South African Journal
of Information Management
[DoE accredited]

Student Work Vol.4(3) September 2002

Fee vs free: order vs chaos?

Merle Ruff
(Anglo American Information Centre, South Africa)
Post Graduate Diploma in Information Management
RAU University
mruff@angloamerican.co.za

Contents

1 Introduction

Information retrieval via the World-Wide Web represents a complete reversal of the traditional fee-based commercial databases. In the fee-based environment, a single search strategy can be used to simultaneously search many different sources. Currently, it is impossible to use a standard search strategy when searching the Web. The information on the Web is stored in digital format on Web servers all over the world. In its unstructured state it would seem as if the originators of Web content do not want the information to be found!

Industry vendors and Web developers have attempted to improve the retrievability of Web documents. These have included the development and constant refinement of powerful search engines and directories. In addition, advanced search tools which track information stored in Web formats have been developed to assist the serious searcher. While these features improve information retrieval, the results still do not compare with the degree of precision and relevancy that has become the trademark of the fee-based database clusters.

More recently, attempts to improve the situation have focused on ways to organize the Web. These include methods of structuring and breaking down the different elements of digital information stored on a Web page. Portalization is also viewed as a mechanism to structure access to Web content. Coupled with this approach are the current initiatives to standardize the manner in which Web documents are stored. The HTML protocol for publishing Web content was the first attempt towards standardization. This has been enhanced by recent developments in the XML language and Metadata standards to organize

Web content.

This article focuses on the following problem:

- What are the major differences between searching on commercial databases versus the Internet?

The sub-problems include the following:

- What are the problems with information retrieval using free search tools?
- What steps need to be adopted to alleviate these problems in the future?
- Are we moving towards a more 'structured' Web?

top

2 Comparison between commercial on-line databases and the World-Wide Web

The appearance of the first fee-based on-line commercial databases in the 1970s was in itself revolutionary. It provided searchers with on-line access to the world's published literature. This development has enabled searchers to access reliable information sources stored anywhere in the world from remote locations anywhere in the world.

The Web consists of Web pages in digital format found on Web servers all over the world connected together by the Internet. The Internet, which became more and more popular during the 1990s, like the appearance of the on-line database aggregates, has resulted in an information revolution such that searching for information will never be the same again!

The on-line database aggregates provide information access to a privileged group of individuals or professionals belonging to organizations that can afford to pay for information. Skills to use these databases require training and experience. Alternatively, information on the Web is available to anyone who has access to a personal computer with a modem and a telephone connection. Searching on the Web 'appears' to be very simple and user-friendly.

The information stored on commercial databases contains bibliographical references to printed information stored in various locations. Although some commercial databases contain information in full-text formats, the information is still stored in a text-based format.

What implications do these differences have on techniques for searching on-line databases vs free Web-based search engines? What are the differences in the nature of information sources (content, structure or size) found on the on-line databases vs the Web?

2.1 Search techniques for retrieval: commercial on-line databases vs search engines

Table 1 summarizes this author's findings with regard to the major differences between search techniques using traditional databases vs the Web.

Table 1 Comparison of search techniques

	Traditional on-line and CD-ROM databases	**World-Wide Web**
Database	Structured and	Unstructured and no

	authoritative	validation
Indexing	Controlled vocabulary Manual indexing	Largely automatic indexing (except for directories). Robots have different strategies for indexing
Fields for limiting	Author, title, descriptor, document type, date	URL, title, header, date, weighted words
Abstracting	Clear and concise summary by humans	Automatically generated. Often not clear or adequate summary
Search	Query-based search Mouse-based. Follows hyperlinks.	Cannot repeat or formalize search
Evaluation criteria	Relevance, precision and recall	Difficult to apply traditional measures. Other criteria: speed, relevance ranking of results, validity of links and lack of dead links, Web interface

2.1.1 Structure vs unstructured

In contrast to on-line databases, CD-Roms and even traditional libraries, the Web is a place where information is unstructured as well as unfiltered. On-line databases contain references to published material that has undergone strict review or has been published in reputable periodicals and newspapers. Although the Web contains a wealth of valuable material, it also contains much unsolicited material and what Bates (1999:xxiv) refers to as 'dreck'. The 'noise' that one encounters searching the Web is infinitely greater than that found on the commercial databases. Therefore the Web searcher must be much more vigilant in limiting numbers of results by using the filters listed in the above table. Most users are unaware of these filters and give up in frustration when presented with a hefty hit list.

Search engines have at least 'borrowed' Boolean logic, which has been used by on-line database searchers for the last 30 years. While the Boolean 'and', 'or' and 'not' can be used in search engines, most search engines have not understood the use of proximity searching. AltaVista is the only search engine that offers this facility. The biggest problem with search engines is that although they use Boolean logic there is no single standard for the way in which you enter your Boolean command. There is one standard with standard terms and punctuation for entering a Boolean search for a commercial database host.

2.1.2 Precision and recall

In 1975, Sarecevic referred to the two most important criteria for evaluating search results (Dong and Su 1997:78), namely precision and recall. Recall measures the number of relevant documents retrieved out of the total number of relevant documents indexed in the system. In traditional databases, recall represents the completeness of a search. Every document is indexed using controlled vocabulary and all documents are searchable. However the number of the Web pages that could be indexed is infinitely larger than commercial databases. Search engines are unable to index or retrieve all the potentially available information.

According to Leighton, 'the Web is as large and unstructured as one is likely to find, so recall is meaningless' (Dong and Su 1997:78). The only criteria for evaluating Web search results are user satisfaction with the completeness of a search. Dong and Su (1997:75) also refer to the fact that 'rapidly increasing Web size and the limited coverage of search engine databases have made recall a difficult measure to apply'.

Precision is used to measure the number of relevant documents retrieved out of the total number of documents retrieved from traditional on-line databases. Precision ratios have been used in evaluating search engine retrieval. Dong and Su (1997:78) claim that 'the output relevance to a user's query is a very important indicator for judging an engine's quality and intelligence'. The problem arises when this measurement is applied. How is precision evaluated. Is it based on the top 10 hits or the top 20 hits? The problem with this approach is that ranking algorithms differ from one search engine to the next. How does the Internet's ranking, which influences the order in which the hits are displayed and hence selected by the user, compare with the user's ranking of search results?

There is a greater degree of certainty as to the relevance and completeness of a search performed on a commercial database. According to Bates (1999:xix) the arrival of the Web has made business research both more complex and easier. While on the Web, one is guaranteed of being able to get hold of for example annual reports of companies or statistics, but it is 'harder to know when you have conducted a reasonably thorough search'.

2.2 Structure of on-line bibliographic information sources vs digital Web-based information sources

2.2.1 Database structure

Commercial databases consist of defined fields with standardized indexing and retrieval mechanisms. Humans program the type of indexing to be applied to each field. The Web consists of words in a document. The words are indexed and not the subjects. Although subject directories are categorized using a method of classification, each subject directory uses its own unique classification system.

2.2.2 Size and coverage

How much of the Web is covered by search engines? It is estimated that there are 500 billion pages on the Internet. Google covers 1 billion pages, therefore representing less than 1% of the Internet. Commercial database hosts, for example Dialog, cover over 570 databases. When a search is performed on Dialog, the researcher is searching the database's entire coverage. The question here is whether database size or accuracy of retrieval of the total coverage is more relevant. The answer must be the latter, since precision and relevancy is the aim of the serious researcher.

In contrast to the commercial databases, the Web contains a lot of 'dreck' (Bates 1999:xxiv). Material on database hosts consists of carefully selected material published by scholarly publications. Lebedev concluded that search engines were no good at finding scientific information. In 1996 the Internet had only 10 to 20% of the documents he could find on INSPEC (Dong and Su 1997:80).

2.2.3 Invisible Web

Search engines miss a tremendous amount of information stored on the Web. This is due to the fact that search engines are 'barred' from retrieving relevant information contained within databases on the Web, paying information, video and audio material. There are

technical reasons for this, particularly the fact that search engines cannot process non-text information, sounds, image, and information stored in Web-accessible databases. The information that search engines are prevented from seeing is referred to as the "invisible Web". In section 4 strategies for finding information on the Invisible Web are discussed

2.2.4 Information content

There is a 'blurring' of commercial databases and Web search engines on the Web. Web sites can consist of Web-enabled versions of previously published books or CD-Rom databases. The commercial database hosts have made their databases searchable on the Web. In this instance, the researcher is benefiting from the best of both worlds. The accuracy of on-line database indexing is linked with the user friendliness of mouse (click and point) interfaces. The horrors of modem searching are a thing of the past. The search statement remains unchanged. However, Web technology is used to search on-line database content in a considerably more user-friendly manner.

top

3 Evolution of the Web: directories, search engines, meta search engines and development of search techniques

Searching on the Web has evolved along with the enhancement of Web search technology. This section discusses the evolution of Web directories, search engines, Meta search engines and intelligent agents.

The first generation search engines created indexes by automatic 'spidering' of Web sites and analysing location and frequency of words. These search engines match words against a search statement without considering how the pages interrelate, that is the context and syntax. Web directories continue to create their indexes manually. The third way of retrieval matches a search statement with the location and frequency of the words. The search engine then rates the relevancy of the results based on the Web sites that have been most used.

Natural language searching was developed to overcome the problem of search engines' lack of consideration of 'the syntactical relationships between search terms and other vocabulary within their indexes' (Green 2000:128). Ask Jeeves, launched in June 1998, was the first of the natural language search agents. Ask Jeeves' search engine matches the user's query against a database of seven million templates of questions. If no match is found, it presents the nearest alternatives. Thereafter, the user can select the option on Ask Jeeves to conduct a meta search across AltaVista, InfoSeek, Lycos and Yahoo.

Another development is 'links-based' analysis. Google is a prime example of a search engine that uses this technique. The Google search engine matches a search statement against the '1-billion or so hyperlinks that weave the Web together' (Green 2000:128). The results are then ranked in importance depending on how many other sites link to them. The theory is that if a Web author has included links to other sites that are considered important, then some form of editorial judgement has been exercised. Furthermore the Google search engine also processes the text around the hyperlink and therefore claims it can analyse far more Web sites than humans who build subject directories. According to Green (200:129), 'in fact unlike search engines that become less useful the larger the index, Google claims to return even better results with a bigger index'.

Google estimates that through this method of links analysis it can reach 300 million Web pages. Green (200:129) claims that 'Google's combination of extensive reach and greater

accuracy of results is its advantage'.

As search engines become more sophisticated, attempts have also been made to create specialized and authoritative collections of Web content. These attempts are exemplified by the appearance of hubs, newsgroups, subject specific directories and intelligent agents.

Hubs are Web pages that guide the user to a list of authoritative sources, for example Focused Crawler. Focused Crawler uses a 'classifier' that evaluates Web page relevance and a 'distiller' that identifies relevant hypertext nodes that point to relevant pages within a minimum amount of links. Newsgroups are the results of individuals or experts in a field who share their knowledge and opinion in specific subject areas. Specialized newsgroup search engines, for example Deja News, are important to searchers needing to seek out experts to solve specific problems. Subject specific directories are Web-enabled versions of commercial databases.

Intelligent agents are sophisticated retrieval mechanisms that shift the power away from Web servers and on to the desktop, leading to greater search and retrieval capability. Agents can search across a wide range of document types and formats. They act as 'true infomediaries' (Green 2000:131). Copernic translates a single search statement for different search engines and simultaneously submits the search statement to the search engines, Web directories and databases.

top

4 Developing a search strategy for the Internet

The above discussion has highlighted some facts about Internet searching. Given our understanding of the nature of digital information stored on the Web, what should we take into consideration when developing search strategies for information retrieval on the Web? With the advent of the Web, searchers are confronted with the choice of using the free or fee-based information sources. There are circumstances when the world of published information on structured on-line database hosts will satisfy a query. However there are situations when Web-based digital sources will provide the 'best' information.

The Internet has led to an increase in the choice of sources that provide information. On the one hand finding information on the Internet is easy and fast. For example, if one needs current weather conditions for anywhere in the world it is easy enough to find. However, if one wishes to research a particular subject, unless one knows which search engine to use, or knows the URL of a specific Web site, one will feel overloaded and frantic by having to browse through hundreds of listings trying to find useful (relevant) information.

4.1 Why develop a search strategy?

Owing to the shifting landscape of information products and services, one should adopt guidelines on when the Web is the best source of information. The fact that searching the Internet comes with its frustrations does not to detract from the fact that there is a lot of good information available on the Web if the tools are used with understanding. It is a well-known principle that search strategies are essential in both the fee-based and the free Web-based environment. This means analysing the aspects of a query, identifying the best sources and formulating an effective search strategy. If the Web is the 'best' source, then search tools whether they are search engines, subject directories or meta search engines need to be selected according to the query.

4.2 What goes into a search strategy for the Web?

Calof (2002:7) has developed an approach for searching the Web, which he refers to as the 'Searching Smarter' approach. Table 2 was compiled by this author to illustrate Calof's seven stages of a search strategy developed for the Web.

Table 2 'Searching Smarter' search strategy

S	**Specify information needs** This refers to first establishing and analysing what it is you are looking for, what information specialists refer to as the 'reference interview'
M	**Match information sources to information needs** Decide which is the 'best' source for retrieving the information you need. If the Web is the 'best' source, then proceed to the next level
A	**Assess Internet approaches** Once the decision is to use the Web, one can choose whether to use a directory, search engine or meta search engine. Decide this based on what the tool accesses, and how it works
R	**Recognize search engine differences** This requires gathering information on how to enter search strategies on different search engines etc.
T	**Think search statements and strategy** This means asking the right question – formulating your strategy with the words that may be used in a Web document
E	**Execute the search** Formulate your search statement, using all the operators, limiters and syntax offered by the search tool
R	**Refine the search** Searching on the Web is an iterative process. Refine the results until you get closer to what you are looking for

Other approaches to searching the Web recommend using a query-based approach. Emphasis is placed on knowing what you are looking for and matching the information need to the appropriate Web tool that will 'best' retrieve the information.

4.3 Effective Web searching: tips of 'super searchers'

A growing number of researchers have become 'super searchers' on the Web. This section focuses on their knowledge of Web sources and their tips for getting 'better' results from Web searching.

When to use the Web as the 'best' source of information
Muchin (1999:152) says that one should 'become a free thinker'. He is referring to the fact that one can often find answers to questions quickly and inexpensively using the tools on the Web. The Web is an excellent source for:

Free news sources and on-line versions of major newspapers

Business and corporate information

On-line shopping

Reference sources on the Web: on-line dictionaries (www.onelook.com) and links to

common encyclopaedias (www.mygo.com), and has links to Britannica and Encarta

Telephone directories

Government documents on government Web sites.

4.3.2 Invisible Web

Price and Sherman (2001:32) define the invisible Web as content that is 'largely hidden from search engines'. There are many professional searchers who have perfected techniques to find this content. To get to Web sites that search engines cannot access, they listen and follow the literature to identify new sites regularly. To get to the home pages of these databases, Price and Sherman suggest that one include the term 'database' in the search statements. Yahoo directories are full of links to invisible Web sources. They insist that relevant Web sites should be bookmarked as soon as they are identified. They conclude that 'once you've built your own virtual reference collection, finding what you're looking for on the Invisible Web will be as easy as using a search engine to navigate the visible Web' (Price and Sherman2001:34).

There are also 'specialized search engines' that access the invisible Web (Calof 2002:29). Calof lists over 20 search engines that access some parts of the invisible Web. Some examples are annotated subject directories, such as the Librarians' Index to the Internet (www.lii.org) and Google's search engine for discussion groups (groups.google.com). Calof recommends the Sherman and Price directory.

4.3.3 Challenges of Web searching

During Bates's (1999:39) interview with Linda Cooper, she commented that with Web searching, 'after you think you've come full circle …a whole new world opens up'. Cooper says that the only way to know when to stop searching on the Web is to refer back to the reference interview and give the client what has been asked for.

top

5 Organizing chaos: development of structure and standardization on the Web

Individuals and businesses want to avoid searching multiple sources of information using different search engines tools. There is a growing need to access a single source that provide all the relevant information. This has led to the idea of creating a 'one stop shop' where information can be collected and accessed from a single point. The information technology industry is starting to address this requirement with their development of Internet portals.

In 2002, the Web is still unstructured and disorganized, unlike the fee-based environment. Users are desperate for more control and standardization. Until recently HTML was the only Web standard. However HTML only describes the structure, design and layout of Web pages. It is being replaced by Xtensible Markup Language (XML). XML describes the actual content of the information. What XML brings to the Web is 'powerful structured searching akin to the database field searching, but on a textual Web page' (Green 2000:132). In the future we will see search engine interfaces offering the option to search under keywords or tag searching (on the XML tags). Are we coming full circle to the level of precision offered by the commercial database hosts?

top

6 Conclusion

This article highlights the major differences between searching on commercial databases versus the Web. It has been shown that the on-line fee-based database hosts store structured information in highly organized databases. This results in high precision and recall and authentic information. The free Web on the other hand still consists of unstructured information, precision is low in comparison and information cannot always be authenticated.

The problem with searching the Web therefore is two-fold: firstly the data are unstructured and secondly searching results in information overload (lack of precision). The major search engines are addressing the unstructured Web by introducing standards such as meta data and XML. Search engines 'have borrowed techniques (best match searching, relevance ranking etc,) from information retrieval research while the subject directories nod towards classification theory' (Poulter 1997:142).

Finally the author poses the question regarding the free Web environment: *Are we moving towards a more structured Web*? The answer must be 'yes' although there is still a long way to go. At the rate that information technology responds to the changing needs of its industry, can we hope to reach this goal within half a decade? In the words of Green: (2000:134) 'Now emerging from its nascent stages, the Web may evolve into a highly organized, vastly diverse and complicated system'.

top

7 References

Bates, M.E. 1999. *Super searchers do business: the online secrets of top business researchers*. Medford, NJ: CyberAge Books.

Calof, J. 2002. *Searching Smarter: intelligence and the Net.* [Unpublished seminar notes presented at author's workshop at RAU on 26 April 2002].

Cooper, L. 1999: Linda Cooper: Independent info pro and end user. *In Super searchers do business: the online secrets of top business researchers*. Medford, NJ: CyberAge Books:31–51.

Dong, Z. and Su, L.T. 1997. Search engines on the World-Wide Web and information retrieval from the Internet: a review and evaluation. *Online and CD-ROM review* 21(2):67–81.

Green, D. 2000. The evolution of Web searching. *Online information review* 24(2):124–137.

Green, D. 2001. Infinitely extensible markup language? *Information world review* (December):34.

Muchin, J.A. 1999. Free thinking: using free sites on the Internet to save time and money. *The Bottom Line: Managing Library Finances* 12(4):150–152.

Poulter, A. 1997. The design of World-Wide Web search engines: a critical review. *Program* 31(2):131–145.

Price, G. and Sherman, C. 2001. Exploring the invisible Web: seven essential strategies.

Online 25(4):32–34.

Disclaimer

Articles published in SAJIM are the opinions of the authors and do not necessarily reflect the opinion of the Editor, Board, Publisher, Webmaster or the Rand Afrikaans University. The user hereby waives any claim he/she/they may have or acquire against the publisher, its suppliers, licensees and sub licensees and indemnifies all said persons from any claims, lawsuits, proceedings, costs, special, incidental, consequential or indirect damages, including damages for loss of profits, loss of business or downtime arising out of or relating to the user's use of the Website.

ISSN 1560-683X

Published by InterWord Communications for the Centre for Research in Web-based Applications, Rand Afrikaans University