QUARTERLY REPORT

to the

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Contract NSR 39-011-076

University of Pittsburgh
Knowledge Availability Systems Center
July 1968

INTRODUCTION

In accordance with the "Reporting Requirements" for Contract NSR 39-011-076, each of six projects is accounted for separately in this report and appear as follows:

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 #4: Small Business Administration Participants Study

Project #5: Demand Searches

Project #6: Study of Fee-Paying Industrial Client Attrition

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--The data base is the approximately three hundred strategies from one retrospective "current awareness" search period for which there is feedback from the user concerning the relevance of the computer citations to the question statement.

Terms Identification--The unexpectedly large number of unique terms used in the strategies (more than half of the terms were used in only one strategy) precluded the use of manual techniques in listing the subterms included under each main (strategy) term and it was found necessary to mechanize as much of this part of the project as possible. Each unique term in the subset of strategies forming the data base was visually equated with a term in the Thesaurus and keypunched. This deck was used to pull and read onto another tape the strategy term (as a main term) and its subterms from a machine-

readable copy of the Thesaurus. This step is still in progress.

Project Plans:

Program for term evaluation--After the main terms from the strategies and the appropriate subterms have been written onto another tape, they will be sorted alphabetically and visually equated with Subject Authority List (SAL) terminology. Appropriate term modifications will be made and the terms re-sorted by line sequence number. This will result in a thesaurally structured SAL composed of terms peculiar to the subset of strategies included in the data base.

Further plans will depend upon the results of single aspect searches for each of the terms used in the sample period.

Project #3

Investigation of Comparative Relevance of Search Results

I. Background

The comparative indexing investigation has been designed to empirically test the hypothesis that certain document surrogates* in machine-readable form can be searched with existing strategies (that had previously been formulated for the specific purpose of searching the index terms in the NASA file) with the same effectiveness as would result from searching the file on index terms alone. Effectiveness will be measured in relation to the traditional recall value, i.e., the ability of the experimental searches to recover a proportional part of all the relevant material in the file that is at least as large as would result by searching the index terms, and precision (screening out documents the user doesn't want).

The experiment that has been designed to test this hypothesis consists

* Title, abstract, first paragraph, last paragraph, index terms, and notation of content.

of the following: a file of 1,200 sets of six document surrogates has been keypunched; a delimited population of 426 actual questions that have been posed to the NASA system by participants in the regional dissemination program has been selected; and two computer programs for text searching are being examined.

II. Progress to Date

Data Base--The data base population was defined in the first Quarterly Report and the completion of the key punching of the sample file (1,200 sets of six document surrogates) was noted in the second Quarterly Report. Also noted in the second report was the key punching of the search terms which constitute the population of search strategies.

Concordance--The decision to form a concordance of the text was influenced greatly by the existing capability of the computer facilities at this University and by the experimental design which may require several runs over the same data base. A concordance can be searched rapidly and economically once it has been formed.

A program was written to take a sample of the data base and form a concordance. This program is written in PENELOPE, a natural language processor which can scan text word by word and sentence by sentence. It is written for and available on the IBM 360/50. Unfortunately, the text had already been keypunched with a format that was designed for a character scan using Fortran II on the IBM 7090. This format broke most of the words at the end of each card image in an unnatural place and requires a concatenation to reform the word if PENELOPE is to be used. Other problem areas involve punctuation, abbreviations, special characters, formulae (superscripts, subscripts, and symbols), and references. At this moment, PENELOPE does not have a masking capability to permit the extracting of root words from words encased with either

a prefix or a suffix, or both. The feasibility of incorporating this feature into the processor language has been suggested and is being studied by two staff members of the Computation Center.

Program to Search the Concordance--The second computer program currently in an experimental stage is one to search the concordance by matching terms in the search strategy with terms in the concordance. A match will then be appropriately scored and weighted. A criterion weight for each strategy can be used to determine when a search strategy has been successfully met by key-words in one of the several document surrogates.

Program of Choice--The program of choice for the experiment is one currently operational at IBM's Technical Information Retrieval Center. This program utilizes the IBM 360/50 normal text system for storing and retrieving documents on a current-awareness basis. Negotiations have been taking place with IBM since last March to secure this program on a pre-release basis (normal release is scheduled for 1 Sept 1968). The program requires a 256 K core storage capability which is not currently available at the University of Pittsburgh although expected in the foreseeable future (six months). We are now awaiting delivery of a documented copy of this program for additional exploration in this direction.

Project #4

Small Business Administration Participants Study

I. Background

To aid in the determination of the suitability of Regional Dissemination Center (RDC) services for small business organizations, a study of the six (6) small businesses who participated in the program sponsered by the National Aeronautics and Space Administration in conjunction with the Small Business Administration is being conducted.

The creation of a "small business profile" is being attempted to determine if there are any indicators that would allow identification of businesses to whom the service would be most beneficial. To do this, data concerning the company background, financial information, employees, and service received from their association with the Regional Dissemination Center is needed.

It is hoped that by comparison within the six small businesses as to those that continued the service after the year and those that did not continue, and, also, by comparison with other companies of the RDC (by means of data colledted for Project #6, Study of Fee-Paying Industrial Client Attrition), insight into the problem will be gained.

II. Progress to date

Data Base--The same collection techniques used in Project #6 (page 7) were performed. Unfortunately, public information was not available for all of the businesses (Table 1). This situation now necessitates the identification of the most desirable and obtainable public information and the contacting of responsible representatives of those companies for which the information is needed. Through previous conversations with Mr. J. Shaundel, SBA, Cleveland, this course was recommended.

Table 1

Data Available for Small Businesses

(X = Data Collected, NA = Data Not Available)

NAME OF SMALL BUSINESS	TYPES OF INFORMATION		
	Service Data from KAS	Observations from Engineering Consultants	Public and Financial Information
Astro Metallurgical Corporation	X	X	NA
Crobaugh Laboratories, Incorporated	X	X	NA
Gilmore Industries	X	X	X
Horizons Incorporated	X	X	NA
Rand Development Corporation	X	X	X
Reuter-Stokes Electronic Components, INC	X	X	NA

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, a considerable number of demand searches were performed both for NASA and their designees. However, due to the reorientation of NASA Contract No. NSR 39-011-064, no funds were expended under the provision of Article I, Item No. 5, of NASA Contract No. NSR 39-0110-76.

Project #6

Study of Fee-Paying Industrial Client Attrition

I. Background

Since one of the objectives of this study is the determination of means of better orienting services offered to current and prospective clients by the examination of the attrition of fee-paying industrial clients, it might be thought that the starting point of the study would be the reasons provided (if any) at the time the decision was made not to renew the service. However, the value of such comments is questionable (also not available for every case). It was decided, therefore, that additional information concerning each client was needed in order to discern any common denominator (or denominators) that could reveal reasons or combinations of reasons for not continuing the service. This additional information falls into two (2) categories: (1) public and financial information; and (2) information on service received from the Knowledge Availability Systems Center (KASC). In addition, it was felt that a comparison between the attrition clients and the clients still active in the Regional Dissemination effort of KASC would be useful in identifying any possible patterns.

The steps necessary to accomplish this task are:

(1) identification of attrition clients and active clients;

(2) identification and collection of additional information about the clients; and,

(3) analysis of collected data.

II. Progress to date

Identification of Attrition Clients and Active Clients. In order to establish the population for consideration, it was arbitrarily decided to include only those fee-paying clients, active or attrition, in the service from May 1964 to March 1968. From the 101 fee-paying clients of this period, 59 were identified as active and 42 were identified as attrition.

Identification and Collection of Additional Information. Both the identification and collection of the additional information was performed separately for the categories identified above.

1. Public and financial information. From available, standard references* types of information of possible importance were identified (Table 2). A form for recording purposes was designed, and the standard references were checked for each of the 101 fee-paying clients previously identified. Of the 59 active clients, complete public and financial information was available for only 42; and for the 42 attrition clients public and financial information was obtained for only 18.

2. Information on service received from the KASC. For purposes of data collection, two areas were identified: (a) the quantitative information from the KASC files, and (b) the qualitative information available from the engineering consultants to the Regional Dissemination activity.

---

*Middle Market Directory, New York: Dunn & Bradstreet, 1968.
Million Dollar Directory, New York: Dunn & Bradstreet, 1968.
Moody's Handbook of Common Stocks, New York: Moody's Investor Service, 1967.
Moody's Industrial Manual (American and foreign), New York: Moody's Investor Service, 1967.
Moody's Stock Survey, New York: Moody's Investor Service, 1967.
Moody's Transportation Manual, New York: Moody's Investor Service, 1967.
Poor's Register of Corporations (Directors and executives), New York: Standard and Poor's Corporation, 1967.
Standard and Poor's Corporation Records, New York: Standard and Poor's Corporation, 1968.
Standard and Poor's Stock Guide (Security owner's), New York: Standard and Poor's Corporation, 1968.
Thomas Register.

a. quantitative information. Data was collected for each query relating to the number of abstracts identified during a search, the number of abstracts forwarded to the user, the evaluation of the user and the number of documents requested (Table 3). Statistics were available from August 1964 thru January 1968 representing 41 searches.

b. qualitative information. Each engineering consultant was requested to complete an inquiry form (Example 1) for each company with which he has had contact, either personal and/or through reveiw of search results. It is hoped that a weighting technique may be applied to the responses for analysis purposes.

The collected question data has been tabulated by individual companies, and means for analysis are being considered.

Table 2

Public and Financial Information
of Possible Interest

Historical and/or Background Data

- Place of Incorporation
- Date of Incorporation
- Major purchases, mergers, etc., comprising establishment of corporation
- Recent expansions in interest, i.e., new product field, companies acquired, etc.
- Rank as to industries in similar product manufacturing
- Nineteen year scale of earnings and dividends

Current Information

- Executive officers
- Location of executive and/or principle office
- Number of employees
- Number of stockholders
- Principle plants and properties
- Business and products
- Complete list of subsidiaries

Current Financial Information

- Wages, salaries and employee benefits
- Complete financial report
- Financial statistics
  - Gross revenues
  - Operation profits (margin percentage)
  - Net income
  - Working capital
  - Senior capital
  - Number of stock shares outstanding
  - Cash flow per share
  - Earnings per share
  - Dividends per share
  - Average yield
- Dividends record
- Estimated increase or decrease per share earnings for next year; i.e., 1967 per share earnings and 1968 estimated per share earnings
- Percentage dip or rise of per share earnings; i.e., 1968 vs. 1967 vs. 1966

Table 3

Quantitative Information from KASC Files of Possible Interest

Retrieval Information

1. Number of abstracts identified during computer search

   American International Aerospace Abstracts
   National Aeronautic and Space Administration Abstract Bulletin

2. Number of abstracts identified during analyst's review

   American International Aerospace Abstracts
   National Aeronautic and Space Administration Abstract Bulletin

3. Number of other types of information identified during analyst's review

   Tech Briefs
   Government sponsored reports
   Department of Defense oriented documents
   Nuclear Science Abstracts

Information Forwarded to User

1. Number of abstracts identified during computer search

   American International Aerospace Abstracts
   National Aeronautic and Space Administration Abstract Bulletin

2. Number of abstracts identified during analyst's review

   American International Aerospace Abstracts
   National Aeronautic and Space Administration Abstract Bulletin

3. Number of other types of information identified during analyst's review

   Tech Briefs
   Government sponsored reports
   Department of Defense oriented documents
   Nuclear Science Abstracts

Evaluation of Information by User

1. Related
   American International Aerospace Abstracts
   National Aeronautic and Space Administration Abstract Bulletin
   Other

2. Not related

American International Aerospace Abstracts
National Aeronautic and Space Administration Abstract Bulletin
Other

3. Peripheral

American International Aerospace Abstracts
National Aeronautic and Space Administration Abstract Bulletin
Other

Documents Requested

1. Hardcopy

American International Aerospace Abstracts
National Aeronautic and Space Administration Abstract Bulletin

2. Microfiche

American International Aerospace Abstracts
National Aeronautic and Space Administration Abstract Bulletin

3. Extra copies either hardcopy or microfiche of

National Aeronautic and Space Administration Abstract Bulletin

Example 1

NASA/RDC Consultant Questionnaire
Knowledge Availability Systems Center
University of Pittsburgh
Pittsburgh, Pennsylvania

Name of consultant: _____

Name of company: _____

Dates of active participation of company: _____

1. Did you participate in the initial presentation to company?
   Check one. Yes_____ No_____

2. Were you involved in creation of initial questions of this company?
   Check one. Yes_____ No_____

3. At time of your first contact with the company, did you consider it to be a good_____ or bad_____(please check one) prospect for the Pitt/NASA/RDC service? Why did you consider it so, e.g., knowledge of file, needs of company, etc.?

   _____

   _____

4. When reviewing the abstracts identified during the computer search as answers to the participant's queries, did you believe that they (check one for each of the following):
   a. were really related to the question?
      Yes_____ No_____
   b. represented a full review of the literature in relation to the question?
      Yes_____ No_____
   c. represented a good sampling of the literature in relation to the question?
      Yes_____ No_____

5. How many personal (telephone or visit) contacts would you estimate you had with this company? _____

6. If company is no longer a participant, did you suspect that they would drop? Check one. Yes_____ No_____
   Why did you suspect that they would drop, e.g., poor service, information file not adequate, change of management in company, etc.?

   _____

   _____

   _____

If your answer to the previous was "yes", did you inform anyone at KAS about your suspicion? Check one:

Yes _____ No _____

7. If company is still a participant, why do you think they have continued, e.g., good service, adequate information in file, good personal relations, etc.? _____

_____

_____

Do you believe that their contract for the next year will be renewed? Check one: Yes _____ No _____