DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Syracuse University School of Library Science Syracuse, New York 13210	UNCLASSIFIED
	2b. GROUP N/A

3. REPORT TITLE

FREE TEXT RETRIEVAL EVALUATION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Final Report 1 February 1971 - 31 January 1972

5. AUTHOR(S) (First name, middle initial, last name)

Pauline Atherton
Kenneth H. Cook
Jeffrey Katzer

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
July 1972	196	6

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F30602-71-C-0185 Job Order No. 45940000	None
	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) RADC-TR-72-159

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
None	Rome Air Development Center (IRDT) Griffiss Air Force Base, New York 13440

13. ABSTRACT

The basic problems this research effort investigated were (1) the development, implementation, and evaluation of algorithms to improve recall levels in interactive, free-text retrieval using a modified version of IBM's Document Processing System (DPS), (2) the development of techniques for increasing the vocabulary capacity of DPS, and (3) unobtrusive statistical data gathering of system use, growth, and cost, through a previously developed computer program. A free-text document data base (DDB) of 46,828 bibliographic citations and abstracts from Psychological Abstracts was developed. Also, two interactively accessible data bases were developed and implemented to provide free-text vocabulary control and recall improvement directly to the user. No intermediaries were used in the retrieval process. These two algorithms were (1) a Vocabulary Data Base (VDB) containing the 106,702 unique free-text terms from the inverted file of processed documents in the DDB, and (2) a Search Data Base (SDB) containing previously submitted user search inquiries to the DDB. A two-month period of experimental use of the entire system with all three data bases by students, staff, and faculty of Syracuse University in fall 1971 provided the required "real-life" field environment. A total of 2399 search inquiries were submitted via the 2741 terminals. The system operated under both the 360/50 and the 370/155 operating systems. The capacity of the DPS vocabulary was increased by the development and successful implementation of computer programs that revised the DPS coding structure of the vocabulary file. The newly developed structure changing the 16-bit coding to 32-bit coding resulted in increasing the vocabulary capacity

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Abstracts						
Indexes						
Information Retrieval						

13. Abstract (continued)

from the former single-file limit of 65,534 to over 4 billion terms. An extensive user-oriented public relations/publicity, instruction, and education package was developed and implemented. This user emphasis resulted in a significantly greater number of registrants and actual users in the system, which did not use search intermediaries, than during a similar fall 1970 period. Results of a controlled cost-performance (recall) study indicated that use of the VDB or the SDB yielded better cost-performance levels, especially at higher recall percentages, than by use of the DDB alone. These results are the initial findings, and would require additional testing to substantiate their validity. Other evaluative techniques included a Semantic Differential attitude scale for interactive retrieval systems, a structured telephone interview of users, and a special number users could call for help in developing search inquiries. The STATPAC program for unobtrusively monitoring and retrieving data on system use, growth, and cost was modified and successfully used to provide evaluative data. General conclusions are that interactive, free-text cost-performance (recall) levels can be improved by direct user control of algorithms providing vocabulary control. Real-life application of these algorithms suggests that the recall improvement available through controlled vocabulary or indexing systems might be obtained readily by the free-text searcher who is provided with techniques such as those implemented in this research effort.

FREE TEXT RETRIEVAL EVALUATION

Pauline Atherton
Kenneth H. Cook
Jeffrey Katzer

Syracuse University

ic

FOREWORD

This report was done by the Syracuse University Psychological Abstracts Retrieval Service (SUPARS) Research Group at the Syracuse University School of Library Science, under contract F30602-71-C-0185, Job Order Number 45940000, for Rome Air Development Center, Griffiss Air Force Base, New York. Mr. Nicholas M. DiFondi (IRDT) was the RADC Project Engineer.

This report represents a continuation of work conducted under contracts F30602-69-C-0013 and F30602-70-C-0190 during the period 1 July 1969 and 31 January 1971.

The current reseach contract covers work accomplished during 1 February 1971 to 31 January 1972, and deals with the development, implementation, and evaluation of new algorithms to improve recall in an interactive, on-line, free-text retrieval system.

Individual authors of different sections include: Section I, Kenneth H. Cook; Section II, Lynn Trump and Mr. Cook; Section III, Sandra Browning, June Brower, Jeffrey Katzer, Patricia Moell, and Peggy Mucia; and Section IV, Mr. Katzer.

Acknowledgement is given to the following individuals: Dean Roger C. Greer and Dr. Allan F. Hershfield of the School of Library Science who provided continued support for the SUPARS project; Mr. William J. Jones, Director, Syracuse University Computing Center, for his cooperation and assistance in the implementation of the on-line system; and to Mrs. Brenda Lefebvre, whose overall editorial and typing made this report possible.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS).

Nicholas M. DiFondi

Approved: NICHOLAS M. DIFONDI
Technical Evaluator

Approved: FRANZ H. DETTMER
Colonel, USAF
Chief, Intel & Recon Division

FOR THE COMMANDER: Fred I. Diamond

FRED I. DIAMOND
Acting Chief, Plans Office

ABSTRACT

The basic problems the current research effort (February 1, 1971-January 31, 1972) investigated were (1) the development, implementation, and evaluation of algorithms to improve recall levels in interactive, free-text retrieval using a modified version of IBM's Document Processing System (DPS), (2) the development of techniques for increasing the vocabulary capacity of DPS, and (3) unobtrusive statistical data gathering of system use, growth and cost through a previously developed computer program. A free-text document data base (DDB) of 46,828 bibliographic citations and abstracts from Psychological Abstracts was developed. Also, two interactively accessible data bases were developed and implemented to provide free-text vocabulary control and recall improvement directly to the user. No intermediaries were used in the retrieval process. These two algorithms were (1) a Vocabulary Data Base (VDB) containing the 106,702 unique free-text terms from the inverted file of processed documents in the DDB, and (2) a Search Data Base (SDB) containing previously submitted user search inquiries to the DDB. A two-month period of experimental use of the entire system with all three data bases by students, staff, and faculty of Syracuse University in fall 1971 provided the required "real-life" field environment. A total of 2399 search inquiries were submitted via the 2741 terminals. The system operated under both the 360/50 and the 370/155 operating systems. The capacity of the DPS vocabulary was increased by the development and successful implementation of computer programs that revised the DPS coding structure of the vocabulary file. The newly developed structure changing the 16-bit coding to 32-bit coding resulted in increasing the vocabulary capacity from the former single-file limit of 65,534 to over 4 billion terms. An extensive user-oriented public relations/publicity, instruction and education package was developed and implemented. This user emphasis resulted in a significantly greater number of registrants and actual users in the system, which did not use search intermediaries, than during a similar fall 1970 period. Results of a controlled cost-performance (recall) study indicated that use of the VDB or the SDB yielded better cost-performance levels, especially at higher recall percentages, than by use of the DDB alone. These results are the initial findings, and would require additional testing to substantiate their validity. Other evaluative techniques included a Semantic Differential attitude scale for interactive retrieval systems, a structured telephone interview of users, and a special number users could call for help in developing search inquiries. The STATPAC program for unobtrusively monitoring and retrieving data on system use, growth, and cost was modified and successfully used to provide evaluative data. General conclusions are that interactive, free-text cost-performance (recall) levels can be improved by direct user control of algorithms providing vocabulary control. Real-life application of these algorithms suggests that the recall improvement available through controlled vocabulary or indexing systems might be obtained as readily by the free-text searcher who is provided with techniques such as those implemented in this research effort.

EVALUATION

The objective of this study was to develop, implement and evaluate methods for increasing vocabulary file space and improving the retrieval effectiveness of a free-text indexed on-line document retrieval system at Syracuse University. The system first operated on an IBM 360/50 computer and recently on a 370/155 computer. The document data base consisted of 46,828 bibliographic citations and/or abstracts from Physchological Abstracts. Vocabulary storage space was increased by developing computer programs to convert the half-word (16-bit) coding scheme as defined by the IBM/DPS program to full word (32 bit) coding. Methods of improving retrieval effectiveness include a vocabulary data base and a search data base as on-line searching aids. Results are reported in terms of nine levels of Recall (the portion of relevant documents retrieved), total retrieval (the number of documents retrieved to achieve a specific Recall level), and cost-performance (the cost incurred to achieve a specific Recall level).

There are several significant conclusions derived from the results of this effort:

1. The conversion from 16 bit to 32 bit coding has increased the limit on a single vocabulary file size from approximately 65,000 words to over 4 billion words. Without this capability, upon reaching the 65,000 word limit a new vocabulary would have to be defined and then created by the free-text indexing program resulting in inefficient use of core due to redundant information between files and longer search cycles.

2. The use of the vocabulary data base as a searching aid results in better cost performance than using the other data bases. Since the vocabulary data base is a portion of the inverted file developed from the original processing of documents, it is relatively inexpensive to add to the system and reduces cost performance by decreasing the need to search the more expensive document data base.

3. Statistically significant differences in total retrieval at all levels except the 10% level of Recall reflected variations in efficient use of the system by search experts. Each expert was equally knowledgable in the subject area, equally trained in the use of free-text retrieval, and used the same information requirement statements to formulate his searches.

However, each chose a different search strategy in the hopes of minimizing total retrieval. This finding indicates that it is difficult for system experts to find or establish efficient search methods. Since most users may be knowledgeable in their fields but not necessarily experts in the use of on-line retrieval, attempts by novice users to formulate efficient search strategies may be difficult to achieve.

4. Total document retrieval is very high at all levels of Recall. Although use of the vocabulary data base or the search data base does reduce total retrieval considerably from that achieved by using the document data base, too much time and effort would be required of the user to scan for relevant documents.

As a result of this study future research can be directed toward establishing better search methods to reduce total retrieval, continue work on the Recall improving algorithms to insure total retrieval improvement does not negate their effect on system effectiveness, and identify differences in the experts methods of searching to determine the impact on the general population of users and make adjustments accordingly.

Nicholas M. Difondi

NICHOLAS M. DIFONDI

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF FIGURES FOR APPENDICES

APPENDIX V

APPENDIX VIII

USAGE OF TERMS

Because the definitions of terms used in the information technology field are not completely standardized nor consistent, the terms in this report are explained below. An effort has been made, where possible, to follow the most consistently used and reasonable meaning to convey a concept. Where specialized or more specific usages of these terms are employed in specific sections of this report, an explanation will be given by the section author.

(a) Delta: The character (upshift "H") on the 2741 keyboard that is used in SUPARS/DPS user interaction to access the document data base and initiate a search inquiry.

(b) Delta S: The character "delta" and the letter "S" (S) which are used in SUPARS/DPS user interaction to access the search data base and initiate a search inquiry.

(c) Delta V: The character "delta" and the letter "V" (V) which are used in SUPARS/DPS user interaction to access the vocabulary data base and initiate a search inquiry.

(d) Dictionary: The internally stored list of unique free-text terms processed by DPS and the document frequency count for each work. The DPS dictionary forms one part of the inverted file.

(e) Document: In this study, the term "document" stands for the bibliographic citation and abstract that are used as a surrogate of the original journal, article, proceeding, book, etc.

(f) Document Data Base (DDB): Consists of SUPARS/DPS processed documents of Psychological Abstracts which are interactively accessible on-line by users. The DDB is one of three data bases available to the SUPARS/DPS user (others are vocabulary data base and search data base).

(g) Document Processing System (DPS): The IBM free-text, batch mode programs that convert machine readable textual data into searchable and retrievable data sets organized in inverted file structure.

(h) Free-Text: Specifically refers in this study to the Document Processing System. The general reference is to an organized system allowing the indexing and retrieval of documents or their surrogates by any of the terms used in a defined text, rather than terms defined by a controlled set of terms.

(i) Information Requirement Statement (IRS): The verbal or written statement of an individual's interest area as generally related to documents or their surrogates. The IRS is the publically verifyable indication of the internally held construct, "information

(j) Label (labelled line): The portion of a search inquiry, such as L1, L2, etc. that identifies and stands for the search words and operators used to act on those words; a label can be used itself in an inquiry as a search word.

(k) Operator: The user language accepts as input one or more keywords which represent the IRS of the searcher. Keywords may be combined with Boolean operators (AND, OR, NOT) or grammatical operators (those which specify the desired proximity of keywords within a sentence or those which specify the root of a word).

(l) Search: The search inquiry, the user/computer interaction, and the printed output, if any. The beginning of a new search inquiry marks the end of a search.

(m) Search Data Base (SDB): Consists of the previously stored and processed search inquiries made to SUPARS/DPS. The SDB is one of three data bases available to the user and was newly developed during the current research. (Others are the document data base and the vocabulary data base.)

(n) Search Inquiry: The user arrangement of words, word combinations and logical operators in a form acceptable as input for machine processing. A SUPARS/DPS search inquiry would consist of the free-text terms combined with Boolean and other logical operators, the request for output, and an "end" statement. Examples of search inquiries are given in Section II.

(o) Search Word: Free-text word(s) or term(s) used as part of a search inquiry. A KEYWORD is a synonym for a search word.

(p) STATPAC: The Statistical Package used in conjunction with SUPARS/DPS to unobtrusively collect, store and retrieve the elements of user interaction and other system parameters such as time, terminal number, cost, etc. STATPAC includes a highly flexible retrieval system in itself which allows the operator to specify and retrieve various combinations of data reflecting user interaction or system performance. In addition to standard summaries printed periodically, the operator could request, for example, a listing of the computer time used for all searches of the document data base by one-time users after a certain calendar date.

(q) SUPARS/DPS: Syracuse University Psychological Abstracts Retrieval Service/Document Processing System. The modified DPS program developed at Syracuse University which allows on-line, interactive searching of free-text data. SUPARS/DPS I refers to the research work conducted from July 1969 to January 1971. SUPARS/DPS II refers to the work conducted from February 1971-January 1972.

(r) Vocabulary: The on-line, interactively accessible dictionary that is stored by DPS. The term "vocabulary" rather than "dictionary" is used to connote the words and terms accessible to the user

that can be used as free-text index terms.

(s) Vocabulary Data Base (VDB): Consists of the on-line, interactively accessible DPS dictionary of free-text terms. The VDB is one of three data bases available to the user and was newly developed during the current research. (Others are document data base and search data base).