DOCUMENT RESUME

ED 441 448 IR 057 719

AUTHOR Khoo, Christopher S. G.; Poo, Danny C. C.; Toh, Teck-Kang; Hong, Glenn

TITLE E-Referencer: Transforming Boolean OPACs to Web Search Engines.

PUB DATE 1999-08-00

NOTE 9p.; In: IFLA Council and General Conference. Conference Programme and Proceedings (65th, Bangkok, Thailand, August 20-28, 1999); see IR 057 674.

AVAILABLE FROM For full text: http://www.ifla.org/IV/ifla65/papers/010-143e.htm.

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Computer Interfaces; Computer Software Development; *Computer System Design; *Expert Systems; *Information Retrieval; *Library Catalogs; *Online Catalogs; Relevance (Information Retrieval); Search Strategies; Subject Index Terms; World Wide Web

IDENTIFIERS Boolean Search Strategy; Intermediaries; *Natural Language; *Search Engines

ABSTRACT

E-Referencer is an expert intermediary system for searching library online public access catalogs (OPACs) on the World Wide Web. It is implemented as a proxy server that mediates the interaction between the user and Boolean OPACs. It transforms a Boolean OPAC into a retrieval system with many of the search capabilities of Web search engines. E-Referencer encapsulates some of the knowledge and searching expertise of experienced librarians. It processes the user's natural language query, maps the query words to Library of Congress subject headings, selects a suitable search strategy, and formulates an appropriate search statement to send to the library system. Based on the user's relevance feedback on the search results, it further selects a strategy for refining the search. (Author/MES)

IFLANET
International Federation of Library Associations and Institutions
Search Contacts
Annual Conference

65th IFLA Council and General Conference

Bangkok, Thailand,
August 20 - August 28, 1999

Conference Proceedings

Code Number: 010-143-E
Division Number: III
Professional Group: Information Technology
Joint Meeting with: -
Meeting Number: 143
Simultaneous Interpretation: *Yes*

E-Referencer: Transforming Boolean OPACs to Web Search Engines

Christopher S. G. Khoo (assgkhoo@ntu.edu.sg)
Division of Information Studies,
School of Applied Science,
Nanyang Technological University

Danny C. C. Poo (dpoo@comp.nus.edu.sg)
Teck-Kang Toh (tohteckk@iscs.nus.edu.sg)
Dept. of Information Systems,
School of Computing, National University of Singapore
Singapore

Glenn Hong (glennhong@nlb.gov.sg)
National Library Board
Singapore

Abstract

E-Referencer is an expert intermediary system for searching library online public access catalogues (OPACs) on the Web. It is implemented as a proxy server that mediates the interaction between the user and Boolean OPACs. It transforms a Boolean OPAC into a retrieval system with many of the search capabilities of Web search engines. E-Referencer encapsulates some of the knowledge and searching expertise of experienced librarians. It processes the user's natural language query, maps the query words to Library of Congress subject headings, selects a suitable search strategy and formulates an appropriate search statement to send to the library system. Based on the user's relevance feedback on the search results, it further selects a strategy for refining the search.

Paper

INTRODUCTION

There is a growing number of search engines that help users locate Web pages of potential interest. Most of these accept queries in natural language, perform fuzzy matching, and display records in ranked order of probable relevance. Some search engines perform query expansion to identify additional terms that are related to the user's query words, as well as query refinement to identify additional Web pages that are similar to those marked relevant by the user.

In contrast, library online public access catalogues (OPACs) on the Web are still difficult to use. Most do not accept natural language queries, requiring users to formulate queries in Boolean expressions. Records retrieved are not ranked, and no assistance with query expansion and query refinement is provided. Current Web interfaces to OPACs are little better than traditional OPACs transplanted on the Web. Borgman (1996) said that most of the improvements to OPACs in recent years were in surface features rather than in the core functionality.

For an increasing number of people, Web search engines will be the type of information retrieval system they are most familiar with. They will find current OPACs archaic and unacceptable, and will ask, „Why can't library catalogues be more like Web search engines?"

E-Referencer is being developed as an answer to this question. It is an expert intermediary system (or expert system interface) that mediates the interaction between the user and a Boolean OPAC. It transforms a Boolean OPAC into a retrieval system with many of the search capabilities of Web search engines. E-Referencer encapsulates some of the knowledge and searching expertise of experienced librarians. It processes the user's natural language query, maps the query words to Library of Congress subject headings, selects a suitable search strategy and formulates an appropriate search statement to send to the library system. Based on the user's relevance feedback on the search results, it further selects a strategy for refining the search.

IMPLEMENTATION

E-Referencer uses the Z39.50 Information Retrieval protocol to communicate with library systems. It makes use of the OCLC Z39.50 client API written in Java (available at the URL http://www.oclc.org/z39.50/#api) and the Java Expert System Shell (available at http://herzberg.ca.sandia.gov/jess/). Technical details of the initial implementation are given in Poo, Toh, & Khoo (1999) and Khoo, Poo, Toh, Liew & Goh (1998). E-Referencer is accessible at the URL http://islab.sas.ntu.edu.sg:8000/E-Referencer/. It currently searches the library systems of the Nanyang Technological University, the National University of Singapore and a few other libraries.

Version 1 of E-Referencer was implemented as a Java applet which could be downloaded and executed by the user's Web browser. It was re-implemented in version 2 as a proxy server between the user's machine (client) and the library system (server). The proxy server is written in Java. A Java applet for communicating with the proxy server is stored on the same machine. Using a Web browser, the user can access and run the E-Referencer applet which connects to the E-Referencer proxy server, which in turn connects to the various library Z39.50 servers. With this model, processing can be distributed between the applet and the proxy program. Processing that requires access to a large knowledge base is performed on the proxy machine, and only the results are sent to the Java applet running on the user's machine. The proxy can also be used to capture user-interface interactions in a transaction log, which can be used for analyzing the effectiveness of E-Referencer and obtaining insights into how it can be improved.

KNOWLEDGE BASE

The intelligence and knowledge in E-Referencer lies in:

1. the conceptual knowledge base that maps free-text keywords to concepts represented by the Library of Congress (LC) subject headings
2. the search strategies coded in the system, including

    a. initial search strategies, used to convert the user's natural language query to an appropriate Boolean search statement
    b. reformulation strategies, used for refining a search based on the results of the previous search statement
    c. the rules for selecting an appropriate search strategy.

Conceptual Knowledge Base

The conceptual knowledge base contains information about which LC subject headings are associated with each free-text keyword. This knowledge base of keyword-subject heading associations was constructed by analyzing about 16 years (1980-1996) of LC catalogue records. For each keyword found in a title, we retrieved all the titles containing the keyword and extracted all the subject headings assigned to these titles. Each of the subject headings was then assigned a score equal to the number of titles that were assigned the subject heading. The raw scores were then normalized by dividing by the highest score (score obtained by the most frequent subject heading). The normalized scores thus reflect how strongly each subject heading is associated with the keyword. As an example, the subject headings that are strongly associated with the keyword Java are given in Table 1. This conceptual knowledge base is used to map users' query words to LC subject headings to use in the search.

Table 1. Subject Headings Associated with the Keyword *Java*

Subject Heading	Raw Score	Normalized Score
Java (Computer program language)	98	1.00
World Wide Web (Information retrieval system)	22	0.22
Object-oriented programming (Computer science)	17	0.17
Java Indonesia-History	8	0.08

Initial Search Strategies

Two initial search strategies have been implemented. *Initial Strategy 1* carries out a keyword search in all fields. *Initial Strategy 2* makes use of the *conceptual knowledge base* described earlier to identify appropriate LC subject headings to use in the search.

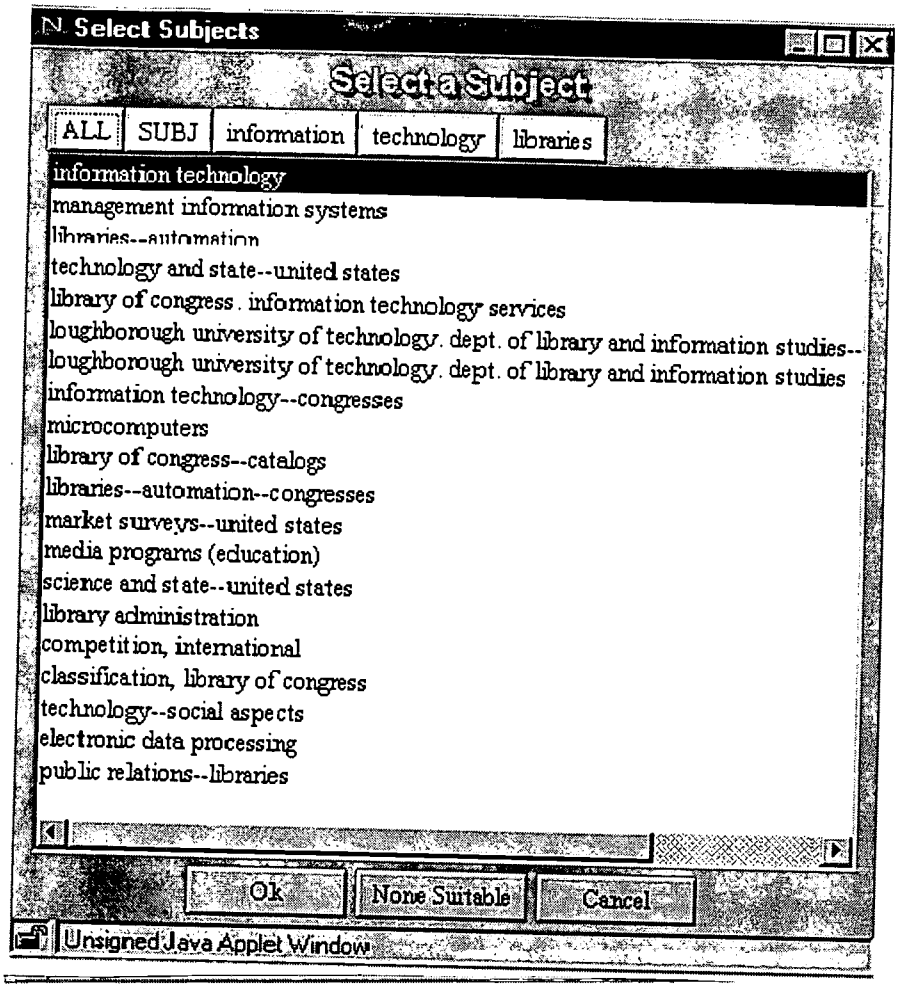

Fig. 1. Subject Headings Displayed by E-Referencer for the Query *Information Technology in Libraries*

Initial Strategy 1 (keyword search) is a simple-minded procedure that removes stopwords, stems the remaining words, and searches for the words in all searchable fields in the library database. Word adjacency is preserved. Punctuation marks and stopwords are replaced with the Boolean AND, but the words in between are retained as phrases.

Initial Strategy 2 (subject heading search) uses the *conceptual knowledge base* to identify twenty LC Subject Headings that are most highly associated with the user's keyword. These are displayed for the user to select. If the user's query contains more than one keyword, the set of subject headings associated with each keyword is retrieved, and the sets of subject headings are then combined. If a subject heading occurs in more than one set, the subject heading is assigned a new score equal to the sum of the scores in the different sets. The 20 subject headings with the highest combined scores are displayed for the user to select. For example, the subject headings that are strongly associated with the query *information technology in libraries* are listed in Fig. 1.

Reformulation Strategies

After the initial strategy is executed, E-Referencer displays the first 20 titles retrieved and prompts the user to indicate which titles are relevant. This is illustrated in Fig. 2. (The user can display more records if the user so chooses.) After the user has indicated which records are relevant, E-Referencer selects and executes one of the reformulation strategies. A reformulation strategy may modify the previous search statement or construct an entirely new search.

Three types of reformulation strategies are used:

1. *Broadening strategies* modify a search statement to make it less constrained in order to retrieve more records. This strategy is appropriate when no record is retrieved by a search, or when most of the records retrieved are relevant and the user wants more records.

2. *Narrowing strategies* modify a search statement to reduce the number of records retrieved. This strategy is appropriate when too many records are retrieved, and the user wants to reduce the set to those records that are more likely to be relevant.

3. *Relevance feedback strategies* analyze the content of the records retrieved to identify terms that are likely to retrieve other relevant documents. Generally, if a term occurs in most of the records found relevant by the user and occurs in few non-relevant records, then it is likely to retrieve other relevant records.

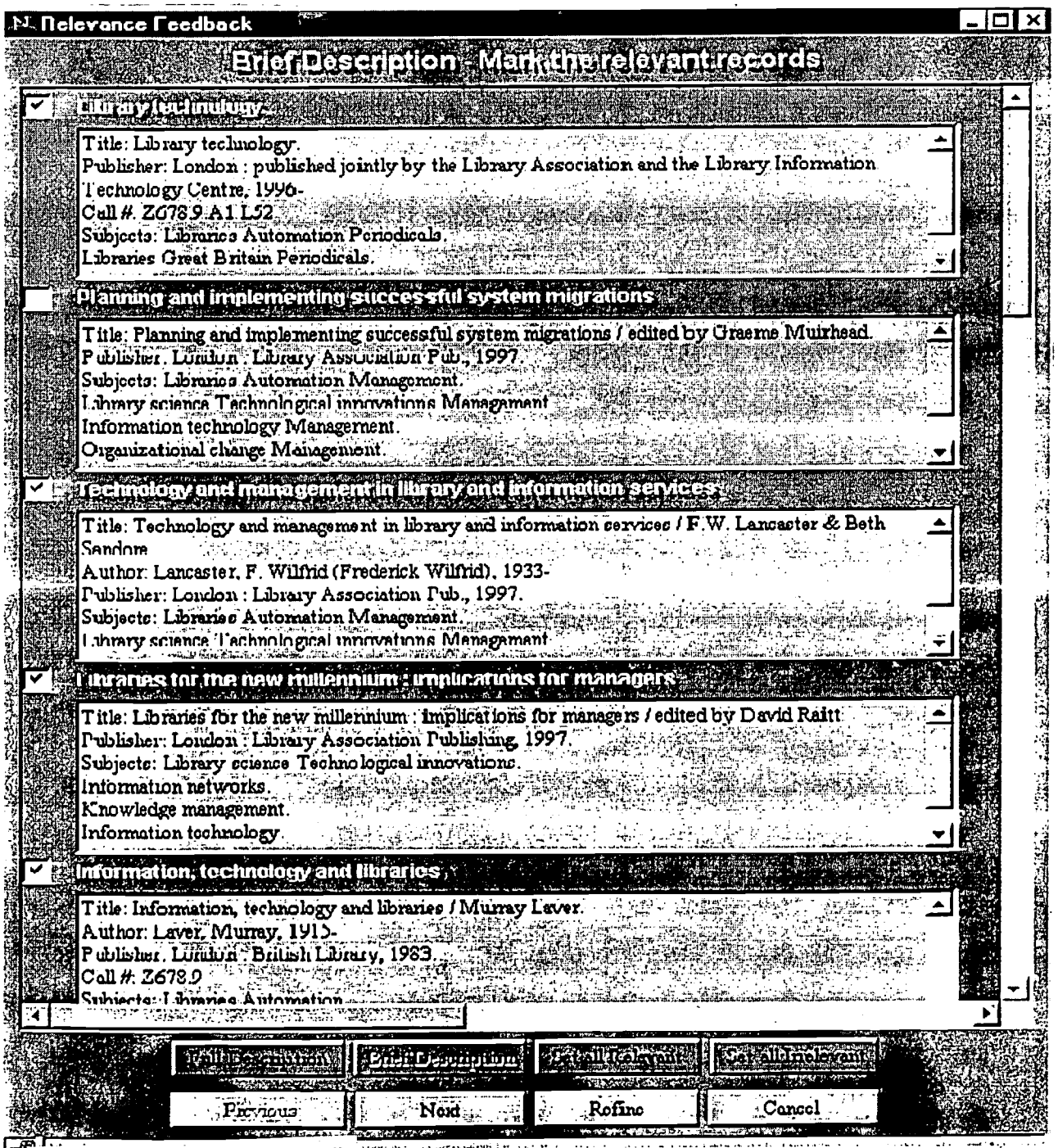

Fig. 2. Search Result Display

Broadening and narrowing strategies are listed in Table 2. For relevance feedback, E-Referencer first compiles a list of keywords and subject headings found in the records displayed to the user. E-Referencer also extracts every combination of two terms from each record. For each term (and combination of terms), E-Referencer calculates a score based on how many relevant and non-relevant records the term is found in. Two formulas are used for calculating the score:

1. Relevance feedback formula 1: number of relevant records containing the term
2. Relevance feedback formula 2: number of relevant records containing the term minus the number of non-relevant records containing the term.

In an earlier study (Khoo, Poo, Toh, Liew & Goh, 1998), we found that these two relevance feedback formulas work well in different situations. We also found that different weights should be assigned to different types of terms. For example, subject headings should be weighted higher than keywords in title. Details of the weighting scheme are given in Khoo et al. (1998).

Table 2. Broadening and Narrowing Strategies

Broadening Strategies
Strategy 1: Convert adjacency operators to Boolean ANDs.
Strategy 2: Search each keyword individually to identify keywords not found in the database. Remove such keywords from the search statement.
Strategy 3: Select every combination of 3 keywords. AND the keywords in each combination. Find the number of records retrieved by each combination of three words. Rank the combinations of 3 words in decreasing order of the number of records retrieved. Start with the combination that retrieved the smallest number of records. Display the records. Then go to the next combination. Display the records and then go on to the next combination. Do this until at least 15 records are displayed.
Strategy 4: Select every combination of 2 keywords. AND the keywords in each combination. Link the combinations with Boolean OR.
Strategy 5: Convert ANDs to ORs.
Strategy 6: Prompt user to enter synonyms and related terms for each keyword.
Narrowing Strategies
Strategy 1: Convert one of the OR operators to AND, and execute the search. Replace the OR operator, convert a different OR operator to AND, and execute the search. Do this for each of the OR operators in turn. Combine all the search sets using OR.
Strategy 2: Convert one of the AND operators to an adjacency operator, and execute the search. Replace the AND operator, repeat the procedure for each of the other AND operators in turn. Combine all the search sets using OR.
Strategy 3: Ask the user for additional keywords to AND to the search.

The rules used to select a reformulation strategy are listed in Table 3. Note that when the initial strategy retrieves fewer than 15 records, broadening strategies are tried in the order listed in Table 2 until at least 15 records are retrieved. E-Referencer displays the records retrieved by the initial strategy first. As the search is broadened, additional records are appended to the end of the display. Since the records retrieved by the broadening strategies are less likely to be relevant than those retrieved by the initial strategy, the result is that E-Referencer displays the records roughly in order of probable relevance.

Table 3. Decision Tree for Selecting a Reformulation Strategy*

**Table 3 is unavailable. Please contact authors.*

EVALUATION AND CURRENT WORK

Experiments with real users are being carried out to evaluate the effectiveness of E-Referencer compared with two Web OPAC interfaces. The results of a preliminary study suggest that E-Referencer's Initial Strategy 1, coupled with the automatic broadening strategy, is more effective than a Web OPAC interface in helping users retrieve relevant records. On average, the subjects retrieved 7.1 relevant records using E-Referencer compared with 5.4 relevant records using the Web OPAC interface. The subjects were 14 graduate students in the Division of Information Studies at Nanyang Technological University. The subjects had taken a course in online searching and were familiar with the Web OPAC interface! An evaluation study involving undergraduate and graduate students in several disciplines is still in progress.

Other work in progress includes:

Exploring how neural networks can be used to identify good search terms to use during relevance feedback.

Developing a browse-search interface to allow searching by browsing a network of LC subject headings and the LC Classification Scheme.

Investigating the feasibility of using Initial Strategy 2 (subject heading search) as a cataloguing aid. Given the title of a new book, E-Referencer can suggest subject headings for the cataloguer to assign to the book.

Extending E-Referencer to search abstracts and full-text databases.

REFERENCES

Borgman, C.L. (1996). Why are online catalogues still hard to use? Journal of the American Society for Information Science, 47, 493-503.

Khoo, C., Poo, D., Toh, T.K., Liew, S.K., & Goh, A. (1998). E-Referencer: A prototype expert system Web interface to online catalogs. In C. Nikolaou & C. Stephanidis (Eds.), Research and Advanced Technology for Digital Libraries, 2nd European Conference (ECDL'98), 1998 (pp. 315-333). Berlin: Springer-Verlag.

Poo, D.C.C., Toh, T.K., & Khoo, C.S.G. (1999). Search interface for Z39.50 compliant online catalogs over the Internet. In Proceedings of the 32nd Annual Hawaii International Conference on System Sciences (HICSS-32), Software Technology Track, Multi Media Database and Internet Mini Track, 1999 (pp. 50-57). New York: IEEE.

Latest Revision: *June 1, 1999*

www.ifla.org

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (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").