is also used in the traditional web search engines. However the spider will be a tool which will be used and trained by the user to search the web for specific types of information resources. The agent can be personalized by its owner so that it can build up a picture of individual likes, dislikes and precise information needs. Over time, an agent will build up an accurate picture of a user information needs. It will learn from past experiences, as a user will have the option of reviewing search results and rejecting any information sources, which are not relevant or useful.

Intelligent Agents come in handy when discussing the performance of Meta-Search Engines.

3. The Meta-lie

Although I recommend that one never stop after searching with just one Web search tool, I also recommend that one avoid Web sites that search multiple tools, i.e. such meta-search engines as Meta-Crawler and Internet Sleuth. Unlike the individual search engines and directories, meta-search engines do not have their own databases; they do not collect web pages; they do not accept URL additions; and they do not classify or review web sites. Instead, they allow the user to type in a search, then submit that search to several Web engines at once. This may appear to be a perfect solution, but it is not as every search tool interprets queries differently. None of the meta-search sites linked to multiple search tools circumvents all of these differences. The user is almost invariably much better off searching each tool separately. The exception would be Copernic, which is a freely downloadable software using a very powerful Intelligent Agent, and providing excellent results to user queries. simply pass one's search terms along, and if the query contains more than one or two words or very complex logic, most of that will be lost. It will only make sense to the few search engines that support such logic (see table of general search engine features).

The **UC Berkley Library** (2000) describes to forms of meta-searching tools: one that has a single query box on its interface, and one they terms as "pseudo-meta-searching" which provide a collection of search boxes for different search engines or a drop-down menu that let's one choose which one among a list of search engines to search.

Although meta-search engines were welcomed enthusiastically by the IT community, even favourable reviews poinetd out that they were far from being able to solve users problems (**Dreilinger & Howe, 1997**). Meta-Search engines are useful if one is looking for a unique term or phrase (enclose phrases in quotes ""); or if one simply wants to test run a couple of keywords to see if they get what one wants (**Liu, 1998**). Most meta-search engines only spend a short time in each database and often retrieve only 10% of any of the results in any of the databases queried. Their development lags behind the technical improvements of individual engines, and a number of search engines do not even allow a meta-searcher anywhere near its databases (an example is NorthernLight). Regardless to warnings by professional searchers, ezines such as *Online* actually promote meta-search engines as an antidote to *"the time-consuming practice of sequential search engine searches"* (**Garman, 2000)**.

4. Finding the Tooth-pick In the Bushland

There are three main ways of searching for information online:

1. Not so long ago, one could search for information on the Net by "surfing" and making educated guesses about which links to follow from a few select Web pages to find the information one needed. As early as 1997 this method was considered entertaining, but not an activity a reference searcher would engage in (**Schankman, 1997**)
2. Subject Directories provide a second, much more useful methods of finding pages dealing with specific subjects.
3. Search Engines – use automatic indexing software to discover, harvest and index Web pages.

The first two search options will be covered elsewhere. This paper concentrates on search engines.

A typical search engine will use one of two search strategies:

1. A depth-first strategy: goes to the page attached to the first link of the starting page, then to the first link on the second page, then to the first link on the third page, etc., evaluates each of these either to store it for a further visit or to discard it. The evaluation is done on the basis of various criteria (indexable text, encountered before, etc.). This continues until the robot reaches a dead end (dead link or page with no links), when the robot takes a step back into the page it visited last and takes on the second link on that page. The depth first strategy creates a relatively comprehensive database on a few subjects.
2. A breadth-first strategy: the robot takes one step backwards after each step forward: it follows the first link on the starting page, then returns to follow the second link on the starting page, and so on till it reaches the last link. It then moves on to the first link on the second page, and so on. Breadth first builds databases touching more lightly on a wider variety of documents.

Neither of these two strategies actually performs predictably over the Web as a whole, although either strategy may be employed to different ends for greater effects. If we wish to create a subject-specific index, our best bet would be to employ a depth-first strategy with an upper bound on the degree of depth. On the other hand, if we wish to build an index that samples the wide variety of the Web, we might well choose a breadth-first strategy, with biases toward short URLs and URLs from previously unexplored servers.

Most queries consist of one or more subject keywords or a natural language statement, either with or without some type of Boolean operator (such as **AND**) to connect the parts. Once the user types and submits the query, the search engine takes over. The search engine's job is to match the query to the appropriate words in the index. Matching is done either through keyword searching or concept searching. Keyword searches have a tough time distinguishing between words that are homonyms or synonyms, with truncation and verb endings (see for example **Jones (1999)** discussion on linguistic search engines such as Dr-Link and Tagrget, which use conceptual searching). Unlike keyword search systems, concept-based search systems try to determine what the query means, not just what it says. In the best circumstances, a concept-based search returns hits on documents that are "about" the subject/theme you're exploring, even if the words in the document don't precisely match the words you enter into the query. Search engines that return concept based results employ intelligent agents (**Barlow, 2001**).

Once the engine finds matching words in the index, it compiles a list of URLs, orders them in some way, and returns the URLs and corresponding summaries to the user. The order in which URLs are presented in the list are often determined by some type of relevancy ranking, an attempt to move the "hits" most likely to be relevant toward the top. This is far from perfect, however.

Some search engines use a complicated set of ranking algorithms to produce a large, but well-ordered set of results. Rather than retrieving only pages with two or more of the three keywords, they would retrieve every page that contained any one or more of the keywords; however, its ranking process would move every page containing all three words to the top of the list, followed by every page containing two, then every page containing one of the query terms. The assumption is that users would prefer to see pages containing all of their query keywords first, regardless of whether their search was broad or narrow, while having the option of looking at a more complete set.

According to *Online* ezine, the basic premise of relevancy searching is that results are sorted, or ranked, according to certain criteria: the number of terms matched, proximity of terms, location of terms within the document, frequency of terms (both within the document and within the entire database), document length, and other factors. More recently, several search engines have begun considering factors such as the number of links made to a page or the number of times a page is accessed from a results list. In the highly-competitive search engine industry, ranking algorithms are closely-guarded company secrets. Most search engine producers, however, give a general description of criteria they consider in computing a page's ranking "score" and its placement in the results list (**Courtois & Berry, 2000**).

There are pros and cons to every method of searching and ranking. Many search tools are designed based on the premise that Boolean language is too archaic and complicated for most of their users; instead these tools provide the opportunity to search with natural language techniques. Boolean queries are certainly inappropriate for some users, especially novice searchers. Many searchers routinely use natural language queries even when such queries are not supported; they reasonably (but mistakenly) assume that a machine can be questioned like a reference librarian . On the other hand, automatic indexing – even with recent advances – is still in its infancy. Experienced searchers will rarely want to give up control of the search process to the machine, especially when Web search tools provide the user with so little information as to how their indexes are built and searched.

Further, every method of statistically ranking documents is only useful in certain contexts. Each of the approaches we have mentioned could present the same set of results in an entirely different order. This is one reason that the same search often produces quite different results with different search tools. This also means that the Help or FAQ information provided for a specific search tool can be misleading and uninformative. *Caveat emptor*!

Search Engine Watch [http://searchenginewatch.com/facts/] provides a comprehensive list of tips for good searching. Good things to remember when searching are:

1. *Have a good vocabulary*
2. *Use Boolean searches*
3. *Check the syntax*
4. *Check for stopwords*
5. *Never stop with one search – use different permutations*
6. *Avoid Meta-tools*
7. *Avoid date searches – one never knows which date the search engine is ooking for*
8. *Be cautious*
9. *Use a variety of tools*

5. The Diary of a Disenchanted Searcher

Ideally using a search engine to find Web pages (or documents) of interest would be as simple as selecting a particular search engine, typing in a query (i.e., search request), and obtaining the results. But as many have discovered (**Want, 1999**) in the real world this does not happen. Much more likely is a response in which thousands and thousands of documents are listed as matches or hits, though in fact most have little or nothing to do with one's query.

Another fact to keep in mind is that no one search engine, however powerful it may be, can cover the entire Web. The Web is just too large and complex and evolving too rapidly for this to happen. So if one fails to find what one is looking for using a particular engine, try another. This does not mean that one search engine is necessarily better than another; it mainly points to the fact that different search engines use different means of exploring and indexing the Web. As **Craigmile and Wohrley (1997**) have found in their paper on cyber-searching, the sheer speed with which the Internet changes renders results from any particular search obsolete at once.

If the keyword used in the search is a common word, or has multiple meanings, the relevancy ranking mechanism in a search engine will often produce totally irrelevant results. While many of the search engines look for keywords in the <TITLE>, URL or the <HEAD> tag, this allows unscrupulous publishers to put innocently looking terms in those tags, just to pop up on every search. Many pornographic sites use this methodology, and AltaVista used to be one of the worst places to do any kind of research, as 50% of what it pulled off the cyberspace were X-rated websites. Meta-tags are now being used less than, say, two years ago, and search engine owners are learning how to increase the list of their stopwords.

Ideally, Help documents explain the syntax and features of a search tool. But most search tools have lousy Help files, something that should not surprise anyone familiar with online Help in other contexts (**Barlow, 2001**). Nevertheless, anyone who uses a Web search tool for any important searches should get into the habit of checking available Help. In particular, users should look for information about available operators, options for how results are ranked, and defaults for how the tool interprets searches.

One of the main problems with Help documentation in the tools we review is that the Help authors frequently seem unfamiliar with the tool. In other cases, the Help information is accurate but disorganised. Good Help files (they do exist) usually center around examples of how to construct a query (especially how to fix a bad one), with careful descriptions of the process of retrieving and evaluating results. An even better Help file would describe both how the index is built and what the search tool's shortcomings are none that we have encountered so far includes such a feature. One of the problems with the increasingly commercial nature of Web search tools is finding an altogether honest one. Users should always check the Help files, then try to verify what they say.

If one gave the same query to each of the major search engines, one is likely to get very different results (**Craigmile and Wohrley, 1997**). The reason for this is that each of the engines has a unique query structure. A query that is optimal for one engine may be abysmal for another. This doesn't mean that the poorly performing engine is bad; it simply means that the query that was given to it was poorly formed for that particular engine. Often, if one repeats the experiment with another query formulated differently, a different engine will get the most relevant results.

One of the most important things to remember is that most users have absolutely no idea about any of this stuff (**Sonnenreich and Macinta, 1998**). In fact, most users go to one engine, type in a few keywords, and see what the results look like. If they find what they want, great. If not, they'll modify their query a little, or try another engine with the same query until something useful comes up. Most users don't have the time or patience to learn the ins and outs of any one particular engine. The truth is, one can generally find what one is looking for with a simple search on Yahoo! (which is not even a search engine).The only times the differences between the engines become important is when one is looking for a very specific piece of information or if one is trying to find quality pages on a topic that is very popular.

6. Search Economy?

Commercial Web search tools presently have two major methods of generating revenue: selling advertising space, and selling the robot software for use on corporate intranets. An obvious third method is charging users for the privilege of searching and retrieving results, either with a fee for each transaction or a subscription for unlimited access. Although **Maze et al. (1997)** did not see this as likely to develop into a viable option for Web indexes, simply because search engines are providing automated discovery and indexing of a hodgepodge collection of documents. The Web is currently too shallow and chaotic a source-and robotic indexing of it too unreliable-for Web search tools to charge a significant amount for usage, a recent article has found that a number of content providers online are looking into this as a viable option **(Dix, 2001).**

Even though the fee-based model is the historical precedent for commercial databases in other environments, Web indexes have not proven themselves worth a fee to their primary audience. Librarians and other professional online database searchers comprise the traditional customer base for fee-for-use database providers such as Dialog. The vast majority of Web users, however, do not search the Web for a living (especially with their jobs hanging on the quality of the results). In fact, neither do most professional searchers, although this is already changing, as original research and other valuable information gradually appears on the Web (**Basch, 1996)**. Web search tools do not yet provide essential content with a quality to match that of traditional fee-based vendors. Obviously, this is not to imply that they are not valuable, just not worth paying for.

An increasingly popular advertising twist ties a company banner to a particular keyword in the search tool's Web index. Every time a user searches with the keyword, the company's ad will appear at the top of the results page. This strategy allows advertisers to improve their chances of reaching a likely customer. Other innovative strategies will become standard practice in the near future if commercial Web search tools are to survive, especially since many advertisers are already beginning to wonder if their returns justify their spending (**Gauntlett, 2000**).

Some search tools have been tempted to try a different means of advertising, by selling positions in user search results. Since this is already a major issue for business websites, a number of companies are offering advice on how to rank highest on search engines [see, for example, **Sizzler Studios (2001)**]. Sort of the evil stepsister to tying banner ads to keywords, this approach basically makes profit a factor in the ranking algorithms. An advertiser could purchase the insurance that if a user search retrieved its Web page from the index, it would automatically receive a high position in the user's results page. There are even software packages for Web-developers, which scour the net's search engines, and find one's pages and then compile a report stating where each page ranks The software then submits and periodically resubmits the pages in question until they get listed, automatically, with no danger of oversubmitting. This software uses "stealth" technology so search engines can't tell that the scouring is not done manually by a human agent [http://www.topdog.com/topdog/info.html]. Such applications sell because many of these automated search engines have delays of up to two months to have a site entered into them - if indeed it ever makes it into the index.

It is still possible for nonprofit and educational institutions, especially universities and research groups, to regain strong presence in Web searching. WebCrawler, Lycos, and HotBot all began as projects by graduate students and computer science departments. Such institutions might continue to be the launching pad for new generations of search tools, even if they all go commercial eventually. Yet none of the present generation of commercial engines will be returning to their educational roots, except perhaps to buy innovations. Further, all of these tools have significant leads on upstart rivals-large indexes of Web pages, which inevitably require lots of time and money to create. Future nonprofit projects will have to incorporate increasingly innovative and powerful technology in order to compete for popular attention.

The most likely scenario is that the current crop of search tools will continue to expand their portfolios of services and products; to spread their eggs among more than one basket. Portals are already part of a few search engines, and they add to the economic variety, although in my opinion they also render user interface crowded and messy.

If the increasing size of the Web proves to be too much for general search tools to handle well, it seems likely that an increasing number of smaller tools will arise that will seek to add more value to specific pieces of the Web [an example of the already extensive list can be seen at http://www.leidenuniv.nl/ub/biv/specials.htm]. As we have seen, searching the general tools routinely produces results sets with many thousands of hits, most of them irrelevant. An interesting article of how major search engines were unable to provide information on a political celebrity shows the need for specialised searching tools (**Guernsey, 2001)**. As the problem worsens, smaller tools may woo users by focusing on and doing a better job with small subsets of the Web Users tired of wading through dozens of irrelevant hits could instead search an index that provides more sophisticated retrieval of Web documents with a common subject. The large general tools might decide to downsize and focus their efforts, rather than expand their services. The problem with this scenario is that the more specialized the tool becomes, the more specialized are the needs it can conceivably satisfy. Specialized tools will draw a smaller audience (or at least fewer searches), since they are designed for specific needs. This limits such a tool's appeal to advertisers, making it more difficult to generate revenue.

Besides software and ad space, search tools may soon begin selling another valuable commodity-demographic information and the search habits of their users. We mentioned earlier that several search tools already tie ads to specific keywords, in an attempt to target advertising to specific users. In a similar move, search tools will begin keeping track of the topics that a user has searched for in the past. Such information could be sold to advertisers, who could then target users with banner ads or e-mail solicitations related to their past searches.

One of the most obvious implications of competition between Web search tools is that they have little incentive to be forthcoming about their abilities and weaknesses. They also have a negative incentive to refer users to competitors, even when referral could be to the user's benefit. We do not mean to imply that any of the current search tools consciously lie about themselves to their users, nor that it is a necessary consequence of doing business. With few exceptions, though, search tools do a far better job of promoting than of explaining themselves. Help files too frequently contain more hyperbole and sly jabs at rivals than thoughtful instruction. Since so much of the search process is hidden from the user (the harvesting process, the indexing, search algorithms, ranking), it is quite difficult to learn how a particular tool behaves. Further, the Web is too elusive to form a solid basis for comparison.

All institutions try to protect their own existence, just like individual people. Search tools with stockholders, customers, products, services, and brand names have powerful incentives to continue to exist and to grow, thus creating conflicting interests within the organization. On the one hand, competition and rapid innovation in technology create a desire to focus efforts on research and development, to constantly improve the service. On the other hand, increasing size and the accumulated capital investments in past efforts create inertia and a desire to shift efforts to maximize the return on what has already been done. Not surprisingly, companies try to do both. Search tools' frequent tinkering and tweaking of their interfaces are perfect examples of this synthesis. It is a lot easier and cheaper to add graphics and better designs to the interface than to rebuild the Web index, overhaul the search engine, or develop new harvesting and indexing robots. Cosmetic changes are also far more obvious to the user, who is inclined to judge the tool by the interface too much anyway. An appearance of continuous novelty and innovation is a big victory for a Web enterprise of any sort, often bigger than actual novelty and innovation. But the real benefit to the user is usually less than other kinds of improvements would provide. Think of it as "virtual improvement."

7. Where Would You Like To Go….Tomorrow?

In a research conducted in 1998, a Business Week/Harris poll found that Internet users spend as much as 50% of their time online searching for information (**Frauenfelder, 1998**). As result of this and similar researches, new improvements have been added to search tools to make them more user-driven (ranking higher sites that have been accessed more often, as well as sites that come up as a result of repeated queries). From Eliza to the current search engines, technology has moved in leaps.

Can the current search engines continue to cope with the size of the web, and if not, what are the alternatives? Estimates place the coverage of the web by search engines to be no more than 40-50% of total web pages. Even more concerning is the way that the human-edited directories are falling behind their listing of new sites or updating dead links, while the economic justifications of employing more people to cope with the demand seem unfeasible. In fact, many of the major search tools are facing financial challenges in maintaining a viable service for web users and are having to look at new revenue streams, including charging websites for inclusion.

Sullivan (2001) predicts that the current financial worries faced by the search engine software market are part of a continuous cycle, due to the proliferation of these tools. According to him, search engines will be born and die, either fading away slowly, or crashing like the Big Bang. **Williams (2000)** postulates that web engines will continue being more interested in making money than in websites, and therefore that heavy-weight companies will prevail over new websites which won't be even indexed. This may render search engines more than useless as search tools, and making the Internet even more chaotic. **Butler (2000)** complains that current search engines are letting the scientific community down, precisely because the rank subjects of mass interest much higher than the obtuse scientific ones. He is hopeful that the new XML (eXtensible Markup Language) will allow more restrictive parameters to be incorporated on search engines. A year later, XML is nowhere there, and nothing indicates that it will magically make search engines more intelligent, just as corporate portals did not prove to be the killer applications that automatically made information manageable (**Brewer, 2001**). Google has proven to be an excellent search engine in comparison to other hopeless monsters such as Infoseek and Lycos, and yet I personally can't say that it is perfect, or that it produces reasonably less irrelevant hits than any other search engine. Its ranking scheme is better, it caches terms and that's about all there is to it. Research carried out in 1999 by two scientists from NEC Research Institute in Princeton, N.J, found that the percentage of the Web indexed by all the search engines had dropped from 60% to 40% between 1996 and 1998 (**Lawrance & Giles, 1999).** Kris Carpenter, director of search products and services for Excite, has been quoted saying in that same year that her company purposely ignores a large part of the Web because of a lack of consumer interest and that the future of search engines lies not in bigger indexes but in more specialized ones, in which having everything on a given subject could be indexed and displayed to viewers (**Dunn, 1999**).

Four years ago, **Pat Ensor (1997)** complained about library users not being as savvy information finders as librarians themselves. This year, the librarians are complaining that the search engines are stealing people from the libraries (**Mayfield, 2001)**. With the current information status (one tries avoiding the term explosion), more and more "lay" people are doing what for ages only librarians knew how to do well: searching and finding information for themselves. Since information hunting is becoming decentralised, it might be a good idea to educate the user, rather than bemoan the current technology situation. Search engines are marvelous pieces of information-seeking software, but despite assertions to the contrary (see discussion in Appendix) they will never become librarians. Intelligent agents or otherwise, these pieces of scripted code are dumb in comparison to humans. Taking the risk of sounding like a Luddite, my final words would caution putting too much hopes in software, and directing all efforts at teaching Mr. Layman to lower his expectations to the reasonably workable level. If all fails, he will end up switching the modem off, and will shuffle to the nearest library.

On a more optimistic note, **Brewer (2001)** says in his recent article that search engines have "*led to increased overall productivity by millions of workers, and thus to our recent global economic expansion*". He obvioulsy gives up on any form of information organisation, subscribing to the idea that information finding (and therefore solution to information chaos) is a redeeming enough aspect for search engines to forgive them all other mishaps. The article considers search engines superior to databases, on account of their ‘fuzzy logic”. Brewer (2001) hopes that XML and distributed systems and intelligent software will enhance search engines in the future.

8. References

Barlow, L. 2001. A Helpful Guide to Web Search Engines. [Online]. Available WWW: http://www.monash.com/spidap4.html

Basch, R. 1996. Secrets of the Super Net Searchers. Information Today Inc, New York.

Brewer, E. 2001. *When Everything is Searchable*. **Communications of the ACM**: 44(3) pp. 53-55

Butler, D. 2000. *Souped-up search engines*. **Nature**: 405(6783) pp. 112-115

Chun, T. Y. 1999. *World Wide Web Robots: An Overview*. **Online and CD-ROM Review**: 23(3) pp. 135-142

Courtois, M. P. & Berry, M. W. (1999). Results Ranking in Web Search Engines. [Online]. Available WWW: http://www.onlineinc.com/onlinemag/OL1999/courtois5.html

Craigmile, B. & Wohrley, A. 1997. *Searching Cyberspace*. In **The Cybrarian's Manual**, Pat Ensor (ed.). American Library Association, Chicago.

de Wet, N. 2000. Evaluating Search-Engine Design: A Study in Human Computer Interaction. [Online]. Available WWW: http://people.cs.uct.ac.za/~ndewet/hci.html

Dix, J. 2001. *How much would you pay to visit your favourite Web sites?* **Computerworld:** 24(40) p. 16

Dreilinger, D. & Howe, A. E. 1997. *Experiences with Selecting Search Engines Using Metasearch*. **ACM Transactions on Information Systems:** 15(3) pp. 195–222.

Dunn, A. 1999. Search engines finding less on growing Web. [Online]. Available WWW: http://seattletimes.nwsource.com/news/nation-world/html98/inet_19990708.html

Emmen, A. 2001. Medical intelligent agent search engine Marvin: a prototype Grid applications. [Online]. Available WWW: http://www.hoise.com/primeur/01/articles/monthly/AE-PR-05-01-19.html

Ensor, P. 1993. *Why Can't a User Be More Like a Librarian?* **Technicalities**: May, pp.10-12

Flemming, J. 1998. Web Navigation: Designing the User Experience. [Online]. Available WWW: http://www.oreilly.com/catalog/navigation/chapter/ch05.html

Frauenfelder, M. 1998. The Future of Search Engines. [Online]. Available WWW: http://www.directhit.com/about/press/articles/industry_std.html

Garman, N. 2000. Meta Search Engines. [Online]. Available WWW: http://www.onlineinc.com/onlinemag/OL1999/garman5.html

Gauntlett, D. 2000. Basic Web Economics: How things work in the 'attention economy'. [Online]. Available WWW: http://www.newmediastudies.com/economic.htm

Google. 2000. Google Wins Wired Readers Raves Award. [Online] Available WWW: http://www.searchengineguide.com/pr/20001013_pr2.html

Guernsey, L. 2001. Mining the 'Deep Web' With Specialized Drills. [Online] Available WWW: http://www.nytimes.com/2001/01/25/technology/25SEAR.html?pagewanted=all&searchpv=tech (needs registration to access)

Jansen. J. 1996. Using an Intelligent Agent to Enhance Search Engine Performance. [Online]. Available WWW: http://www.firstmonday.dk/issues/issue2_3/jansen/

Jones, K. 1999. *Linguistic Searching Versus Relevance Ranking: Dr-Link and Target*. **Online and CD-ROM Review**: 23(2) pp. 67-80

Lawrence S. & Giles, C. L. *Accessibility of information on the web.* **Nature**: 400(6740) p. 107
Liu, J. 1998. *Guide to Meta-Search Engines*. **BF Bulletin (Special Libraries Association. Business and Finance Division)**.107:17-20.

Mayfield, K. 2001. Ask the Librarian, Not Jeevs. [Online] Available WWW: http://www.wired.com/news/culture/0,1284,40308,00.html

Maze, S. et al. 1997. Authoritative Guide to web Search Engines. Neal Schuman, New York

McGuire, M. et al. 1997. The Internet Handbook for Writers, Researchers and Journalists. The Guilford Press, New York

Roesler, M. & Hawkins D. T. 1994. *Intelligent agents: software servants for an electronic information world*. **Online**: July, pp. 18-32

Schankman, L. 1997. *Beyond Surfing: Serving Information to Our Patrons*. In **The Cybrarian's Manual**, Pat Ensor (ed.). American Library Association, Chicago.

Sizzling Studios. 2001. The Truth About Search Engines. [Online] Available WWW: http://www.sizzlingstudios.com/nl/01apr/searchengines.htm

Sonnenreich, W. and Macinta, T. 1998. Web Developer.com Guide to Search Engines. John Wiley & Sons, New York.

Strom, D. & Tittle, E. 1997. Maintaining Your Website. [Online] Available WWW: http://www.lanw.com/myw/text.htm

Sullivan, E. 2001. The End For Search Engines? [Online] Available WWW: http://www.searchenginewatch.com/sereport/01/02-theend.html

The Library, University of California, Berkeley. 2000. Meta-Search Engines. [Online] Available WWW: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html

Want, R.S. 1999. How To Search The Web. Want Publishing, New York

Westera, G. 2001. Comparison of Search Engine User Interface Capabilities. [Online] Available WWW: http://lisweb.curtin.edu.au/staff/gwpersonal/compare.html

Williams, S. 2000. Will Search Engines Let Us Down? [Online] Available WWW: http://www.cowleys.com.au/news/archive/20000218.htm

Yee, M. M. 1991. *System Design and Cataloging Meet the User: User Interfaces to On-line Public Access Catalogs*. **Journal of the American Society for Information Science:** 42(2) pp. 78-98.