Msc in information management

Business, Management

This essay is aimed to analyse why most World Wide Web search engines provide best match searching as their principle retrieval method with Boolean searching playing an auxiliary role. The World Wide Web has revolutionised the way in which people access information, and search engines are are widely used by people to find useful information on the Web. There are pros and cons lying in both best match retrieval system and Boolean retrieval system. Comparisons of these two retrieval methods show that the performance of best match searching is generally stronger than Boolean searching in an onlineenvironmentfor general uses.

Nevertheless, as long as in some circumstances Boolean searching provides more effective and accurate performance, a replacement is not seen. Main Contents The Importance of Searching The World Wide Web Seeking information is an activity fundamental to all human beings. Throughout history, one of man's primary concerns has been to satisfy his information needs. With the development of new technologies, people's behaviour with regard to accessing information has been greatly changed. Since the Internet was invented in 1969, it has been growing rapidly and now has extensions in every corner of the globe (Poutler (1997)).

The World Wide Web (WWW) has revolutionised the way in which people access information, and has opened up new possibilities in areas such as digital libraries, and the dissemination and retrieval of scientific information. The Internet is proving to have important implications in areas as diverse aseducation, commerce, entertainment, and medicine andhealthcare. With the help of web search engines, people can efficiently search the large

amount of information available. The volume of information on search engines has been exploding in the past years.

As a huge amount of information has long been available in libraries, the revolution that World Wide Web has brought is rather an improvement in the efficiency of accessing information. The results of GVU's April 1998 WWW user survey indicate that about 86% of people now find a useful Web site through search engines, and 85% find them through hyperlinks in other Web pages; people now use search engines as much as surfing the Web to find information (Kobayashi, M. ; Takeda, K. (2000)). This indicates the importance of the role that web search engines are playing for people accessing the information on the Internet.

What is a World Wide Web Search Engine? Poutler (1997) defines a 'World Wide Web search engine' as a retrieval service consisting of a database (or databases) mainly describing resources available on the World Wide Web (WWW), search software and a user interface also available via the WWW. " Archie", the earliest Internet search engine appeared at McGill University. It allowed keyword searches of database of name of files available via FTP (File Transfer Protocol). A global network of Archie servers was set up and each server offered local access to a " mirror' of the original Archie database.

Only substring filenames could be searched. However a single Archie search could turn up to references to a file stored on many different sites, then the searcher could retrieve the nearest copy of this file by FTP. (Poutler (1997)) Gopher, which was created at University of Minnesota in 1992, later allowed the creation of menus and links from items in these menus to either files or

to menus on other Gopher servers. This was the forerunner of World Wide Web. A search engine called Veronica was developed for searching on Gopher.

It supported standard Boolean operators and it would default a multiword search without operators to Boolean AND. (Poutler (1997)) Today, web search engines use software robots to survey the Web and build their databases. Web documents are retrieved and indexed. When a user enters a query at a search engine website, the input is checked against the search engine's keyword indices. The best matches are then returned to the user as " hits". Most web search engines offer two different types of searches-" basic" and " advanced" searches.

A " basic" search normally uses a st matcbeh retrieval method. This is the principle retrieval method of most Web search engines. A " basic" search requires the user simply to enter a query without sifting through any pull-down menus of additional options. The output of the retrieved records relevant to the query will be ranked in a relevancy order, which is also the order that is likely to be browsed. Depending on the particular search engine, the " basic" search may sometimes be fairly complex. An " advanced" search normally uses the Boolean retrieval method.

Most web search engines offer this as an auxiliary retrieval method with which a user can refine the search. As there are many different Web search engines, options may vary from one to another. However, generally, a user is required to construct a query by using Boolean logic operators in order to specify the query term. Web Search Engines and Information Retrieval

Models According to Turtle and Croft (1991), There are three main information retrieval models. They are: the exact match model, the vector space model, and the probabilistic model.

Boolean searching belongs to the exact match model. (It is named after the British mathematician George Boole). (Cohen (2001)) Boolean logic refers to the logical relationship among search terms. There are three logical operators in Boolean logic: AND, OR, and NOT. Boolean AND means that all the terms that user specifies must be contained in the retrieval of records. OR means that at least one of the terms that user specifies must be contained in the retrieval of records. NOT means that the terms that user specifies 'not' must not be contained in the retrieval of records.

A Boolean search system accepts Boolean queries as input and uses the normal interpretation of first-order logic augmented with proximity operators to select a set of documents that satisfy the query. Normally a user utilises Boolean searching to refine the search and to specify what form he would like his results to appear in, and whether he wishes to restrict the search to certain fields. However it requires sophisticated constructing of query forms. Best match searching, on the other hand, covers the vector search and probabilistic models.

As Prof. Willett prompted in his lecture, several other terms are also widely used to refer to what I shall refer to as best match searching. These include vector processing, probabilistic retrieval, and ranked-output searching. Best match is considered as a convenient and simple searching method. In a best match retrieval system a user inputs a search query to a certain information

retrieval system, and the system ranks the documents in the database in order of decreasing similarity with the query and then displays the results to the user.

Simple entry and ranked output are the main features of best match searching. Natural language query processing is regarded as a constituent of best match retrieval. Comparison of Boolean and Best Match Information Retrieval Systems In order to evaluate the performance of different information retrieval systems, Koenemann and Belkin (1996) indicate that two standard measures of retrieval effectiveness are precision (the number of relevant retrieved documents over the total number of retrieved documents), and recall (the ratio of relevant retrieved documents to the total number of relevant documents).

Meanwhile Kobayashi and Takeda (2000) say that most of the measures that have been proposed to quantitatively measure the performance of classical information retrieval systems can be straightforwardly extended to evaluate Web search engines. Thus a three-way trade-off between the speed of information retrieval, precision, and recall is recognised. Koenemann and Belkin (1996) say that users in all types of information retrieval systems face the central difficulty of effective, interactive formulation (and reformulation) of the queries that represent their information problems.

Web search engines are no exception. Belkin and Croft (1987) state that from experimental studies it has been known for some time that in terms of recall and precision performance measures, best-match, ranked output retrieval techniques are in general superior in non-interactive settings to

exact-match systems, such as commercial Boolean information retrieval systems. The Boolean retrieval system has been used for several decades, but it has always been subject to certain criticisms. Cooper (1983) says that Boolean systems have some serious drawbacks.

The first drawback is that the Boolean language is confusing to the novice. This may be true. Since a Boolean search system only accepts Boolean queries as input, a user is required to construct the query formulations by Boolean logic operators (AND, OR, and NOT). Constructing a query formulation in Boolean form poses evident difficulties for those inexperienced users or those who have not been trained to operate Boolean logic. Salton et al. (1983) write, In operational information retrieval, Boolean query formulations are used to express the customers' information needs.

The standard rules of Boolean logic may not, however, provide an ideal environment for the formulation of effective search requests... Unfortunately there exists much evidence to show that ordinary users are unable to master the complications of Boolean logic, and even professional indexers and searchers find it difficult to construct consistently effective index representations and search statement. Similarly, Willett (1988) argues that although the great majority of current retrieval systems are based upon Boolean searching, there are severe problems associated with the use of such a retrieval model.

He indicates that the first major disadvantage is the difficulties associated with the formulation of the query using the Boolean operators AND, OR and NOT, since end-users are normally unable to formulate good queries and

require the assistance of trained intermediaries. The aforementioned criticism on the difficulties of constructing queries for Boolean retrieval system explains one of the reasons why most of the Web search engines do not use Boolean searching as a principle retrieval method.

The early Web search engines were volunteer-run and co-operative efforts, so this user-oriented problem might not be an outstanding conflict. Now, however, most Web search engines are commercial products and host advertisements in order to pay the costs of running and maximise profit. Trying to get the highest hits and acting the role of the most popular and successful web-searching site becomes the primary goal of all Web search engines. " Welcome! As the world's first Internet Butler, I'm always at your service.

I've made it my mission to humanize the online experience by making it easy to find the most relevant information, products and services. " This is the greeting on the Web search site called 'Ask Jeeves'. Regardless of whether this particular web site is successful or not, making efforts to be " humanise" is one way of ensuring success among strong competition from other Web search engines. This is one of the reasons why most of the Web search engines would not put a sophisticated, difficult to handle Boolean searching (or advanced searching) on their front pages, because Boolean searching is not 'natural' for the majority of users.

As a constituent of the best match retrieval method, natural language query processing was introduced for the first time in a commercial online environment in 1992. West Publishing Company launched their WIN

(Westlaw Is Natural) in that year. (Pritchard-Schoch (1993)) Westlaw is a large commercial system that provides access to U. S. legal materials. Practically, WIN provides the facilities to use plain English to access the massive case law collection, and the output of the records would be ranked by the likelihood that they will be judged to match the information need.

Later, Turtle (1994) evaluated and compared the performance between natural language and Boolean query based on the WIN product. Four different measures were used in Turtle's test: precision at standard recall points, raw precision and recall at a fixed rank cutoff, the relative performance of individual queries in the test set, and precision at all ranks up to some maximum value. Turtle reached the conclusion that natural language searching would present users with more relevant documents and would present these documents to the user in rank order of those most likely to be browsed.

On all of the evaluation metrics discussed, natural language searches consistently produced better rankings. West has tested these techniques extensively on other material types (e. g. statutes, law review articles, administrative codes). Natural language searching performs consistently better than Boolean for all of the material types tested. Turtle typically mentioned ranking in his test and this is another criticism on Boolean search. Bookstein (1985) comments,

A user may make a request a form A and B, but cannot indicate that term A is more important for his search or that it is, say, twice as important as term B. The corresponding output limitation is that documents are simply

retrieved or not retrieved, where as a ranking of documents according to the degree of relevance to the inquiry is believed by many to be desirable. Along with the increase in the volume of documents on the Web, the problems posed by lack of ranking become increasingly severe.

With a simple Boolean query input, a Boolean system could easily return an unmanageable number of results. Willet (1988) also points out that one limitation of Boolean searching is that the retrieval operations result in a simple partition of the database into two discrete subsets: those records which satisfy the query and those which do not. Thus all of the retrieved records are presumed to be of equal usefulness to the searcher, and there is no mechanism by which the may be ranked in order of decreasing probability of relevance.

Despite the aforementioned shortcomings, Boolean searching is not going to disappear. After his comparison of natural language versus Boolean query (1997) Turtle concluded that natural language searching has strong performance, but Boolean query languages will not disappear anytime soon. For some queries and some kinds of materials Boolean techniques give better results. Furthermore, some users prefer Boolean queries in some circumstances. Commercial systems will need to support both query types to be successful.

Westlaw's WIN still has the ability for users to easily switch between natural language and Boolean query. The DIALOG searching is still based on the Boolean logic approach. In general, the best match retrieval has shown better performance levels for general searching of full-text files as well as

being more user-friendly to an untrained user. Nevertheless, the exact match models continue to be useful for know-item searching as well as bibliographic and field searching.

Best match searching has not yet reached the level at which it could take the place of the Boolean system altogether. For example, in a case where two different worlds share the same spelling, a well-formed Boolean query could easily eliminate the unwanted matches while a best match retrieval may well return with completely irrelevant results. Boolean searching as an exact match model is suitable for user who wants to carry out a specific search with the option of formulating sophisticated queries, especially when the relevancy ranking is not such an important issue.

Regarding Westlaw's WIN, Jeffery Riffer considers it an enhancement of his ability to retrieve relevant document, but does not see it as a replacement for Boolean logic searching. Meanwhile Ed Marrod, a seasoned searcher says more explicitly " You can tell, within a few searches what it's doing and if I wanted to do the search that way, I'd do it in Boolean. " (Pritchard-Schoch (1993)) Conclusion Market forces heavily predominate the design and evolution of the major commercial search engines. It is unsurprising that about 85% of Internet users surveyed claim to use search engines and search services to find specific information.

Commercial considerations play an important role in determining the popular employment of best match searching as the default search mode. Since the majority of users of a large, general-use search engine will be untrained, casual users, it makes sense that the default setting offers the type of search

that best matches the needs of these users. Here, best match searches display clear advantages over Boolean systems - most importantly the simplicity and user-friendliness of the user interface (especially when " natural language" searches are offered) and the ranking of results according to relevance.

However, as long as there are remain circumstances in which a Boolean search is more effective and a customer base that is prepared to use it, it makes sense for large search engines to make this option available as an auxiliary search method. Furthermore, the same survey which showed the impressive 85% preference for search engines as the method of data retrieval, also indicated that users are not satisfied with the performance of the current generation of search engines; the slow retrieval speed, communicationdelays, and poor quality of retrieved results (e.

g. noise and broken links) are commonly cited problems. In such a setting, the current trend towards increasing computer literacy as well as increasing reliance on the internet for many types of serious research (as opposed to casual domestic " web-surfing") might even mean an increase in the number of people who are willing to learn how to use Boolean searching in order to make such research more effective.

Bibliography

Belkin, N. J. ; Croft, W. B. (1987). " Retrieval techniques". In: Williams, M. E. (Ed.), ARIST, Ch. 4, pp. 109-145.

Bookstein, A. (1985). " Probability and fuzzy set applications to information retrieval". ARTIS, 20.

Cohen, L. (2001). " Boolean Searching on the Internet". Online[Online] http://library. albany. edu/internet/boolean. html [Accessed 17 December 2001].

Cooper, W. S. (1983). " Exploiting the maximum entropy principle to increase retrieval effectiveness". Journal of the American Society for InformationScience, 34(1).

Kobayashi, M. ; Takeda K. (2000). " Information Retrieval On the Web". ACM Computing Surveys, 32, 144-173.

Koenemann, J. ; Belkin, N. J. (1996). " A case for interaction: A study of interactive information retrieval behaviour and effectiveness" Common ground: CHI96 Conference Proceedings, 1996, 205-212.

Poutler, Alan (1997). " Design of World Wide Web search engines: a critical review". Program, 31(2), 131-145.

Pritchard-Schoch, T. (1993) " Natural Language Comes of Age". Online, 17(3), May 1993, p. 33-43.

Salton, G., Buckley, C., ; Fox, E. A. (1983). " Automatic query formulation in information retrieal". Journal of the American Society for Information Science, 34(4).

Turtle, H. R. ; Croft, W. B. (1991). " A comparison of Text retrieval Models". The Computer Journal, 35 (3), 279.

Turtle, H. (1994). " Natural language versus Boolean query evaluation: a comparison of retrieval performance." In: Croft, W. B. ; van Rijsbergen, C. J. (eds) Proceedings of the Seventeenth International Conference on Research and Development in Information Retrieval. Pp. 212-220. London: Springer Verlag.

Willett, P. (ed.). (1988). Document Retrieval System. London: Taylor Graham.