DOCUMENT RESUME

ED 121 302 IR 003 280

AUTHOR Schipma, Peter B.
TITLE Research Toward Enhancing Retrieval Effectiveness.
SPONS AGENCY National Science Foundation, Washington, D.C. Office of Science Information Services.
PUB DATE 9 Mar 76
NOTE 13p.; Paper presented at the Annual Meeting of the National Federation of Abstracting and Indexing Services (Columbus, Ohio, March 9, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS Bibliographic Coupling; Bibliographies; Computer Programs; Cost Effectiveness; Data Bases; *Indexing; *Information Retrieval; *Information Systems; Library Automation; On Line Systems; *Relevance (Information Retrieval); *Search Strategies; Thesauri; Vocabulary
IDENTIFIERS Computerized Searches; ITT Research Institute

ABSTRACT

Two programs addressed the development of information retrieval systems useful to a variety of users. One program was designed to examine aspects of indexing and information display. Experiments with different indexing and display systems revealed that for good relevance judgements, a display with full citation, keyboards, and abstract is necessary. However, the cost of searching abstracts is high, and adequate retrieval can be made using titles and keyboards. The second program was undertaken to improve computer-search quality and efficiency for large bibliographic files. An algorithm was generated whereby citations are clustered for presentation to the user. Preliminary findings indicated that a term map is required which would project found vocabulary up the hierarchy toward greater generality. (CH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED121302

RESEARCH TOWARD ENHANCING RETRIEVAL EFFECTIVENESS

PETER B. SCHIPMA, IIT RESEARCH INSTITUTE

(PRESENTED AT NFAIS ANNUAL MEETING, 9 MARCH 1976
COLUMBUS, OHIO)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CS003880

This Research was Reported by
NSF Office of Science Information Service

INTRODUCTION

IIT Research Institute has conducted several research programs under contracts and grants from the National Science Foundation Office of Science Information Services. While the earlier programs (1968-1972 period) were primarily concerned with developing viable information retrieval systems, recent work has concentrated on improvement in the quality and cost-effectiveness of retrieval. These programs have been aimed toward the discovery of information useful to a variety of systems and system operators. I shall discuss two of these projects briefly today. One has been completed, and detailed information is available in the Final Report. The other program is current - a semi-annual report is now available, and a Final Report will be distributed at the end of this year.

STUDY OF INDEXING AND INFORMATION DISPLAY

The first project was conducted from November 1973 to April 1975.

Program Goals

Studies of indexing and display have been performed for many years. They have ranged from studies of the value of content indexing as opposed to citation indexing to studies of automated indexing. Many of these failed to be definitive because no user community had available one large data bank containing all of the variables to be tested.

Using the facilities of the Computer Search Center (CSC), IIT Research Institute (IITRI) was able to overcome some of these limitations. The recently completed program was possible because of:

- the existence of large machine-readable bibliographic record files representing the same set of documents in different but related ways*

*For this program, a bibliographic record is comprised of citation information (author and location, title, source (e.g. journal), source subelements (e.g. volume, issue, pagination, date)), plus additional information such as abstracts, index terms, molecular formulae, etc.

- the existence of software capable of performing identical manipulations on each of the above,
- the existence of well-tested user questions, and
- the existence and availability of people familiar with the data base(s) and software system.

Data bases used in this study were the Chemical Abstracts Services Condensates, CASIA and CBAC - all converted to IITRI-format for searching purposes.

Program Strategy

This was a three-experiment program designed to quantify several aspects of indexing and information display.

1. Effect of Record Completeness on Relevancy Judgement

   This experiment tested the effects of completeness of records on retrieval efficiency - the same set of questions was used to search several versions of the same data base. These versions differed in the amount of material available for searching (titles only, plus index terms, etc.)

II. Effect of Indexing Methodology on Retrieval Efficiency

   This experiment studied several indexing methods for their effects upon retrieval efficiency.

III. Effect of Information Display on Relevancy Judgement

   This experiment tested the effects of completeness of display on relevance judgement. The total set of retrieved documents, found by all search methods, was judged for relevance. Only certain parts of the records were given to each reviewer, to determine the effects of record completeness on their judgement of relevancy.

Program Findings

Experiment #1 - Effect of Record Completeness on Relevancy Judgement

Three levels of record completeness (citation information only,

citation information plus keywords, and citation information plus keywords and abstract) were searched.

This experiment showed that the more complete a record, the more likely is its selection as a hit. The overall average selection for citation information was 31.68%; for citation information plus keywords was 54.50%; and for citation information plus keywords and abstracts was 76.29%. Adding keywords increased the likelihood of getting a hit by 23% while adding the abstracts also increased the likelihood of a hit by another 22%. The addition of all the other elements (index terms, Registry Numbers, molecular formula, etc.) added another 24%.

While CSC costs are not absolutely congruous to those of others, the data in Figure 1 indicate the expected increase in cost with increase in the size and complexity of a data base. Keywords return quite a bit in performance for little incremental cost. The searching of full abstracts provides a similar increase in performance at a considerable increase in cost. The dollar figures are normalized to $100.00.

Figure 1

	CITATION INFORMATION	CITATION INFO. + KEYWORDS	CITATION, KEYWORDS, + ABSTRACT	FULL RECORD
PERFORMANCE LEVEL	31.7%	54.5%	76.3%	100%
COST	$36.00	$43.00	$74.00	$100.00

OVERALL PERFORMANCE/COST FOR SEARCHES OF FILES AT GIVEN LEVELS OF RECORD COMPLETENESS (USING CSC COST FIGURES).

Experiment #2 - Effect of Indexing Methodology on Retrieval Efficiency

This experiment related indexing methods to retrieval efficiency. The major portion of the work was carried out on two files. The CA Condensates file was used to represent unstructured, uncontrolled indexing and the CASIA file (the time-ordered version of the CAS Integrated Subject File) was used to represent controlled unstructured indexing.

To obtain a baseline measure for appearance of terms in given data elements, both the Condensates and the CASIA files were searched for single terms. The percent found uniquely for each data element was recorded for each term. Data for this study are given in Figure 2.

One year of both Condensates and CASIA were searched. It was shown that the closer a term was to a chemical name, the more often it was found in CASIA. On an overall average, for this sample of terms the Condensates keywords were the more discriminatory field.

One very important fact emerged. CASIA did not replace citation data and keywords for search purposes. This confirmed previous findings by both IITRI and the University of Georgia in their studies of the CAS Integrated Subject File (ISF). While subject indexing was of benefit when searching chemical names, it was poorer than citation and keyword information for searching subject concepts.

Searches were made of the CASIA file for the questions previously searched against the other versions of the data base. Two important facts were obvious from the results:

- many more records were extracted from the data base, which had not been found via searches of the citation and keyword data, and
- very few of the records identified via citation and keyword searches were the same as those identified via the index search.

While a total of 7128 records had been identified by searches of all versions of the citation and keyword information, the searches of the index file (CASIA) identified 4988. But only 828 of these 4988 were common with those contained in the 7128. Thus, there were actually 11,288 identified records from the sum of the searches. In point of fact, searches of citation and keyword data alone perform considerably better (63.15%) than those of index data alone (44.19%). To insure complete retrieval, both are required.

TERM	CONDENSATES % PRESENT			CASIA	CONDENSATES % UNIQUELY PRESENT			CASIA % UNIQUELY PRESENT
	TITLE	KEYWORD	ABSTRACT		TITLE	KEYWORD	ABSTRACT	
Prostaglandin	32.54	32.92	15.84	18.58	6.73	1.92	2.24	2.24
Penicillin	11.03	29.31	33.45	26.03	0.95	0.48	11.90	1.43
Norepinephrin*	9.32	20.80	30.33	39.45	1.32	4.42	7.51	26.27
Dopa	16.33	22.55	32.50	28.54	1.16	0.93	10.23	2.79
Teratogen*	5.32	57.79	36.50	0.00	0.00	22.00	19.33	0.00
L-Dopa*	22.34	38.83	38.46	0.00	2.50	9.17	5.83	0.00
*Oroxyphenylalanine	36.36	6.06	18.18	36.36	45.45	4.55	0.00	22.73
Neurotransmitt*	23.21	42.86	32.14	0.00	15.38	33.33	23.08	0.00
Biogenic amine*	35.96	38.20	24.72	0.00	22.38	24.19	17.74	0.00
Nitroso	11.61	16.08	23.08	49.09	0.91	0.00	3.64	20.45
MEAN:	20.40	30.54	28.52	19.81	9.68	10.10	10.15	7.59

* Indicating truncation

Figure 2

Experiment #3 - Effect of Information Display on Relevancy Judgement

The output from all the search types conducted in Experiment #1 was summed. Each was printed in three formats:

- titles and citations only
- titles, citations and keywords, and
- titles, citations, keywords, index terms and abstracts

Printouts were distributed to IITRI scientists in such a way that each display mode for each profile was evaluated by a different scientist.

The results were consistent. Two profiles had no hits and were discounted. Of the remaining 21 profiles, 15 showed one pattern and six showed another. The most common pattern, obtained in 15 of the 21 profiles, was that the Titles Only display mode gave the poorest Recall* and highest Noise** while the Title Plus Keyword display mode gave better Recall and less Noise. The All Fields display mode, by definition, had total Recall and no Noise. In general, however, the Title Only and Title Plus Keyword display modes were fairly similar and poor in relationship to All Fields display mode. This strongly indicates the need for full index terms and abstracts in display of records to assure good relevance judgement.

The other six profiles, while also showing relatively poor performance by both the Title Only and the Title Plus Keyword display modes, showed a seeming anomaly in that the Title Only display mode resulted in better relevancy judgements than the Title Plus Keyword display mode. Analysis of the profiles provided the answer. They were similar in that Title Only left a number of ambiguous cases, so those were selected. The keywords, here, worked only in a negative sense. They removed some ambiguous cases, but didn't add more specific records.

*Recall is a measure of the degree of potential performance (relevant records selected).

*Noise is a measure of confusion (irrelevant records erroneously selected).

Experiment #3 strongly indicates the need for full abstracts as well as citation information, titles, keywords and index terms as display items. While keywords improve titles somewhat, a large percentage of relevant records will be missed if abstracts and index terms are not present in the display.

Major Program Implication

- For efficient retrieval evaluation, a display including full citation, keywords and abstract is necessary.
- However, good search results can be obtained from a system with titles and keywords available for searching. The addition of abstracts to the searching field, while increasing the search capability some, greatly increases data base preparation, up-date and manipulation costs.
- A cost-effective, efficient search system should have a capability to search titles and keywords, combined with a display capability including full citations, keywords and abstracts.
- Introduction of the CASIA file for on-line searching would offer a valuable tool to the chemical research community. It may be possible to extend this state ment to indicate that index information in general (for any data base) will enhance the utility thereof, but the data were only obtained for a chemical data base.

ENHANCING THE RETRIEVAL EFFECTIVENESS OF LARGE INFORMATION SYSTEMS

This project was begun in June of 1975 and is scheduled for completion by November of this year.

Program Goal

The research goal is to improve computer-search quality/efficiency for bibliographic files. The original proposal emphasized a two-step process:

1) A standard Boolean (or other)search of high recall resulting in a large initial retrieved set (RI).
2) A cluster analysis of RI to sort records into categories so as to reduce user evaluation time without sacrificing quality (precision and recall).

Prior to this grant, IITRI had written several clustering programs incorporating unique criteria for term associations. Initial goals of the grant were to test these programs in a statistically meaningful manner using Chemical Abstracts and Engineering Index. Initial results indicated that the disparity between machine and human relevancy judgements contains two large factors that are amenable to machine solution at a level less complex than syntax analysis. The two factors are:

- Term synonyms (several terms with similar meanings)
- Term ambiguity (one term with different meanings depending on context)

These two factors limit search quality for Boolean and for clustering methods. For the former, the user must try to specify all synonyms in the original profile - which is a task of diminishing returns since some of the synonyms will occur only at very low frequency. For clustering, the algorithm attempts to identify synonyms based on the occurrence patterns of words. While it does work, it is also clear that it cannot be perfect because the occurrence patterns of words do not contain enough information, for the small retrievals involved, to define synonyms precisely. Thus, two different means of overcoming these two word definition problems are being explored. These methods may be incorporated into either clustered or non-clustered retrievals:

- Have the user evaluate a sorted list of terms derived from RI (as a part of a standard search)
- Construct a term map so that synonyms and ambiguities may be automatically simplified.

Both of these methods may prove to be compatible with an on-line environment.

It seems probable that the future of information retrieval from bibliographic files lies in the direction such that machines will more closely approximate the processes that occur in the mind of the manual searcher. Historically, the progression has been:

- look for the occurrence of a list of words
- look for combinations of words from a list

- group citations according to combinations of words and present the groups to the user (standard clustering)
- group citations obtained by test against a list of words according to the combinations and present them to the user (IITRI Algorithm)

Our current activity adds to this list:

- group citations obtained by a test against a list of words according to the combinations, taking into account synonyms and ambiguity.

This last step is in the direction of syntax and meaning because it involves enabling the computer to work with definitions. It is related to automatic indexing and may provide a mechanism for automatic or assisted profile generation. As the cost of computer storage and operation continues to fall, relative to other costs, the number of operations per search that are economic, rises, and it seems but a matter of time until the computer operates at a level of syntax/meaning.

Current Activities

Testing of the Clustering Algorithm

On the basis of some preliminary clustering runs against Engineering Index and Chemical Abstracts criteria were established for evaluation of the Algorithm. That is, what kind of profiles (Boolean terms and logic) and retrievals should be used to test the sensitivity of the Algorithm to jargon, relatedness of concepts and retrieval size. These decisions have been largely completed.

Characteristics of the Initial Retrieved Set - RI

While IITRI has studied term frequencies for whole data bases, it has not previously studied the distribution of the vocabulary within the set RI, the initial retrieval. It was expected that the distribution of the "found" vocabulary would be very different from the vocabulary of the whole data base that would have high relative frequencies for terms related to the search terms. Thus, programs were run to generate some sample distributions from Chemical Abstracts and Engineering Index retrievals. The results showed that the relative frequencies were too low for direct user evaluation and that an intermediate mapping or grouping is required.

Vocabulary Decomposition

In an effort to design a module to enable the user to make intermediate vocabulary judgements (i.e. evaluate the found vocabulary) it is desirable to know whether there are any simple rules that distinguish the key words from the others. In an effort to characterize those words, we have manually analyzed a retrieval set and isolated the minimum set of words on which an accurate relevancy judgement could be made. We are currently examining that vocabulary in detail.

Planned Activities

Preliminary findings indicate that what is required is a term map - constructed manually on the basis of meaning, that can map a specific term such as "gimbals" back to the level of "navigation". That is, the map would project the found vocabulary up the hierarchy towards greater generality. At the more general levels, term frequencies would be expected to be greater, so that the number of user evaluations that would be required would be relatively smaller.

Another fact of the word map/projection process is that it would allow the found vocabulary to be sorted to that link to which it is relevant. That is, suppose in an A & B type search, the A terms are plants and the B term are air pollutants. The program may find "Tree" in the found vocabulary and it could then associate it via the map with the A link.

If the word map and the term link assignment are available, the scenario of a search would then be as follows:

1. User specifies links and logic (example A & B)
2. Computer finds initial retrieved set (RI)
3. Computer finds found vocabulary of RI (example Tree)
4. Computer uses word map to reduce found vocabulary to an appropriate level of generality
5. Computer groups found vocabulary according to links
6. User specifies link to be expanded (example - breakdown by plant terms and keep all air pollutant terms).
7. Computer prints out a list of the retrieved sets (Rn) to be obtained for each of the examples of found vocabulary associated with the A link.

1) Tree 167

2) Bush 202

(etc.) (etc.)

8) User specifies which of the Rn subsets he wishes to obtain and has those printed.

This scenario obviates many of the problems. The synonym problem is handled explicitly by the word map. Term ambiguity may be handled by building limited associations into the map, the key questions now are:

1. Are our preliminary results of general validity?
2. Can the required file access and the computations be be done in times compatible with an on-line environment?
3. How expensive would it be to construct a functional word map?

We will continue work toward definitive answers to these questions.