DIRECT SEARCHING OF A CD ROM DATABASE BY LIBRARY USERS : A STUDY

David F C Chan

Abstract

Library users searching for information from CD ROM databases normally do so without any help from search intermediaries who have the necessary knowledge and experience required for the job. Existing user interfaces to CD ROM databases provide little to help the library users to carry out their searches. Informal retrieval requires cognitive skills and these should be identified before an effective user interface could be designed. This paper presents the findings of a study to discover the problems and difficulties faced by library users searching for information from a CD ROM database. The cognitive skills and knowledge required for successful information searching, and the type of help they require during the process are identified and discussed. The findings will facilitate the designing of an intelligent user interface that will help library users to retrieve information from CD ROM databases on their own.

INTRODUCTION

In the past, searching of online databases were normally conducted by search intermediaries, the reference librarians. With the availability of databases on CD ROM and the supposedly user-friendly interfaces, searchers are now expected to be able to conduct searches without any help from the intermediaries.

A study was conducted to discover the problems and difficulties faced by library users searching a typical CD ROM database without the help of a search intermediary. The study also identified the skills and knowledge required for successful searching by the library users, and the type of help they need when searching for information. The findings will facilitate the designing of an intelligent user interface which can help the searchers achieve their goals, that is getting as many relevant citations as possible from the CD ROM database. The study was not concerned with the mechanics of system use and syntactic details which are very system dependent. Most existing user interfaces do more to ease the mechanical rather than the conceptual problems of use (*Borgman*, 1986). The search process as well as the capabilities required at the various stages should be understood before any meaningful intelligent help

facilities could be incorporated into the user interface.

METHODOLOGY

The think aloud protocol technique was used to elicit detailed information on how end users carry out their searches of a bibliographic database. Users of the Computer Select (CS) Database on CD ROM were asked to provide running commentaries of their actions and thoughts while they were retrieving information. The sessions were tape recorded and the recordings were analyzed later.

Recordings of the think aloud protocols were also used during subsequent interviews with the users when references to particular actions or thoughts were made. The expert searcher, the reference librarian, could also be asked to listen to the recordings and then comment on the performances of the end users. The correct search tactics could therefore be elicited from the expert.

In this study, the thinking aloud sessions of twenty students and academic staff accessing the Computer Select (CS) database were recorded. A short interview was carried out immediately after each session.

The CS database comes with the Lotus Bluefish Searchware (product of Lotus Development Corporation & Ziff Communications Company) on CD ROM. CS contains the latest articles from 40 computer journals and abstracts from 110 other publications; descriptions and specifications of over 68,000 hardware and software products; profiles of over 11,000 companies; and a glossary of computer terms (Lotus, 1991). Two types of searches are possible: word search and field search. For free text searching, word search is used. The user enters a search query consisting of words and phrases, and logic and proximity operators. These words and phrases can be located in any part of the text including the title. Wildcard (in words and in phrases) searches are possible. In a field search the system looks for documents that meet the specific value indicated in one or more fields. The value of each field is normally selected from a wordwheel. There is a topic field which provides the user with a list of controlled topics to choose from.

The initial search is made on the entire selected domain. Subsequently, the output of a particular search can be selected and used as a restricted (narrower) domain. Word searches can be combined but not a word search and a field search.

TASKS ANALYSIS

Searching of the database involves a numbe of tasks. The study shows that special skill and knowledge are required for each task End users who do not have sufficient know ledge of the subject of interest and who d not have the necessary skills for informatio searching will have the most difficulties i getting what they want. The difficulties c the searchers in performing the tasks ar discussed below:

Devise and use a search strategy

A search strategy is a plan for the who search. A user with a well thought-out pl will indicate that he knows exactly what h needs and how to get what he wants. The us who has a search strategy is normally o who has knowledge of and experience information retrieval, has some knowled of the subject area, and has probably access the database previously. A user who lac knowledge or experience in any one of the areas does not know what to expect and is r

likely to have a plan. When searching is done through a search intermediary, some form of planning is normally imposed. A search request form has to be completed by the user, with the assistance of the search intermediary. This forces the user to think very carefully about his request. For example, the user is required to supply the following information to the search intermediary:

a. Purpose of search
b. Detailed statement of request
c. Important terms (broad and narrow), synonyms and keywords
d. Known relevant papers and authors
e. Specific year, sources and other limitations

It was observed that users doing their own searching of the CD ROM database had no search plan at all. There was no requirement (imposed by a search intermediary) for them to come prepared even just with a partially completed search request form. They were expected to do the searching themselves and get what they want by trial and error. No charges were involved and they were allowed to have free access to the CD ROM and use it in the reference library for as long as they wish.

End users want access to an information retrieval system because they recognise that there is an anomaly (or gap) in their know ledge of a particular subject. It is therefore not possible for them to specify precisely what they do not yet know or do not know enough. (*Belkin,* 1982; *Hildreth,* 1989).

Express his requirements in the form of a Boolean search statement which is combination of terms and logic operators

Users using the CD ROM database are not only expected to know what they want. They have to translate their requirements into a search statement acceptable to the IR system. A search statement/query consists of a combination of search terms and Boolean operators. Search terms are keywords used to index the documents; the user therefore has to use search terms from a controlled vocabulary. Boolean operators allow the user to make his searches more specific but they have to be used correctly to reflect the user's actual requirements.

In a free-text search, the user is free to use words most likely to appear in the titles or texts of the documents. He need not restrict his choice of words (or search terms) to those in the index (or controlled vocabulary). Proximity operators can also be used in the search statement to narrow the search. For example, the user can indicate that the two words or phrases should appear in the same sentence or within the same paragraph.

Free-text searching has the advantage of locating terms not yet in the indexer's vocabulary but the approach is less precise unless the search is properly formulated. In fact, in order to have more precision in his search, the searcher is expected to make greater use of Boolean and proximity operators, and also terms which are not too common or ambiguous.

We observed in our study that although some users were able describe their requirements, they seemed to have difficulties expressing their requirements in the form of Boolean search statements. It is quite apparent that some users could not translate their requirements into Boolean search statements which consists of only terms and Boolean operators. In fact it looked like they tried to avoid the use of operators.

All the users started their searches cautiously with either one term or a very simple search statement with two terms and an AND opera-

tor. They did not use more than one AND operator in any of the subsequent searches. The OR and NOT operators were hardly used. In all cases, the end users understood the functions of Boolean operators in the search statement but had no confidence (and was therefore reluctant) to use them.

In our study, most users using the CS Database were researching into areas that were new to them. They entered whatever words or phrases that came to their heads. They have limited knowledge of the subject area and therefore could not think of enough terms or topics to help them in their search.

In a word search, unless the user supplies the right combination of terms and operators, the retrieved documents will be largely irrelevant and useless. The retrieval system therefore assumes that the user can use the right terminologies because no help is provided. In a field search using topics, the user must know exactly the topics he wants. He may not be able to spell the topics correctly but he must first come up with the topics. The wordwheel which displays the topics in an alphabetical order will only help the searcher to get the the correctly spelt term and perhaps help him to identify terms with the same stems. It would not be helpful to the user who does not know which topics to use. If the term is not in the controlled vocabulary, the user is expected to find an alternative term (synonyms, abbreviations, acronyms, etc.). An online thesaurus, meant for searchers and not indexers (*Bates*, 1986), would be more useful in helping searchers with his choice of terms but searchers will have to be taught how to use it.

A searcher who wanted documents on computer supported cooperative work (CSCW) could not get what he wanted because he did not know the right terms to use in his search. He was only trying his luck using various other unrelated terms. Another searcher looking for documents on information strategy planning in legal firms also had the same problem. She was unable to get even one citation. Notes on the two cases are appended.

Employ search tactics to achieve his goal

In the past, when bibliographic databases were only accessible to search intermediaries, there was no direct searching of the databases by the end users. Tactics were employed by the intermediaries during the process of searching. Their goal was to obtain for the user as many relevant documents as possible and speedily.

Some search tactics used by experienced searchers in on-line searching are:

a. Keeping track of the search paths one has followed and of desirable paths not followed up or not completed.

b. Designing an indirect route through the files to reach the desired information.

c. Putting all elements of the query in the initial search formulation or adding one or more elements to an already-prepared search formulation.

d. Minimizing the number of elements of the query in the initial search formulation or to subtract one or more query elements from an already-prepared search.

e. Making the search formulation broad (or broader) by including synonyms or conceptually parallel terms.

f. Making the search formulation precise by minimizing (or reducing) the number

of parallel terms but retaining the more descriptive terms.

g. Rejecting, in the search formulation, items containing or indexed by certain undesirable term(s).

h. Moving upward hierarchically to a broader term.

i. Moving downward hierarchically to a more specific term.

j. Moving sideways hierarchically to a coordinate term.

k. Seeking additional search terms by looking at neighbouring terms, whether proximate alphabetically, by subject similiarity, or otherwise.

1. Examining information already found in a search in order to find additional terms to be used in furthering a search.

m. Trying alternate affixes, whether prefixes, suffixes or infixes (morphological variations).

n. Reversing or rearranging the words in search terms in any order or all reason able orders.

o. Searching for terms logically opposite from that describing the desired information.

p. Searching under a different spelling. (*Bates*, 1979; *Smith*, 1989).

The tactics used by a search intermediary are a reflection of his skill and experience in information retrieval. The use of search tactics can also be regarded as a form of creative problem solving (*Katz*, 1987) which requires cognitive skills (*Chen*, 1991). However, the search intermediary cannot be knowledgeable in too many fields. He also depends on search aids like the thesaurus, classification indexes and even dictionaries. The thesaurus which provides lead-in terms ('see' references, acronyms, abbreviations, etc.), broader terms, narrower terms, related terms and synonyms (or preferred terms), is indispensable to searchers of bibliographic databases.

Most library users are poor searchers because they have never been trained in information retrieval which was, in the past, a job of the search intermediary. We found that some users could not retrieve anything relevant while others could only retrieve a small part of what the database has to offer. They just do not have the necessary skills and experience in searching databases.

The end users observed in the study seemed to know that a search can be expanded or narrowed depending on the outcome of their earlier searches but only very simple tactics were employed. The user who entered both the acronym ISDN and what it stands for in full, could not tell why the number of citations retrieved were less than his earlier search which used only the acronym. However, he knew that the use of a more specific term will narrow his search. Another user found the topics listed in the retrieved documents to be useful in a subsequent search. However, he did not make a note of them.

An end-user thesaurus, mentioned earlier in this report, will be useful if the searcher has to use some of the tactics discussed in this section. However, even if the thesaurus in its traditional format is available on-line, it is doubtful whether the end-user will actually refer to it to help him formulate and to modify his search. The linear, alphabetical arrangement and one-dimensional concept relation-

ship presentation require training to use and may not be effective in the cognitive sense. The end user who does not have sufficient knowledge of subject area will still find it difficult to explore the thesaurus.

Determine whether the items retrieved are adequate and relevant

Whether the search is conducted by the search intermediary or by the library user himself, the searcher is the one who ultimately decides whether the retrieved citations are sufficient and relevant. The turnaround time for a search done by the search intermediary is high. The end-user will have to wait (unless the search can be done immediately for him) for a day or two before he gets a list of citations and/or abstracts. He then has to read the citations to decide whether he is getting what he wants. The availability of abstracts can help him in this evaluation task since titles in most cases do not tell much. If most of the citations retrieved are irrelevant, too general or the number of citations are inadequate, he has to approach the search intermediary for a second search to be done.

In a CD ROM information retrieval environment, the responses to the searcher's queries are immediate. The titles, abstracts or even full texts of documents displayed. The searcher can decide there and then whether his search has been successful, i.e., whether he has obtained enough of relevant materials. If not, the system normally provides the facility for him to copy the materials onto a diskette for him to make his evaluation later and maybe at a different place.

It has been observed that the library user read the list of titles first to have a feel of the type of documents retrieved. In most cases, the titles did not convey much and he had to browse the texts displayed on the screen. He would read the summary if there was one. If he had to read the full text, he usually looked out for the occurrences of the specified terms (highlighted) in the texts. This was to ensure that the terms were used in the right context. In one case, the library user noted that the topics on the descriptors list were relevant and then read the text to confirm that the document was relevant to him. Besides indicators of subject content, other indicators that could have been used to determine whether a document is relevant are the source, the author's name, date of publication and bibliography or cited papers.

It was also observed that the searcher would not normally attempt to read all the documents to conclude whether he had been successful. Because he had very little time, he would at most read only the first few documents to reach a conclusion. Computer Select does not present documents in any order of relevance because there is no way to measure the degree of relevance of each retrieved document. In a Boolean retrieval system, as long as a document meets the search criteria, it is considered relevant. There is no need for the searcher to indicate, when formulating or modifying his search, which term is more important than the others. Therefore, all retrieved documents are assumed to be equally relevant. Also, all rejected items are considered equally irrelevant. Probabilistic retrieval methods using weighted terms and output ranking are used in some information retrieval systems to tackle this problem.

In a free text search, the presence of the specified terms in two documents does not mean that they have the same subject content. The terms in one document may be used in a context which is different from what the searcher has in mind. The searcher has to ensure that the terms are used in the context that he is interested in by reading at least the sentences or the paragraphs where the terms occur.

Decide whether to modify his search

Whether the searcher wants to modify his search depends on the results of his previous search(es). The output of his previous search could be inadequate, a large part of it could be irrelevant and/or useless or not specific enough. The user will have to reformulate his search statement using better terms and operators or even to change his search domain. The tactics as discussed above could be used for search modification.

CONCLUSION

Designers of user interfaces for CD ROM bibliographic database systems have certainly simplified the syntax and mechanics of system use. However, the searcher who has no training whatsoever in information retrieval and very little subject knowledge will have difficulties in performing the various required tasks in the information retrieval process.

Active information retrieval research and development work are being carried out in the following broad areas to tackle these problems:

a. Enriching the entry or 'lead-in' vocabulary (*Congreve,* 1986).

b. Natural language processing: interpretation and translation (*Vickery,* 1987; *Porter,* 1980; *Doszkocs,* 1983; *Mitev,* 1985; *Smith,* 1989).

c. Improving browsing and navigation facilities for exploration (*Wade,* 1988; *Duncan,* 1987; *Croft,* 1987).

d. Search strategy selection and retrieval techniques ("post-Boolean") (*Robertson,* 1976; *Porter,* 1988).

e. User modelling (*Brajnik,* 1987; *Brooks,* 1985) and knowledge requirement & representation (*Vickery,* 1986; *Smith,* 1989; *Brooks,* 1985).

Hildreth (1989) is of the opinion that the following techniques pioneered in information retrieval and artificial intelligence should be incorporated into future information retrieval systems: natural language processing; direct or indirect mapping/linking of free text terms to terms in the controlled vocabulary; flexible, heuristic retrieval strategies; probabilistic retrieval with weighted-term, combinatorial searching and the ranking of output; and relevance feedback procedures for automatic query expansion and modified search strategies.

Future information retrieval systems should satisfy four requirements: easier, system assisted search entry and query formulation; automatic assistance with failed searches through query reformulation or modified search strategy; improved extended-Boolean retrieval methods; and graphical techniques to present terms and their relationships into the context of the actual search process (*Hildreth,* 1989).

The aim of an information retrieval system is to retrieve documents that are likely to be relevant to the user. According to Cooper, a document may be relevant to a query but may prove to be useless for various reasons. Retrieval success cannot be measured using the traditional recall-precision measurements but only by determining the utility to the users of the retrieved documents (*Cooper,* 1976). Also, information retrieval should be an interactive and cooperative process, a process that fully engages the user in order to meet a common goal. The system should assist the user in performing the various tasks effectively and efficiently. The more a system knows about the user, the better it can offer help.

Meaningful engagement of the searcher will be required at various points (query formulation, relevance assessment, query modification, etc.) in the information retrieval process. The system should assist the user in his tasks. Assistance should be provided on a syntax (form of the query) as well as a semantic (contents of query and the underlying concepts) level (*Alberico,* 1990). The help facilities offered to assist the searcher are primarily to help the user to retrieve relevant information speedily. The goal of educating or training the user is secondary. The user modelling techniques used in intelligent tutoring systems are perhaps not quite applicable in this context. The searcher is not there to learn as his intention is to obtain very quickly as much relevant information as possible.

REFERENCES

Alberico, R and Micco, M. *Expert Systems for Reference and Information Retrieval.* London:Meckler,1990. pp. 234.

Bates, MJ. Information search tactics. *Journal of the American Society for Information Science,* 30(4):205-14, 1979.

Bates, MJ. Subject access in online catalogs: a design model. *Journal of the American Society for Information Science,* 37(6):357-76, 1986.

Belkin, NJ et al. ASK for information retrieval: Part I — Background and theory. *Journal of Documentation* 38(2):61-71, 1982.

Borgman, CL. Why are online catalogs hard to use? Lessons learned from 6 information-retrieval studies. *Journal of the American Society for Information Science,* 36(6):387-400, 1986.

Brajnik, G, Guida, G and Tasso, C. User modelling in intelligent information retrieval. *Information Processing and Management,* 23(4):305-20, 1987.

Brooks, HM et al. Problem descriptions and user models: developing an intelligent interface for document retrieval systems. *Advances In Intelligent Retrieval: Informatics 8,* London:Aslib, 1985.

Chen, H et al. Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management,* 27 (5):405-32, 1991.

Congreve, J. Problems of subject access: automatic generation of printed indexes and online thesaural control. *Program,* 20 (2): 204-10, 1986..

Cooper, WS. The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness. *Information Processing and Management,* 12:367-75, 1976.

Croft, WB and Thompson, RH. I3R: a new approach to the design of document retrieval systems. *Journal of the American Society for Information Science,* 38 (6):389-404, 1987.

Doszkocs, TE. From research to application: the CITE natural language information retrieval system. *Research and Development in Information Retrieval (Lecture Notes in Computer Science Series. 146),* Berlin: Springer-Verlag, 1983.

Duncan, EB and McAleese. Intelligent access to databases using a thesaurus in graphical form. In *Proceedings of the 11th International Online Information Meeting,* 1987. pp. 337-87.

Hendry, IG et al. INSTRUCT: a teaching package for experimental methods in information retrieval, Part 1: The user's view. *Program,* 20(3): 245-63, 1986.

Hildreth, CR. Intelligent interfaces and retrieval methods for subject searching in bibliographic retrieval systems. *Advances in Library Information Technology, no. 2.* Washington, DC: Library of Congress, 1989.

Katz, WA. 1987. *Introduction to Reference Work. Vol 2: Reference Services and Reference Processes.* 5th ed. New York: McGraw-Hill,1987.

Lotus Development Corp. & Ziff Communications Co. Online user's guide. *Computer Select (Standalone) : Version 2.40. Document No. 171555.* 1991.

Mitev, NN and Walker, S. Information retrieval aids in an online public access catalogue : Automatic intelligent search sequencing. *Advances in Intelligent Retrieval : Informatics 8.* London:Aslib, 1985.

Porter, MF. An algorithm for suffix stripping. *Program,* 14(3):130-37, 1980.

Porter, M et al. Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute. *Program,* 22 (1):1-20,1988.

Smith, PJ et al. Knowledge-based search tactics for an intelligent intermediary system. *ACM Transactions on Information Systems,* 7(3):246-70, 1989.

Vickery, BC. Knowledge representation: a brief review. *Journal of Documentation,* 42(3):145-59, 1986.

Vickery, A and Brooks, HM. PLEXUS - The expert system for referral. *Information Processing and Management,* 23(2):99-117,1987.

Wade, SJ and Willet, P. INSTRUCT: a teaching package for experimental methods in information retrieval, Part III - Browsing, clustering and query expression. *Program,* 22(1):44-61,1988.

Walker, S and Jones, RM. *Improving Subject Retrieval in Online Catalogues: Stemming Automatic Spelling Correction and Cross-Reference Tables.* British Library Research Paper, 24, 1987.

THE AUTHOR

DAVID F C CHAN is the Director of the Centre for Computer Studies at the Ngee Ann Polytechnic, Singapore. He holds an M.Sc. degree in Computational Science from the University of St. Andrew's, U.K. and is a Fellow of the British Computer Society. He has nearly eighteen years of experience in management consultancy, information systems devlopment and management, and education. His R&D interests are in the areas of knowledge-based systems, user-interface design, information storage and retrieval, and software engineering.