QUARTERLY REPORT

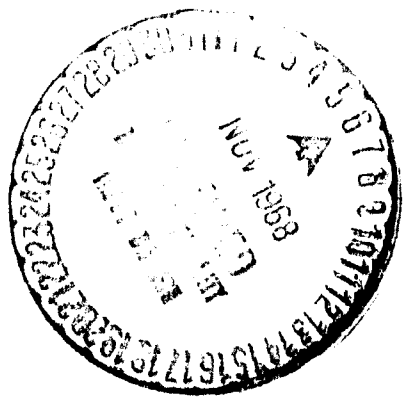
to the

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Period ending 10-15-68

Contract NSR 39-011-076

University of Pittsburgh
Knowledge Availability Systems Center
October 1968



INTRODUCTION

Under the provisions of Contract No. NSR 39-011-076, "Contract for special experimental projects involving information systems and technology utilization," each of the six projects involved are to be accounted for separately in reporting. However, in this quarterly report, only four of the six projects are reported since, in response to a request from the National Aeronautics and Space Administration (NASA), the Knowledge Availability Systems Center (KASC) is furnishing final reports separately for projects #4, Small Business Administration Participants Study, and #6, Study of Fee-Paying Industrial Client Attrition, at the end of October or early in November 1968.

The four projects covered by this report are:

Project #1: Investigation of Means for Improving Exploitation of Indexes

Project #2: Thesaural Development to Permit Improved Relevance Predictability of Searches

Project #3: Investigation of Comparative Relevance of Search Results

Project #5: Demand Searches

Project #1

Investigation of Means for Improving Exploitation of Indexes

I. Background

In the initial development of this study that has as its objective the "...consolidation of similar searches in order to achieve production economies and the improvement of confidence in exhaustiveness of searches for industry over and above what may be achieved through searches of the indexes as provided by the government," the need for a natural

text searching technique was recognized.

Since Project #3, Investigation of Comparative Relevance of Search Results, also has need of a technique for searching natural text, progress on this project is reported only once (page 3).

Project #2

Thesaural Development to Permit Improved Relevance Predictability of Searches

I. Background

The objective of this study is to attempt to predict the relevance of information disseminated to industry. With the release this year of the NASA Thesaurus, a first step has been to determine the applicability of this Thesaurus to the pre-thesaural file.

II. Progress to date

Data Base--As a data base, two hundred and fifty seven strategies used during one "current-awareness"--retrospective search period at the University of Pittsburgh -- Regional Dissemination Center were selected (from the more than seven hundred strategies searched during that period) using the following criteria: 1) the strategy must have been applicable to only one question statement; 2) the computer search must have resulted in citations some or all of which were judged relevant by the analyst and forwarded to the user; 3) the user must have returned an evaluation sheet indicating the "relatedness" of the computer citations to his question, and, 4) the question must still have been active after the effective date of the Thesaurus.

Term Identification--This set of strategies included a total of more than 2,300 terms of which more than half were unique. Because the sample period predated the effective date of the Thesaurus, the terms were in Subject Authority List (SAL) form. They were compared with the Thesaurus and the most closely related thesaural term was assigned as an equivalent term.

Program for term evaluation--This list of terms was keypunched and used to read from a tape of the Thesaurus onto another tape the strategy terms as main terms and all of the appropriate subterms. This tape was then sorted alphabetically and listed. The resulting list consisted of approximately 22,000 terms of which more than 8,000 were unique. The most closely related Subject Authority List term was used to replace the main terms and the subterms and the list was re-sorted by line sequence number. Because of the space limitations imposed by resequencing by line number, an additional 2,000 "used for" entries subsumed by terms on the list other than main terms (the original strategy terms) were not included on the printout of the thesaurally arranged Subject Authority List.

Future plans--The subterms are being used for single aspect searches of the sample period and more than half of those searches are now complete. It is planned to complete the searches during the next quarter. Also, a form is to be designed which will permit convenient recording of the results of the searches so that an anaylsis of the results may proceed.

Project #3

Investigation of Comparative Relevance of Search Results

I. Background

In previous quarterly reports* the usefulness of index entries

*Knowledge Availability Systems Center, University of Pittsburgh, "Special Experimental Projects Involving Information Systems and Technology Utilization," Quarterly Progress Reports to the National Aeronautics and Space Administration, Nos. 1 and 2 on Contract No. 39-011-076, Dec.67 and Apr.68.

as the basis for retrieving information packaged in units called "documents," was seriously questioned. Not only has the reliability of the procedure been questioned, but also, the efficiency and effectiveness of conducting literature searches in this manner. This doubt led to the formation of an hypothesis that selected document surrogates can be keypunched in machine-readable form and searched with existing search strategies (with some systematic alteration and/or modification to accommodate the differences in the structure of index terms vs. text terms). For example, a document index entry might be "holography," yet this term may never occur in the text proper of this document; rather, one might find such terms as "hologram(s)," "holographic," "three-dimensional imagine," etc.

It is hoped that the output of searches on the text of "document surrogates" will have a significant increase in recall value (citing a greater proportion of all the relevant material in the file) and in precision (not citing documents that the user does not want) than can be achieved with index entries alone.

II. Experiment design

To test the hypothesis, a 360/model 40 computer will be employed at various stages of the experiment. The first computer task will be to scan the text of the document surrogates, and to reorganize it in a more useful form. The form of this reorganization will be a listing of each occurrence of each word in the surrogate texts (excluding "common words" which have only a functional use in the sentence). Accompanying the occurrence of each word will be identifying information (i.e., location co-ordinates reflecting information along four-dimensions; that is, they uniquely identify a word by indicating which document, which surrogate type, which sentence number, and the ordinal position of the word in the sentence). This identifying information will allow

for hypotheses to be constructed in relation to the textual environment and contiguity associations of each word.

Since most computer programs that form concordances take strings of characters delimited by blank spaces as their basic "words," some anomalies occur which must be corrected either by increasing the level of sophistication of the computer program, or by some form of post-editing on the concordance. The latter has been chosen as the alternative to be used in this experiment. An example of the problem can be seen in such text terms as "Q SWITCH (LASER)" and "ET AL" in which parts of words appear in the concordance as whole words as follows:

AL
ET
LASER
Q
SWITCH

Words which carry no useful information of themselves can be stricken from the concordance (e.g., AL and ET). Other words that carry information can be rejuxtaposed by either boolean logic in the strategy, or by employing word distance restrictors in the strategy, or by using the word locator co-ordinates.

A "word" and its location co-ordinates form a concordance "line." These lines are then alphabetized according to existing lexical rules for forming dictionary entries. The alphabetized listing (concordance) will then form the searchable data base to be employed in conducting the experiment.

A concordance is a very efficient data base to search. A computer can employ a "binary" search procedure which is extremely fast in searching long lists. For example, if a word is in a list of 16,000 terms, it can be found using the binary search technique in fourteen (14) computer tests/comparisons.

A series of search "passes" will be made on the concordance using and indeterminate number of search strategies (search formulas). Information as to how to best formulate the search strategies to be able to search the new medium (text of document surrogates) will hopefully precipitate from an examination of the concordance itself. Some searching and experimenting will have to be done manually on the concordance to gain a working knowledge of its contents (i.e., the classes of words, their interrelationships and their intra-relationships). This knowledge should give clues as to how to systematically alter and/or modify the strategies. One piece of information available from the concordance that should be of interest is the number of text words that are also terms in the NASA Thesaurus and/or the pre-thesaurus "Subject Authority List." These would be text words that overlap with the index terms and can be used directly in strategy formulation.

The questions for which searches will be made in this experiment will be selected from a delimited population of 426 actual questions used in the NASA Regional Dissemination Program at the University of Pittsburgh. Only those questions for which adequate user feedback information, as to the potential relevancy of the document identified for a given question, will be included in the test sample. One search pass has already been made; i.e., terms in the strategies have already been matched against the index terms of the documents in the experimental file. Those documents meeting the required search logic have been cited. This search, which has been labeled the criterion search, will serve as the bench mark against which the output of all experimental searches will be compared and evaluated.

III. Progress to Date

Data Base--The data base population is:

(a) File--The text of 1,195 sets of document surrogates (i.e., titles, abstracts, first paragraphs, last paragraphs, index terms, and notations of content) have been keypunched into machine-readable form. This is roughly a 20% random sample of the monthly accession to the NASA file for the April 1967 search period.

(b) Strategies--A concordance of all the search terms constituting the delimited population of search strategies has been constructed and put into machine-readable form. This listing has 1790 terms; 910 were single-word terms, 880 were multiple-word terms.

Pilot Searches--Several pilot searches were done manually on the experimental file to find parameters which may serve as guidelines for the alteration and/or modification of search strategies. Two important findings of these manual searches were reported previously* and are summarized briefly below.

(a) Multi-word descriptors are occurring with greater frequency in the search strategies being composed by subject specialists/analysts. This basically reflects the increased use of multi-word descriptors for the indexing of NASA documents. However, the frequency with

---

* ibid, Report #2, pp 2-4.

which multi-word descriptors are found in natural text is very low. It does not appear that an easy transition from index-term searches to text-term searches can be performed.

(b) In most cases an exact match of strategy term and text term does not occur. However, in most of these cases a terminal modification of an (s) or (es); i.e., a change in number; is all that is required. In the remaining cases, a more complex modification would be required to exact a match.*

It is anticipated that the concordance will give additional insight into these two problems as they relate to searching natural text, especially scientific-technical text as found in the NASA file.

Concordance Program--Orginally it had been proposed to write a program to form the concordance from the text of the document surrogates. A program was written in PENELOPE, the University of Pittsburgh's Natural Language Processor, for implementation of the University's 360/model 50 computer. However, the PENELOPE compiler, because of its recent adaption for use on the 360/50 system, was the source of unexpected difficulties. Program patches were inserted to bypass all the known and identifiable difficulties. After several months of debugging, this program was determined to be inefficient in its present form and has been shelved.

Contact has been made with ASPEN Systems and negotiations begun for the compilation of the concordance. The system of programs used at

---

* Op. Cit., Report #1, p. 24.

ASPEN was written in 360 OS Basic Assembler Language and is currently used in text processing and searching applications.

IV. Time Schedule

Estimated time for completion of the Concordance is one month from the submission date of the machine-readable text to ASPEN.

Upon receipt of the concordance, it is reasonable to estimate 6 months to examine the concordance, to formulate modified strategies, and to complete the searches (May 30, 1969).

It is also reasonable to assume that all the necessary data analysis can be completed in 4 months upon completion of the searches. Estimated completion is September 30, 1969.

Project #5

Demand Searches

Under the provisions of the subject contract item, KASC is to provide retrospective searches on demand to NASA responsible personnel or their designated addressees.

During the reporting period, no funds were expended under the provisions of Article I, Item No. 5, of NASA Contract No. NSR 39-011-076.