DOCUMENT RESUME

ED 411 859 IR 056 686

AUTHOR Brunelle, Bette S.
TITLE Smart Systems, Smart Searches.
PUB DATE 1996-00-00
NOTE 6p.; In: Online Information 96. Proceedings of the International Online Information Meeting (20th, Olympia 2, London, England, United Kingdom, December 3-5, 1996); see IR 056 631.
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Access to Information; Computer Software; *Computer Software Evaluation; Evaluation Methods; Foreign Countries; Indexing; *Information Retrieval; *Information Seeking; Information Sources; Librarians; *Relevance (Information Retrieval); Statistical Analysis; *User Needs (Information); *World Wide Web
IDENTIFIERS Boolean Search Strategy; Cluster Based Retrieval; *Search Engines; United Kingdom

ABSTRACT

Almost overnight, the World Wide Web has made the solution of "classic" information retrieval problems a pressing commercial goal. The various information retrieval solutions being offered on the Web are quite familiar to the librarian. The various Web search sites make use of "traditional" inverted indexes, manual indexing, automatic indexing based on statistical models, relevance ranking, and document clustering. All of these statistical techniques have been used, usually along with a "traditional" Boolean Search engine, in various commercial information retrieval software products, and each has its strengths and drawbacks. This paper examines some of the strengths and weaknesses of the different search systems in terms of Web searching. The paper looks at the following search systems: Alta Vista, Yahoo!, InfoSeek, and Excite. All of the current Web search systems are hampered by the sheer size and diversity of the Web, which makes it difficult to add value to "documents" in the tradition of indexing and quality control, and also hampered by the Web's stateless nature which makes refining searches cumbersome and time-consuming. The search systems described have different strengths; depending on the search and the searcher, it might be better to opt for precision (Yahoo!); the large number of terms indexed (Alta Vista); the browsing serendipity of related documents (InfoSeek); or the all-around performance of clustering techniques (Excite). (Author/SWC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 411 859

Smart Systems, Smart Searches

By:

Bette S. Brunelle

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B.P. Jeapes

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

T2056986

BEST COPY AVAILABLE

Smart systems, smart searches

Bette S. Brunelle
Ovid Technologies, USA

Abstract: *Almost overnight, the World Wide Web has made the solution of 'classic' information retrieval problems a pressing commercial goal. Interestingly, in spite of a prevalent attitude summarised in a recent issue of Wired that 'information retrieval is really only a problem for people in library science — if some computer scientists were to put their heads together, they'd probably have it solved before lunch time,' the various I/R solutions being offered on the Web are quite familiar to the librarian. The various Web search sites make use of 'traditional' inverted indexes, manual indexing, automatic indexing based on statistical models, relevance ranking and document clustering. All of these statistical techniques have been used, usually along with a 'traditional' Boolean Search engine, in various commercial information retrieval software products, and each has its strengths and drawbacks. This paper will examine some of the strengths and weaknesses in terms of Web searching, and close with some observations on the current state of affairs and predictions for the future.*

Keywords: Information retrieval, statistical techniques, search techniques, advanced searching, relevance ranking, thesaurus, document clustering, Internet, indexes

1. Introduction

Some of the most-often used sites on the Web are those such as Yahoo!, InfoSeek, Excite and Alta Vista that provide search engines and indexing systems which ease the process of navigation on the Web. Interestingly, in spite of the optimistic attitude that 'information retrieval is really only a problem for people in library science — if some computer scientists were to put their heads together, they'd probably have it solved by lunch time,' (Ref 1) the various I/R retrieval solutions on the Web are simply implementations of various Boolean and statistical techniques that have already been developed by computer scientists, and much discussed at librarian conferences, for the past twenty years. As with any I/R indexing schemes or retrieval systems, each has its strengths and drawbacks — and none can claim to have 'solved' I/R.

A survey of a few of the Web search systems corresponds to a summary of the major search indexes and statistical techniques currently available. As will become clear, all of them are missing features which are common with CD-ROM or 'traditional' online systems. The systems do, however, break ground in attempting to let statistical processes and 'behind-the-scenes' manipulation of user input substitute for some of the set manipulation with which professional searches are accustomed to refining searches. Because the Web is a stateless system lacking the notion of a session, building on prior search sets is not possible. Thus the Web — by its nature a casual, end-user system — is the ideal testing ground for search techniques that by-pass complex Boolean logic.

2. Searching on the Web

A number of the Web search services rely on the most-familiar professional I/R retrieval system of all — the inverted index. 'Traditional' Boolean systems are based on the inverted index — all significant (non-stopwords) words in a document (in this case a Web page or part of a Web page) are posted in a list (the index) which includes the word and a pointer to the document which contains it. Some indexes also show which words are adjacent or near to each other in a document.

Unlike traditional inverted-index databases, which must store not only the index but also the actual text for each document, Web search indexes only have to store the index and a pointer to the document — the document is stored out on the Web somewhere and can easily be accessed separately from the search system. Because of this the storage requirements for Web databases are not as great as for the traditional ones but, given the size of the Web, the inverted indexes themselves can quickly become very large scale — at the current 20% growth rate, by 1998 a complete Web inverted index will exceed the almost 29 terabytes of information in the current US Library of Congress (Ref 1). What performance limitations there might be to indexes of this size is unknown.

Of course, at present none of the Web search systems claim to index the entire Web and it is difficult to know exactly what *is* indexed from service to service. Some index only Web pages; some include Listservs and Usenets. Some claim to index every word on a page and some index only the titles, keywords and/or the first paragraph. There is often no notion of fields corresponding to the document structure. Even where some field

structure is offered, the fields have to do with the structure of the Web (title, URL, host) rather than with the structure of a particular document. Distinctive data elements of a type of document, say 'book reviewer', are not available, nor are the kinds of delimiters often made from these data elements, for example 'book review.'

The Boolean sites typically offer AND/OR/NOT searching, although some have more extensive operators. Rules of entry vary from system to system and are often found only in help text. Many Boolean sites use techniques such as automatic truncation to increase retrieval — this can be both bewildering (when retrieval does not match input) and frustrating (when an already high noise level is increased).

Other Web search systems do not allow Boolean operations at all but rely on statistical processing or concept-based searching, using techniques which can provide fairly good results but in which it can be particularly difficult to predict (or change) results. This is both because the exact algorithm of a particular implementation of a statistical technique is proprietary and because the processes are generally based on document structure rather than on rules of user input. In the absence of Boolean input rules, delimiters or document structure elements it can be very difficult to fine-tune such a search.

Finally, in the end-user world of the Web there is no tradition of exhaustive documentation for sources and structure, nor would such a thing presumably be welcome. But it doesn't take much experience with 'traditional' search systems to know that crude, unrefined searches of a collection of undocumented provenance is going to provide uneven retrieval for many kinds of searches.

2.1. Alta Vista

The phenomenon of a large ratio of false drops for every 'hit' is apparent in a system like Alta Vista, particularly if used in Simple mode as, presumably, it would be by most end-users. In Alta Vista there are two modes of operation, Simple and Advanced. In Simple mode there are no Boolean operators but there are techniques to indicate phrase and proximity, as well as to indicate prohibited words, and the system is basically a Boolean system with a classic inverted index.

Results in Simple mode are relevance-ranked, which makes up somewhat for the diffuse retrieval. In relevance ranking, documents which are statistically most likely to be relevant appear at the top of the document list. 'Relevance' is a statistical technique and there are many variations on how to calculate it. Generally speaking, the words in a query are compared against a document and the more words from the query that appear in the document, the higher its relevance is judged to be. Variations on this scheme include weighing words in a particular part of a document (say the title) more heavily; weighing words that occur often within a collection (such as 'company') less heavily; or ranking documents in which the query words that appear together, or nearby, are more highly ranked than documents in which the query words are widely scattered.

Relevance ranking can mean many things. In some respects what is really relevant to the user can only be known by the user. Much of what he or she would consider determinants of relevance, for example currency and provenance (article source and author affiliation), could be implemented simply as sorts or filters — but this would require more complexity of either the user interface or of the system itself. However when a search is extremely focused, such as a search for a known person's home page, relevance ranking in the statistical mode works pretty well. This kind of search pulled up several references to the page and its surrounding pages with the actual home page URL in the top ten of ranked items.

For a very focused search such as the one just described, Alta Vista has a major advantage — it indexes far more of the Web than its competitors and is thus the system most likely to locate a known item. It also has several advantages which commend it to experienced searchers. It has, for a Web system, fairly extensive Boolean capabilities and syntax in its Advanced mode. These capabilities include some field qualification, truncation and wildcards. Thus an experienced searcher can use nested Boolean logic to perform searches not unlike those on traditional systems.

As may be inevitable when so many sites are indexed, there are many duplicates retrieved with Alta Vista. And the Advanced mode general subject search, although 'good' compared to some of the other Boolean Web systems, is only about as 'good' as any search in a very large general database with no subject headings and no subject focus.

2.2. Yahoo!

Another way to structure information, long known to librarians, is to impose order on a collection in the form of an indexing scheme with an intellectually coherent outlook. At Yahoo! a staff of indexers determine how to classify each Web site. Each site goes into a regional hierarchy and is cross-linked to an appropriate topic. Thus a United Kingdom library would be listed under the *United Kingdom* with a cross-link from *libraries*. The classification 'scheme' developed by Yahoo! is unique to Yahoo! and is designed to at some point 'have captured the breadth of human knowledge.' This is a concept well-known to librarians and the limitations are also well-known: any attempt at ontology includes bias and for optimal use of a classification scheme, the user has to be familiar enough with it to think like the indexer — exactly what the prototypical 'end-user' is considered unwilling to do.

On the other hand, the power of a good classification scheme, or controlled vocabulary, is widely understood as a major advantage over keyword searching. A search on 'uk libraries' in Alta Vista retrieves a variety of library-related sites, most of them having nothing to do with the UK — on Yahoo! you see a selection of index terms which include *UK:Public:Libraries Page*, *Reference:Libraries:Countries:United Kingdom* and *Index — UK Higher Education and Research Libraries*, and links from any of these descriptors give right-on documents. Generally speaking, a Yahoo! search tends to have more precision than a search on a Boolean system. However, if all you

are looking for is a known item such as a company name, going through a controlled vocabulary is not the most efficient route. Also, if your topic is very precise then the chances that it will exactly match an index term diminish, as does the precision of retrieval. Finally, manual indexing is an inefficient, time-consuming process which tends to get behind the collection — especially if the collection is the size of the Web and growing exponentially.

In spite of Yahoo!'s superior indexing, a recent review of Web search engines in *Online* dismissed Yahoo! from consideration because the search engine was 'not powerful enough to handle our queries effectively ' (Ref 2) — presumably in part because of an absence of fully-indexed documents with Boolean capabilities. And in fact, Yahoo! apparently realised these limitations because searches which do not get a 'hit' in the Yahoo! Index are automatically then run in Alta Vista. A search, for example, for a known home page brought up the Alta Vista results.

2.3. InfoSeek

InfoSeek is also an inverted index to Web Pages — however, once a simple index search is performed, documents are presented along with 'related topics' as well as a link from each retrieved document to 'Similar Pages.' A search for UK libraries on InfoSeek includes a rather odd array of related topics, from Law of the United Kingdom, to Novell NetWare, Maps and Veterinary Medicine. A 'Similar Pages' link from a document called *Web Resources in the UK* brings up a mixed bag of documents including *General Web Resources*, *Other World-wide resources* and *UK National Web and FTP Sites*. Generally speaking, one or two Similar Sites in InfoSeek may be of interest but the relevance is often marginal.

Although InfoSeek has no information on its own Web Page about it, it is reasonably certain that a statistical technique is being used to find related topics and similar pages. There are many variations on statistical analysis techniques but in a general way they work the same — a document or a collection of documents is scanned and statistics on the words in the document are used in various ways. For example, co-occurrence of words in a document or a document collection can be used to determine related words: if 'senator' and 'legislator' co-occur often then they might become related concepts. Of course, since computer analysis and not human intervention is being used, and since words or phrases can co-occur even when they are not related, then the results can be spotty.

There are many techniques to fine-tune automatic indexing and if statistical techniques are used, say, to summarise the content of the document then the statistical analysis may include a step which weighs terms differently, depending on where in the document they are located, or it will certainly weigh how frequently they occur in the collection against how frequently they occur in the document, for example:

> 'Keywords are typically chosen according to two standard measurements:
>
> (1) Words which appear frequently in a document are candidate keywords for that document.
>
> (2) The candidate keywords which occur infrequently in the rest of the collection are retained as keywords' (Ref 3).

The finding of Similar Pages could be based on something as simple as matching related concepts or be as fine-tuned as statistical clustering and categorisation — a technique in which documents are divided into subgroups according to their keywords and which is, in fact, the basis of the system surveyed below.

Generally speaking InfoSeek results in a subject search are interesting, if not exactly precise, and the system of related documents is quite good for browsing. At the time of this writing, InfoSeek is beta testing 'InfoSeek Ultra' which adds power with real-time updating, a number of semantic techniques (the system will find 'mice' if 'mouse' is input) and the ability to identify phrases, search exact case and so on. However, both of the InfoSeek systems failed to locate the known home page.

2.4. Excite

The Excite system uses the Architext search engine which indexes pages on the Web by concept. Although the details of implementation are proprietary, this is an example of the creation of a classification scheme through document clustering. Because document clustering techniques work from the actual collection, the subject categories are said to reflect 'the real world': if two words suddenly become related that were not related before, such as 'oj' and 'trial', then the subject categories will begin to reflect that immediately rather than having to wait for a human indexer to assign a new indexing term. And because clustering works on the entire collection, and because the collection in this case is large, then some ambiguities of meaning ('oj' as a celebrity or 'oj' as a citrus drink?) tend to be resolved by the clustering technique itself — if the user enters 'oj verdict' as a search then the documents that tend to cluster with 'oj and verdict' are likely all to concern concepts like 'trial, murder, sensational, century, lawyers' and so on, and not as likely to concern 'citrus, Florida, growers.'

For best results with a concept-based search, the ideal search will contain several concepts or a single unambiguous concept — a search on 'oj' alone doesn't gain any disambiguation by clustering. (However, if you want something on 'orange juice' on the Web it's best not to use 'oj' at all. The overwhelming results from 'oj' on any of the systems is OJ Simpson-related — for practical purposes, in the real world, 'oj' *is* now an unambiguous concept.)

Like InfoSeek, Excite offers a 'find similar' capability called 'query-by-example.' If an Excite hit is particularly good then the system will bring up similar items, relevance ranked (with the original good hit at the top). And if

all these tools aren't enough, Excite provides the user with the option to search by keyword or to use the statistically-created concepts. And all searches start as Boolean searches, using a 'fuzzy AND', which means that both AND and OR are used — with a higher weight being given to AND.

On Excite, the 'uk libraries' search retrieved, at the top of the list, items that were highly relevant — not as many relevant items as Yahoo! but more items, and more consistently relevant than the other systems. However a concept-based model, particularly one in which the concepts are generated strictly statistically, is still prone to errors and is stronger on recall than precision (although the known home page search did retrieve the home page and it was at the very top of the list!). It also can be difficult to determine why a particular item was retrieved and thus difficult to adjust one's search to correct problems.

Excite has a very nice display which includes a full abstract for most items.

4. Conclusion

The systems surveyed were good examples of the current state of Web searching. The InfoSeek Find Similar and Excite! concept search capability were not as precise as the human indexing of Yahoo! nor as imprecise as the 'simple mode' Boolean searching of an inverted index such as Alta Vista — but Alta Vista's 'advanced' Boolean did quite well also.

However, all of the current Web search systems are hampered by the sheer size and diversity of the Web, which makes it difficult to add value to 'documents' in the tradition of indexing and quality control, and also hampered by the Web's stateless nature which makes refining searches cumbersome and time-consuming. Depending on the search and the searcher, it might be better to opt for precision (Yahoo!); the large number of items indexed (Alta Vista); the browsing serendipity of related documents (InfoSeek); or the all-around performance of clustering techniques (Excite).

As end-users become more sophisticated about their needs, it seems likely that some Web search systems will begin to differentiate by topic areas, actively taking on more of the role of database producer, somewhat as Yahoo! is already doing with its indexing. It may also be that these very large systems are able to overcome the limitations of statelessness, much as some of the smaller, proprietary systems have already done, which would allow for delimiters and more ways to add value to the search process. I would predict that over time, Web search systems will come to resemble existing proprietary systems, and vice versa. The ideal searching environment, not easily realised either in twenty years of online or a few of the Web, combines a simple interface with sophisticated behind-the-scenes techniques. The Web, for the present, mandates a simple interface and there is a lot of experimentation with techniques.

Once the ease and power of statistical techniques are combined with the ability to refine user input, and reliance on the particular structure of documents or databases, then searching will get smarter — the goal is a system which works as well as an experienced librarian whose main business has always been to solve 'the information retrieval problem' and who has done it quite well overall.

Bette S. Brunelle
Ovid Technologies
333 7th Avenue
New York
NY 10001
USA
Tel: +1 (800) 950 2035, x 246
Fax: +1 (212) 563 3784 (fax)
E-mail: betteb@ovid.com

References

[1] Steinberg, S.G. (1996) Seek and ye shall find (maybe), *Wired*, **4**(05), 108–114.

[2] Zorn, P., M. Emanoil and L. Marshall (1996) Advanced Web searching, tricks of the trade, *Online*, **20**(3), 15–28.

[3] Munson, J. and B. Thornburg (1996) Taking advantage of advanced searching, *Proceedings of the Seventeenth National Online Meeting*, New York, Information Today, Inc., pp. 263–268.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

ERIC Full Text Provided by ERIC
(3/92)