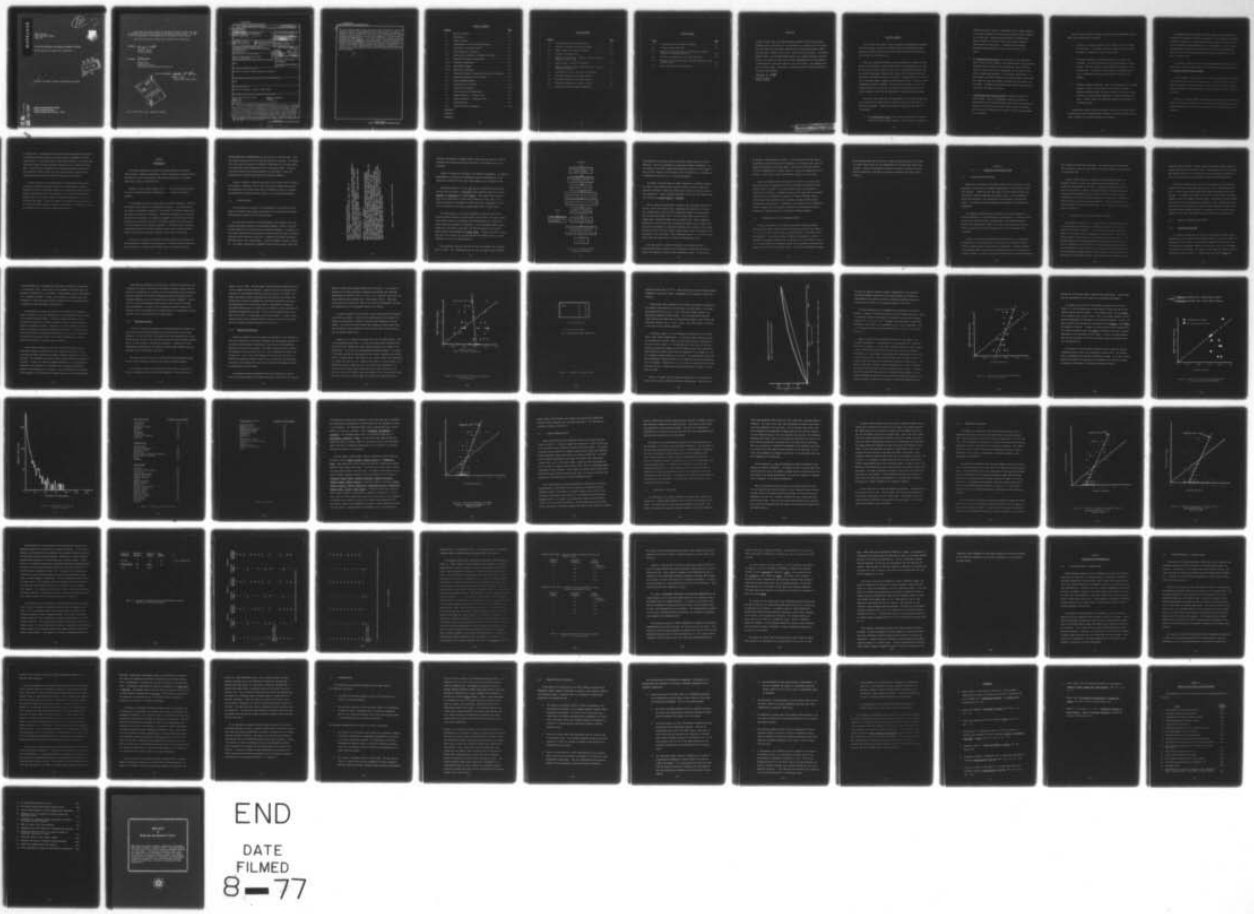
AD-A041 845 PATTERN ANALYSIS AND RECOGNITION CORP ROME N Y F/G 5/2
STATISTICAL METHODS FOR TECHNICAL DOCUMENT RETRIEVAL.(U)
JUN 77 H M HERSH, J M MORRIS, K J MORRIS F30602-76-C-0135
UNCLASSIFIED PAR-77-8 RADC-TR-77-217 NL

1 OF 1
AD A041 845



END

DATE
FILMED
8-77

12 2

RADC-TR-77-217
Final Technical Report
June 1977

STATISTICAL METHODS FOR TECHNICAL DOCUMENT RETRIEVAL

Pattern Analysis & Recognition Corporation

Approved for public release; distribution unlimited.

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed by the RADC Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public including foreign nations.

This report has been reviewed and is approved for publication.

APPROVED: George D Petit

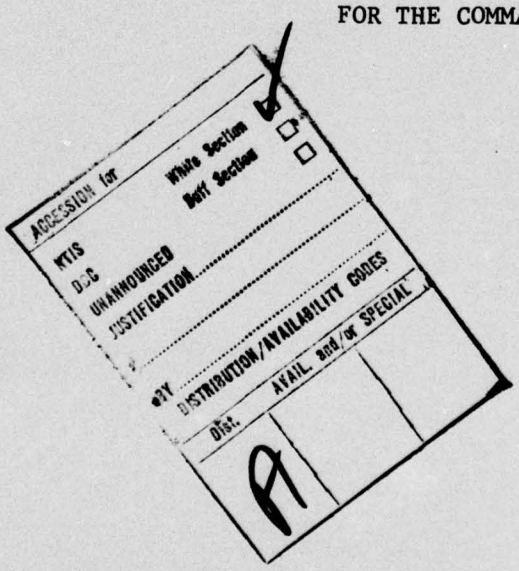
GEORGE D. PETIT
Project Engineer

APPROVED: H Davis

HOWARD DAVIS
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office



Do not return this copy. Retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: RADC-TR-77-217	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): STATISTICAL METHODS FOR TECHNICAL DOCUMENT RETRIEVAL.	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report.
6. PERFORMING ORG. REPORT NUMBER: PAR [redacted] 77-8	
7. AUTHOR(s): Dr. Harry M. Hersh, Miss Katherine J. Morris; Dr. John M. Morris, Mr. James R. Wilson	8. CONTRACT OR GRANT NUMBER(s): F30602-76-C-0135 new
9. PERFORMING ORGANIZATION NAME AND ADDRESS: PATTERN ANALYSIS AND RECOGNITION CORPORATION, 228 Liberty Plaza, Rome NY 13440	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62702F; 45940115
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRDT), Griffiss AFB NY 13441	12. REPORT DATE: June 1977 (17701)
	13. NUMBER OF PAGES: 87 (12 85 p.)
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: Same	15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: George D. Petit (IRDT)

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Associative Retrieval System
Clustering
Identity Associations
Stemming

This report describes experiments which used the

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Statistical methods for document retrieval include techniques for automatic generation of a dictionary, development of a thesaurus, and analysis of the data base through the use of correlations among selected words. The RADC Automatic Document Classification On-Line (RADCOL) system is a tool for testing various statistical procedures for document analysis and retrieval, and for the design of operational systems. ~~In the experiments reported here, which utilized the~~ RADCOL system; it was found, as had been predicted, that procedures for clustering word stems did not provide substantial savings in space and time, and → (cont on pg ii)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

- i - 390 101

(cont to R.i)

that an unclustered thesaurus gave improved retrieval capabilities. Three new versions of the system were implemented, with weights of 0.0, 0.5, and 1.0 assigned to identity correlations (correlations of wordstems with themselves). Because of superior performance of the system using 1.0 correlations, a simplified version of the retrieval technique was recommended for use with Science and Technology ~~(S&T)~~ abstracts. In the simplified system, automatic thesaurus generation would be eliminated, and a large technical vocabulary would be used. Retrievals would ~~utilize~~ direct correlations between queries and documents. The experiments ~~reported here~~ are believed to be the most comprehensive series of tests of statistical retrieval methods ~~ever~~ performed on a data base of realistic size. Further experimentation is recommended to determine the applicability of statistical methods to other types of intelligence data bases and user requirements.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EVALUATION

A study has been done on the RADC Automatic Document Classification On-Line (RADCOL) system to determine its effectiveness as a retrieval tool for Science and Technology (S&T) data bases, specifically the Central Information, Reference and Control (CIRC) system at FTD. Three versions were tested: The initial RADCOL system, an unclustered RADCOL system, and a weighted document retrieval system. The outcome of these tests has been recommendations for the design of a simplified more rapid system based on the third version: the weighted document retrieval system. This effort is included as part of TPO Thrust R3D, Intelligence Data Handling.

George O. Petit

GEORGE D. PETIT
Project Engineer

EXECUTIVE SUMMARY

Over the past three years, Pattern Analysis and Recognition Corporation (PAR) has performed a series of experiments and tests employing the RADC Automatic Document Classification On-Line (RADCOL) system, which was developed for RADC by Informatics, Inc.

RADCOL was intended specifically for the retrieval of Science and Technology (S&T) abstracts in subject-matter areas to be designated by the user. For the purpose of these tests, a data base of 4000 abstracts selected from the Central Information, Reference and Control (CIRC) collection was used. The initial implementation was in the JOVIAL J3 language under GCOS on RADC's HIS 635 computer system. Because of difficulties with the J3 compiler, PAR's programming was done in FORTRAN. The HIS 635 was upgraded to an HIS 6080 during the course of the project, but this had no substantial effect on system operation, other than an improvement in timing and responsiveness.

The goal of this effort was the development of justified recommendations for the design of an operational S&T document retrieval system based on statistical analysis. Three major versions of the system were tested for this purpose:

1. The initial RADCOL system, which employed routines for clustering word-stems around "concept centers". This approach is intended to

identify clusters or groups of words which tend to appear together, in the same documents, and which serve to identify topics contained in the data base. Retrievals are performed by determining the topics contained in the user's query, and locating documents which contain the same topics. A weighting scheme permits evaluating these documents as more or less relevant to the query.

2. The unclustered RADCOL system was developed to test suggestions made in PAR's initial proposal, which indicated that the clustering procedures used by RADCOL could be eliminated, with a resulting saving in start-up time and no substantial loss in terms of additional storage space required, or in retrieval effectiveness. In this system, an automatic thesaurus is generated which contains the four words which are most highly correlated with each word in the dictionary. Documents are retrieved which contain words which are correlated with words in the query.

3. A weighted document retrieval system was simulated, using the RADCOL system. This is a considerably simplified approach, in which documents are retrieved whenever they contain words which are identical to the words in the query (after removal of suffixes), and are weighted according to the number of times that they appear in the document.

Several other statistical methods and functional characteristics of the RADCOL system were tested, including:

1. Routines for removing suffixes to form "stems" for use in forming the dictionary. For example, endings are removed from the words COMPUTER and COMPUTING to form the stem COMPUT.

2. Statistical measures for selection of stems to be used in the dictionary. The RADCOL system used the "Dennis measure" for this purpose. This is a statistic which estimates the value of a stem for retrieval. For example, COMPUT would be a valuable stem for use in a technical data base, while AND and THE would not be valuable.

3. Relevance feedback techniques. When a set of documents is retrieved, some will be more or less relevant to the needs of the user. Feedback techniques permit the user to improve the quality of the retrieval by giving heavier weightings to relevant words and concepts. Several methods for improving retrievals are included in the RADCOL system.

Experimental methods included operational testing of the RADCOL system, together with studies comparing human evaluation of relevance against statistical estimates of relevance developed by the system.

The RADCOL system has provided a useful facility for the testing and evaluation of statistical retrieval techniques. It was possible to implement a variety of methods, and to test these against a large data base, with the goal of developing design parameters for an operational S&T document system. The experiments reported here are believed to be the most comprehensive series of tests of statistical retrieval methods ever performed on a data base of realistic size.

The outcome of these tests has been recommendations for the design of a simplified, more rapid system based on the third methodology described above, the weighted document retrieval system.

In this system, the data base would be searched for documents containing stems identical with those in the user's query. Documents would be "correlated" with queries in the sense that they would receive a higher weighting when they contained a greater number of stems which were the same as those in the query.

Because of the large number of technical terms contained in the S&T data base, such a system appeared to be more effective than one which attempted to locate general topics or subject areas through the automatic generation of a thesaurus.

The following versions of the RADCOL system were used in these experiments:

System 1. S&T data base consisting of 4000 abstracts in representative subject areas from FTD's CIRC data base. This was the version as originally delivered by Informatics, Inc., to RADC, with some minor modifications made by PAR.

System 2. Another version of System 1, as modified by PAR, in which the data base consisted of 1800 Indications and Warnings (I&W) messages.

System 3. Similar to System 1, but with function words (THE, AND, etc.) and some obvious misspellings manually removed from the dictionary.

Section 4. A modification of System 3, with clustering algorithms removed. Because of problems in data storage this version did not produce usable results.

Version A. An experimental system with clustering algorithms removed and with batch entry (to avoid problems encountered in attempting to modify the interactive portions of the system). Weights (correlations) for stems with themselves (the identity correlations) were set to 1.0.

Version B. Same as Version A, with identity correlations set to 0.5.

Version C. Same as Version A, with identity correlations set to 0.0.

Another (unnumbered) version simulated the Weighted Document Retrieval System by including only the identity correlations of Version A.

On the basis of experimentation reported here, it is recommended that a weighted document retrieval system be implemented for S&T abstract analysis and retrieval. Such a system would include facilities for Boolean retrievals (i.e., retrievals containing combinations of stems utilizing the Boolean connectives AND, OR, and NOT), and for weighted retrievals as described above. The ability to search for titles, authors, and topic tags, as well as other information contained in CIRC document headers, should be included.

Of equal importance in the development of a usable system design is a careful survey of user needs. For that reason, it is recommended that a study of the needs of current and prospective S&T data system users be included in any future development of retrieval system designs.

The statistical methods used by the RADCOL system have not been widely tested on other data bases of significant size. For this reason, care should be taken in extending the recommendations of this study to other data bases and user environments. The S&T data base used here, with a high proportion of specialized scientific and technical terminology, did not produce statistical associations which improved the quality of retrievals.

Other data bases, such as those used in Indications and Warnings (I&W) intelligence analysis, are different in character from the S&T data base which was used for this study, since they do not employ a technical vocabulary

of similar size. Concurrent work with I&W data bases suggests that statistical methods may prove valuable for locating groups of messages in related subject areas. In an earlier study, using RADCOL software, it was found that statistical methods provided effective retrievals from a simulated I&W data base. Further work in development of the Message Extraction Through Estimated Relevance (METER) system has shown that statistical methods can be implemented within the time constraints required by I&W analysis.

The next step will be the development of an operational METER system, employing a substantial data base derived from actual message traffic, to determine the type of processing required by a statistically-based system for maximal effectiveness in the I&W environment. The RADCOL experimentation reported here will provide a number of alternative approaches to be used in the METER development. During the design, implementation, and development of the operational METER system, alternative approaches will be tested to provide an effective tool for the use of intelligence analysts.

SECTION 1

INTRODUCTION

This report presents the results of continued testing of the RADCOL (RADC Automatic Document Classification On-Line) associative retrieval system. The results of previous statistical and operational testing of the RADCOL system can be found in RADC-TR-75-208.

Section 1 of this report presents an overview of the system, with emphasis on its operation when the clustering of content stems is included or excluded.

The implementation of the system using the Central Information, Reference and Control (CIRC) data base is discussed in Section 2. Also presented are the results of operational testing of two versions of the RADCOL system (with and without content stem clustering). Queries were presented to the systems, abstracts were retrieved, and subjective judgments of the relevance of these abstracts were evaluated. The use of clustering was found to be detrimental to the performance of the retrieval operation. In addition, basic questions concerning associative retrieval technology arose, which required answering before an operational associative retrieval system could be developed.

In Section 3, testing of associative retrieval methods was continued to further assess the viability of this approach to automatic document retrieval. Major portions of the system were rewritten here to eliminate code which

malfunctioned due to modifications to the system or to the data base. Also, the topical structure of the data base was objectively analyzed. This permitted a more precise evaluation of retrieval effectiveness by allowing queries to be directed toward prespecified areas in the document space. Abstracts were then retrieved using queries generated in this manner. Again the abstracts were evaluated using subjective relevance judgments.

Section 4 presents a detailed discussion of the current state of associative retrieval technology and of further testing and possible modifications and extensions to the system which might be desirable in order to implement an effective and efficient retrieval system in an operational environment.

1.1. SYSTEM OVERVIEW

In this section an overview of the methodology of the RADCOL retrieval system is presented, with specific application to the Science and Technology (S&T) abstracts selected from the CIRC data base.

The CIRC subset consists of 32,000 abstracts of S&T documents from the Foreign Technology Division, Air Force Systems Command. However, only 4000 of these abstracts were actually used for retrieval purposes. An analysis of this subset indicated that the abstracts covered a very wide range of subjects, although certain substantive areas (e.g., computers) were better represented than other areas (e.g., lake ecology). A typical abstract is shown in Figure 1-1. The header, consisting of information such as title, author, topic tags

AAN11558850009 05$021270 $YUR*FD C
TITLE AUTOMATIC PROGRAMMING MACHINE, RIGA, USSR -U-
AUTHOR NONE
TOPIC TAGS -UR- COMPUTER APPLICATION, PUNCHED PAPER TAPE, INFORMATION PROCESSING, COMPUTER INPUT UNIT, COMPUTER OUTPUT UNIT, AUTOMATIC CONTROL SYSTEM/[U]ERPA ALPHANUMERIC PRINTER
DOCNR UR 9024 70 000 000 0000-0000 SOURCE STROITEL'NAYA GAZETA SOVIET TRI WEEKLY 2 DEC. 70

ABSTRACT [U] GP-O- ABSTRACT.
THE EPRA AUTOMATIC ELECTRONIC PROGRAMMING AND WRITING MACHINE WAS CONSTRUCTED BY THE RIGA CENTRAL PROJECTING AND ENGINEERING OFFICE FOR MECHANIZATION AND AUTOMATION. THE MACHINE IS INTENDED FOR THE STORAGE OF REPEATING TEXTS ON A PERFORATED TAPE AND FOR AUTOMATIC COMPOSITION OF COMMERCIAL CORRESPONDENCE. IT CAN BE USED FOR PREPARING AND TRANSMITTING INFORMATION IN AUTOMATED CONTROL SYSTEMS, FOR PREPARATION AND OUTPUT OF INFORMATION FROM AN ELECTRONIC COMPUTER. IT PRINTS WITH A SPEED OF 600 SIGNS PER MINUTE. THIS INCREASES CONSIDERABLY THE WORKING CAPACITY. THE ERPA CAN CARRY OUT PROOF READING AND EDITING WORK.

Figure 1-1 Typical Abstract From CIRC Data Base

(key words and phrases), document number, source, and date were not used for retrievals but could be printed with the text of the abstract when it was retrieved.

Figure 1-2 presents an overview of the retrieval methodology. As detailed descriptions of the system are contained in previous documentation of the system and in RADC-TR-75-208, only a brief summary will be presented here.

Initially, stopwords (e.g., a, the, of) are removed from the abstracts and each word is reduced to a raw stem by the stemming routine (for example, computer and computation both become comput). There were 58,886 unique raw stems in the CIRC subset, of which many occurred only once. Most of the single occurrences are traceable to either spelling or typographical errors, or to very specific words such as proper names, dates, or model numbers.

The Dennis measure [1] is then calculated for each raw stem, and the stems are ranked according to their value on this measure. The Dennis measure is an index of how well a particular stem discriminates (in a statistical sense) among the documents in a data base. In the implemented RADCOL system, using the 32,000 abstract data base, the 5000 raw stems with the highest Dennis measure were designated as content stems. (Content stems were selected from the 32,000 abstract set, rather than the 4000 abstract subset, to give stability to the set of selected stems.)

All non-content stems are next removed from the abstracts and a concordance is formed. This concordance takes the form of a matrix which indicates

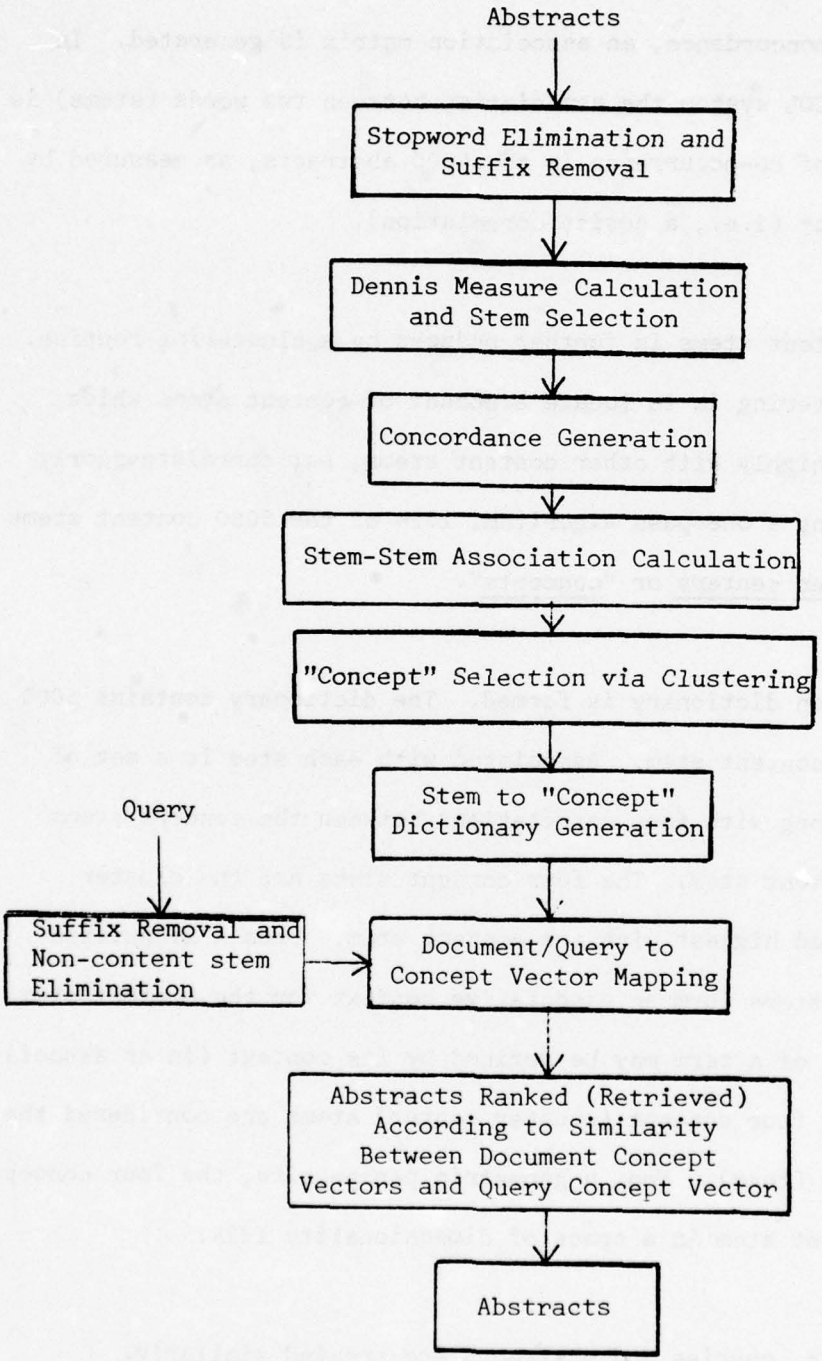


Figure 1-2 Overview of RADCOL Retrieval Methodology

the frequency of occurrence of each of the 5000 content stems in the 4000 abstracts. From this concordance, an association matrix is generated. In the context of the RADCOL system the association between two words (stems) is taken as their degree of co-occurrence in the 4000 abstracts, as measured by a normalized dot product (i.e., a cosine correlation).

The number of content stems is further reduced by a clustering routine. The object of the clustering is to locate a subset of content stems which associate (correlate) highly with other content stems, but correlate poorly among themselves. Using a one-pass algorithm, 1324 of the 5000 content stems were selected as cluster centers or "concepts".

Next an association dictionary is formed. The dictionary contains 5000 entries, one for each content stem. Associated with each stem is a set of four concept stems, along with four correlations between the concept stems and the particular content stem. The four concept stems are the cluster centers which correlated highest with the content stem. From a linguistic perspective, the four stems form an associative context for the content stem and, since the meaning of a term may be defined by its context (in an associationistic theory), the four concept (cluster center) stems are considered the definition of the stem (term). From a geometric perspective, the four concept stems locate the content stem in a space of dimensionality 1324.

From this point on, queries and abstracts are treated similarly. Document/query concept vectors are next generated by replacing the content stems in an abstract or query by their dictionary entries. In other words,

an abstract is replaced by its context. If a stem occurs more than once in an abstract, each occurrence will be replaced by a dictionary entry, and the weights (correlations) are simply added. The result is that each abstract can now be represented by a concept vector in a space of dimensionality 1324. Moreover, query concept vectors simultaneously exist in the same space.

Since the abstracts and the query exist as vectors in the same space, document retrieval is performed by simply retrieving the abstract(s) whose concept vector is closest to the query concept vector (as measured by a normalized dot product). It should be noted at this point that the retrieval operation is not a dichotomous decision as in boolean retrieval systems. Theoretically every abstract can be retrieved, as long as there exists a corresponding concept vector. Retrieval in this context must be considered a matter of degree, and an evaluative function (the normalized dot product) indicates the appropriateness of a particular retrieved abstract.

1.2. INVESTIGATION OF THE CLUSTERING PROCEDURE

Previous evaluations of the RADCOL system raised several questions concerning the benefits derived from the clustering procedure (see Figure 1-2). As part of the current effort, a modified version of the RADCOL system was generated. (The current version of RADCOL is designated as system 3. Therefore, the modified version will be referred to in this report as system 4.) In this system, the clustering routine was circumvented by allowing each content stem to be its own cluster center. The resulting differences between

the two systems were that the latter's concept dictionary used all 5000 stems as cluster center stems, and the dimensionality of the resulting vector space was 5000. Section 2 described the evaluation of these two systems (with and without clustering) on the basis of retrieval performance and overall effectiveness.

SECTION 2

EVALUATION OF THE RADCOL SYSTEM

2.1. SYSTEM OPERATION EVALUATION

Before the two versions of the RADCOL system could be evaluated, it was necessary to verify that the original implementation of the system was operating correctly. Toward this end, each step of the data base generation and the abstract retrieval operation was checked by hand. By this procedure, errors were detected in the calculation of the association (correlation) matrix. The incorrect code was subsequently corrected, and the matrix was regenerated.

The remaining calculations and procedures of the system appeared to function reasonably, if not efficiently and conveniently. It should be noted at this point that the original implementation of the RADCOL system was planned as an experimental system. As a result, many ancilliary features were not included, and the user interface was more of a challenge than a convenience.

Overall, the system was performing as intended, with a few exceptions. An examination of the original code indicated that the relevance feedback feature, though partially coded, was never implemented. Relevance feedback is a procedure where a query vector is redirected in the vector space after a retrieval attempt. The user of the system indicates which retrieved abstracts

were relevant and which were irrelevant. The query vector is then automatically modified so that it points toward the relevant abstracts and away from the irrelevant ones.

Another problem concerned a randomly occurring condition where, for seemingly arbitrary queries, the system would become unresponsive. The system was irrecoverable in this state. Attempts were made to locate the cause of this problem. However, the intractability of the on-line overlay structures, the unreliability of the JOVIAL J3 (WW2.1) compiler, and the peculiarities of the GCOS operating system made it impossible to correct the problem unless the on-line system was rewritten. Thus the problem was endured throughout the initial testing phase. This situation will be addressed again in Section 3.2.

2.2. MODIFICATION OF RADCOL TO ELIMINATE CLUSTERING

As mentioned in Section 1.2., a second RADCOL system (actually, system 4) was produced by eliminating the clustering procedure from the original version of the system (system 3). In the cluster generation program, CLU1, a subset of the content stems is designated as cluster centers. The list of cluster centers is then output to the dictionary generation program. In the modified version, the clustering algorithm is by-passed, and every content stem is considered a cluster center. The dictionary generation program then uses this expanded cluster center list to form new dictionary associations, and the remaining programs generate the new version of the data base in the same manner as the original version. Note that the only modification to the

system was that the subset of content stems which become cluster centers is actually the entire set of content stems. The data bases of the two systems were identical prior to the clustering (or pseudo-clustering) operation; in fact, both systems used the same content stems and association matrix.

It should be mentioned that one of the motivations for the clustering procedure was that storage space would be saved over a nonclustering procedure. In fact, however, the differential storage requirements were insignificant. For example, the number of entries in the association dictionary remained at 5000, the number of content stems. The number of cluster centers associated with each dictionary entry remained at four, although they could now be drawn from the total set of content stems; and since the size of the dictionary entries remained constant, the maximum size of a concept vector remained at four times the number of unique content stems in an abstract.

2.3. TESTING OF RETRIEVAL EFFECTIVENESS

2.3.1. Evaluative Methodology

In a typical paradigm for the testing of information retrieval system performance, a set of queries is initially produced. One or more judges then determine, for each query, a set of relevant documents from the data base. The queries are subsequently submitted to the system, and a set of documents is retrieved for each query. The results of the retrieval operation are then tabulated as shown in Table 2-1. Usually the constructs of recall and

Document Relevant?	Document Retrieved? No	Document Retrieved? Yes	
Yes	A	B	A + B
No	C	D	C + D
	A + C	B + D	

$$\text{Recall} = \frac{B}{A+B}$$

$$\text{Precision} = \frac{B}{B+D}$$

Table 2-1 A Typical Retrieval Evaluation Paradigm

precision are used to summarize the table. Recall is the proportion of relevant documents in the data base that are retrieved. Precision is the proportion of retrieved documents that are actually relevant to the queries (as designated a priori). If the retrieval system is working optimally, cells A and D in Table 2-1 will be empty, and the values of recall and precision will both be 1.0. That is, every relevant document will be retrieved and every irrelevant document will be rejected.

Two points should be noted about this approach to the evaluation of an information retrieval system. First, there are many alternative summary statistics for the 2x2 table. For example, the Φ coefficient (a product-moment correlation for two dichotomies) can summarize the table with a single value whose interpretation is well known. Thus the Φ coefficient might be valuable as a general performance indicator. On the other hand, the measures of recall and precision give more specific information about the functioning of the system; in particular, they can be considered measures of the comprehensiveness and selectivity, respectively, of the retrieval operation.

The second point about evaluating information retrieval systems by a 2x2 table is the assumptions underlying the two variables of interest. In Boolean type information retrieval systems, a document is either retrieved or not retrieved. Retrieval in this case must be considered a true dichotomous variable. Relevance, however, is the result of human judgments. It has been argued elsewhere (e.g., [2], [3], [4]) that natural language concepts such as relevance are not dichotomous nor precise, but inherently vague. That is,

certain documents can be unambiguously classified as relevant or irrelevant to a particular query. But for some of the documents in the data base, the question of relevance will be uncertain. Thus relevance is not a dichotomous, but a continuous variable. Further, it is possible to reliably locate the documents in a data base along a scale of relevance (for a particular query) by ratings or other psychometric techniques [5].

The advantage of allowing the relevance of documents to be evaluated along a continuum is that the performance of the information retrieval system can be measured more accurately. In terms of system performance, it is probably more important to retrieve documents that are clearly relevant as opposed to marginally relevant. Similarly, it is more important to reject clearly irrelevant documents than somewhat relevant ones. A statistic such as the point biserial correlation is sensitive to this type of information, and would give an accurate index of the overall performance of a Boolean system when relevance is measured along a continuum.

As was mentioned previously, the idea of document retrieval being a dichotomous variable is certainly true for a Boolean type system. In an associative retrieval system such as RADCOL, however, retrieval is a matter of degree. The abstracts are output in order of their estimated relevance to the query. In fact, the index of estimated relevance (i.e., query-abstract correlation) can alternately be interpreted as the degree of importance attached to the retrieval of an abstract. Thus the retrieval of an abstract is not measured on a dichotomous nor ordinal scale, but on an interval scale.

Since both the variables in the 2x2 table, retrieval and relevance, can be measured on interval scales for an associative retrieval system, the 2x2 table can be generalized to a scatter plot of query-abstract pairs, where the abscissa represents the retrieval (estimated relevance) continuum and the ordinate corresponds to the (continuous) human relevance judgments. Statistics such as the product-moment correlation can then be used to evaluate system performance. Not only is more information available through this procedure (since the variables are no longer binary, but continuous), but the statistic used is reliable and easily interpretable.

2.3.2. Experimental Design

A set of 25 queries was generated by selecting abstracts at random from the data base and creating questions whose responses would include these abstracts. It was felt that if the systems were functioning properly, these target abstracts would be the first ones retrieved by their respective queries. The set of test queries is contained in Appendix A. (Queries 15-25 were subsequently eliminated from the analysis as their target abstracts were not contained in the 4000 abstract data base.)

Each query was input to both the clustering and nonclustering systems, and the first 64 abstracts retrieved were saved for further analysis.

The relevance judgments were performed by two raters, in order to establish a reliability level for the human judgments. Each rater saw the

queries, one at a time. For each query, raters were also presented with one of the 5 highest retrieved abstracts. Neither the retrieved ranking of the abstract nor its estimated relevance value were known to the raters. The raters were asked to read a query-abstract pair and rate how relevant the abstract was in terms of providing the information requested by the query. The ratings were on a 5 point scale, where a 1 represented a completely irrelevant abstract, and a 5 implied that the abstract exactly contained the information requested by the query. In all, 80 query-abstract pairs were seen by each rater, 45 from system 3 and 35 from system 4. (Several of the original queries never retrieved abstracts from one or both of the systems due to the software problems discussed in Section 2.1.)

2.3.3. Results and Discussion

Before an analysis of the two systems was attempted, it was important to assess the reliability of the raters' judgments, for the agreement between the two human raters necessarily represents an upper bound for the agreement between the raters and the systems. The product-moment correlation between the judgments of the two raters was 0.78. This correlation was highly significant ($df = 78$, $p < .001$), and although it was lower than the levels of inter-rater reliability commonly found in the psychological and educational literature, it did demonstrate that the two raters were assessing relevance in approximately the same manner.

The ratings for each query-abstract pair were averaged over the two raters, and plotted against the estimated relevance calculated by the systems.

Figure 2-1 shows the resulting scatter plot for system 3. As a visual inspection of the figure indicates, relevance and retrieval are completely uncorrelated in system 3 ($r(43) = -0.03$, $p > .10$). Moreover, only one of the target abstracts was retrieved with a rank greater than 64. Since hand calculations had verified that system 3 was correctly implemented, the problem appeared to be in a conceptual feature of the system.

A stepwise analysis of the retrieval results indicated that the clustering routine was suspect. Recall that the associations in the system 3 dictionary are drawn from the subset of content stems designated as cluster centers. But since the cluster centers are themselves content stems, there will be entries in the dictionary for each cluster center stem, as well as each non-cluster center stem.

Figure 2-2 is a schematic dictionary entry for the RADCOL system. STEM is the dictionary entry, and C_1 - C_4 are the cluster center stems that associate (correlate) highest with STEM, ranked according to the weight of the associations. The W_i are the corresponding association weights. The system 3 dictionary entries can be grouped into two classes, depending upon whether STEM is a cluster center. When STEM is not a cluster center, the mean weight, W_i, within this class is 0.16. However, when STEM is a cluster center, C_1 is identically equal to STEM, and thus W_1 is the identity correlation, 1.0. In this case, the mean weight across the cluster center entries is 0.39. That is, a query (or abstract) stem that happens to be a cluster center stem will have over twice as much weight in the resulting concept vector as a

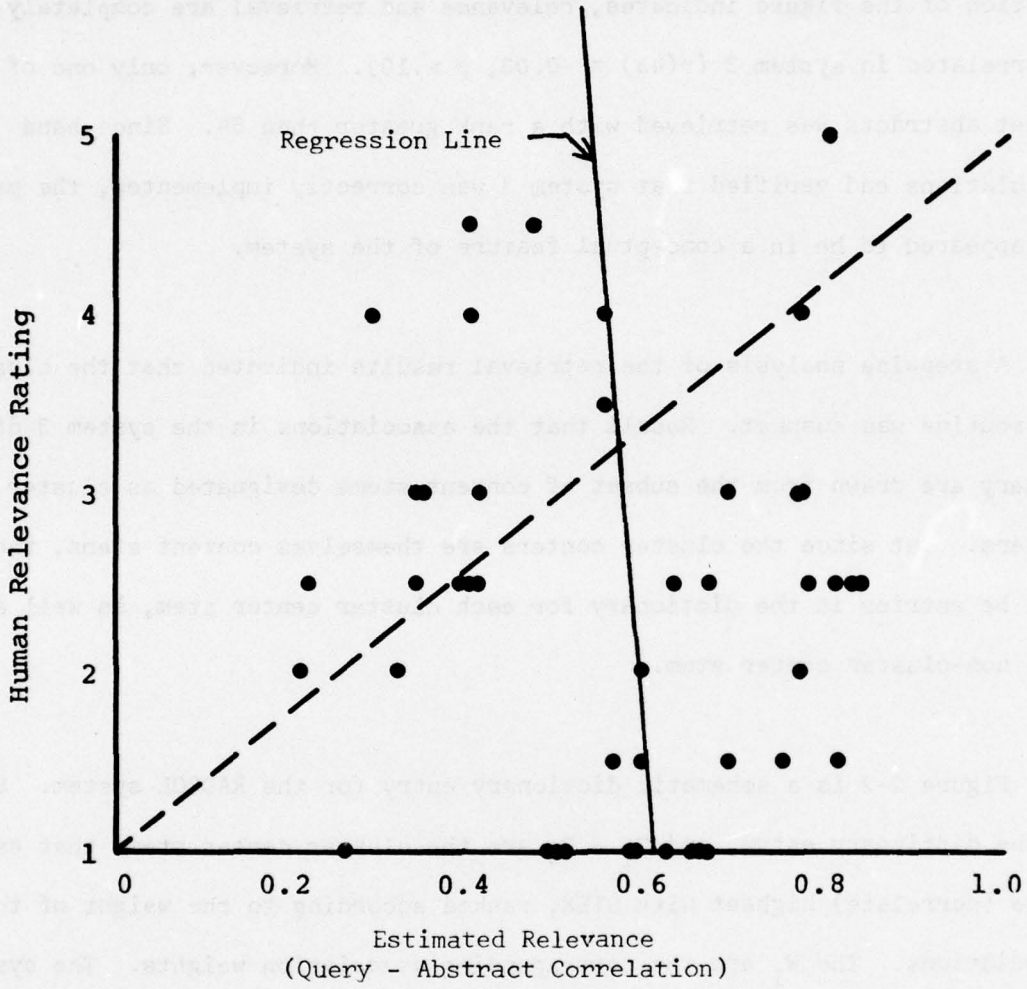
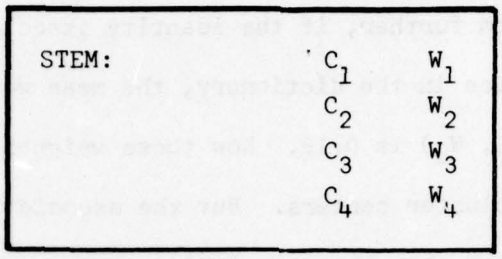


Figure 2-1 RADCOL System 3 Retrieval Performance
($r=-0.03$, $p>0.1$)



$$W_1 \geq W_2 \geq W_3 \geq W_4$$

C_i = Cluster Center Stem

W_i = Association between STEM and C_i

Figure 2-2 Schematic Dictionary Entry

noncluster center stem ($p < 10^{-6}$). These stems will have this disproportionate weighting in the concept vector, independent of the strength of their associations.

Pursuing this notion further, if the identity association is ignored for the cluster center entries in the dictionary, the mean weight of the remaining associations (W_2, W_3, W_4) is 0.18. Now these weights represent the associations among the cluster centers. But the associations among the cluster centers are higher than the associations between cluster centers and noncluster centers (0.16 vs. 0.18, $p < .001$). This relationship is contrary to the goal of any clustering algorithm.

In addition, Figure 2-3 is the cumulative distribution of the cluster center stems, ranked alphabetically. If they were chosen uniformly across the distribution of content stems, the distribution would be represented by the solid line of Figure 2-3. The implication of this figure is that the particular clustering algorithm used in the RADCOL system has a strong bias toward selecting cluster centers from the beginning of the alphabet ($p < .005$ by a Kolmogorov-Smirnov goodness of fit test). One might possibly argue that the distribution of important stems is not necessarily uniformly distributed over the alphabet. Perhaps this is so, but then the content stems would have been selected with a similar bias, and the distribution in Figure 2-3 would still be linear.

Overall, it appears that the clustering algorithm is not working as it should, and the result diminishes retrieval effectiveness. This is not to

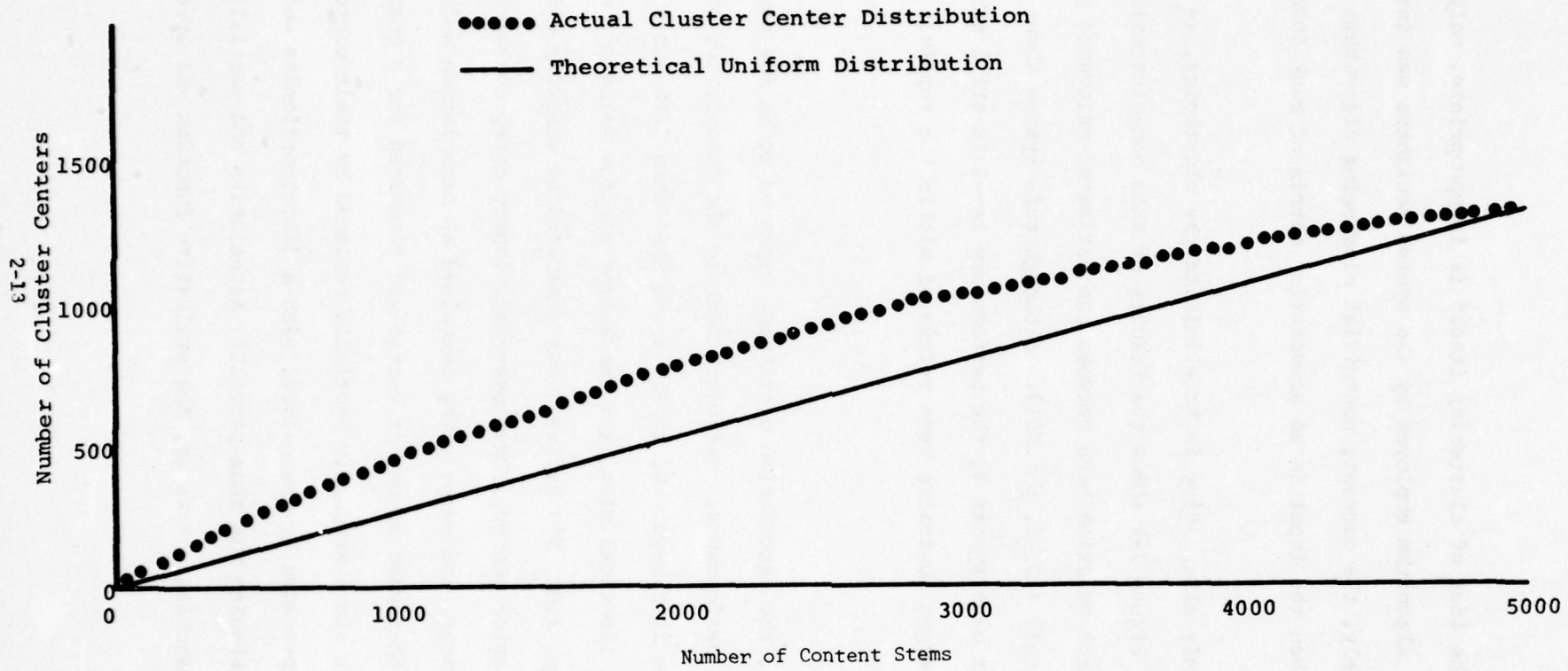


Figure 2-3 Cumulative Distribution: Number of Cluster Centers vs. Number of Content Stems

say that the idea of clustering itself is inappropriate, only that the particular algorithm employed by the system designers was ineffective. Unfortunately, the current, nontrivial clustering algorithms become prohibitive when the input is an association matrix of rank 5000.

The only alternative is to eliminate the clustering, as was done in system 4. Figure 2-4 shows the results of this non-clustering system. There is a moderate relation here between the estimated relevance and the human ratings ($r(33) = 0.45$, $p < .001$). Although this system shows a significant improvement over system 3, the performance level is still suboptimal. Only 5 of the target abstracts were retrieved within the top 64: only 3 within the top 10.

Again, the association dictionary appeared to be the source of the suboptimal performance. Referring back to the description of a dictionary entry shown in Figure 2-2, it was shown that when STEM was a cluster center, C_1 was the identical stem, and the weight was the value of the identity correlation, 1.0. Now system 4 was generated by allowing every content stem to be a cluster center. Thus every dictionary entry contained itself as a cluster center, and every entry contained an association weight of 1.0. The result is that when a concept vector was generated for a query or abstract, the content stems were only partially replaced by their context. The stems were also replaced by themselves, with a disproportionate weight. The retrieval mechanism was thus partially associative and partially boolean, i.e., identity matching. However, the evaluative function was operating on the

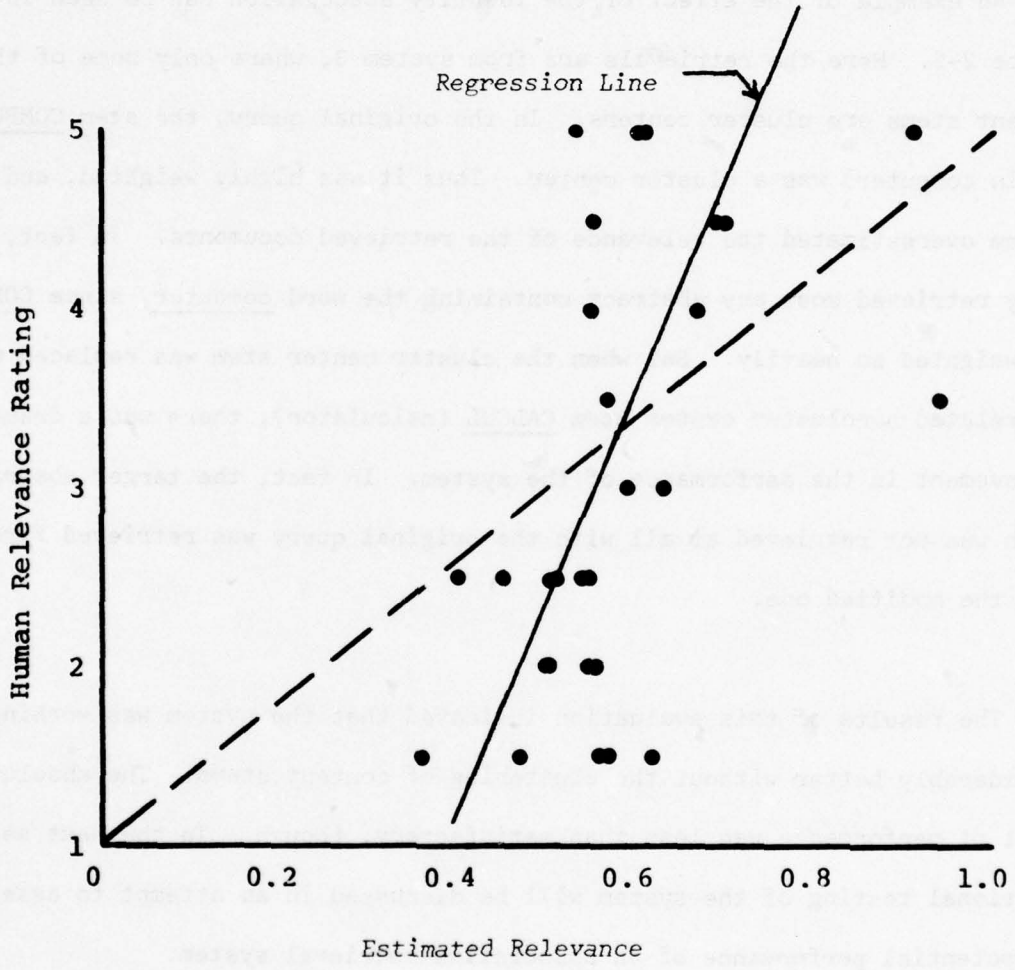


Figure 2-4 RADCOL System 4 Retrieval Performance
($r = 0.45$, $P < .001$)

premise that the concept vectors contained only associations. This discrepancy was considered to be the cause of the suboptimal performance.

An example of the effect of the identity association can be seen in Figure 2-5. Here the retrievals are from system 3, where only some of the content stems are cluster centers. In the original query, the stem COMPUT (as in computer) was a cluster center. Thus it was highly weighted, and the system overestimated the relevance of the retrieved documents. In fact, the query retrieved most any abstract containing the word computer, since COMPUT was weighted so heavily. But when the cluster center stem was replaced with the related noncluster center stem CALCUL (calculator), there was a dramatic improvement in the performance of the system. In fact, the target abstract which was not retrieved at all with the original query was retrieved first with the modified one.

The results of this evaluation indicated that the system was working considerably better without the clustering of content stems. The absolute level of performance was less than satisfactory, though. In the next section additional testing of the system will be discussed in an attempt to assess the potential performance of an associative retrieval system.

QUERIES { Computer Program (for) Steam Power Control
Calculator Program (for) Steam Power Control

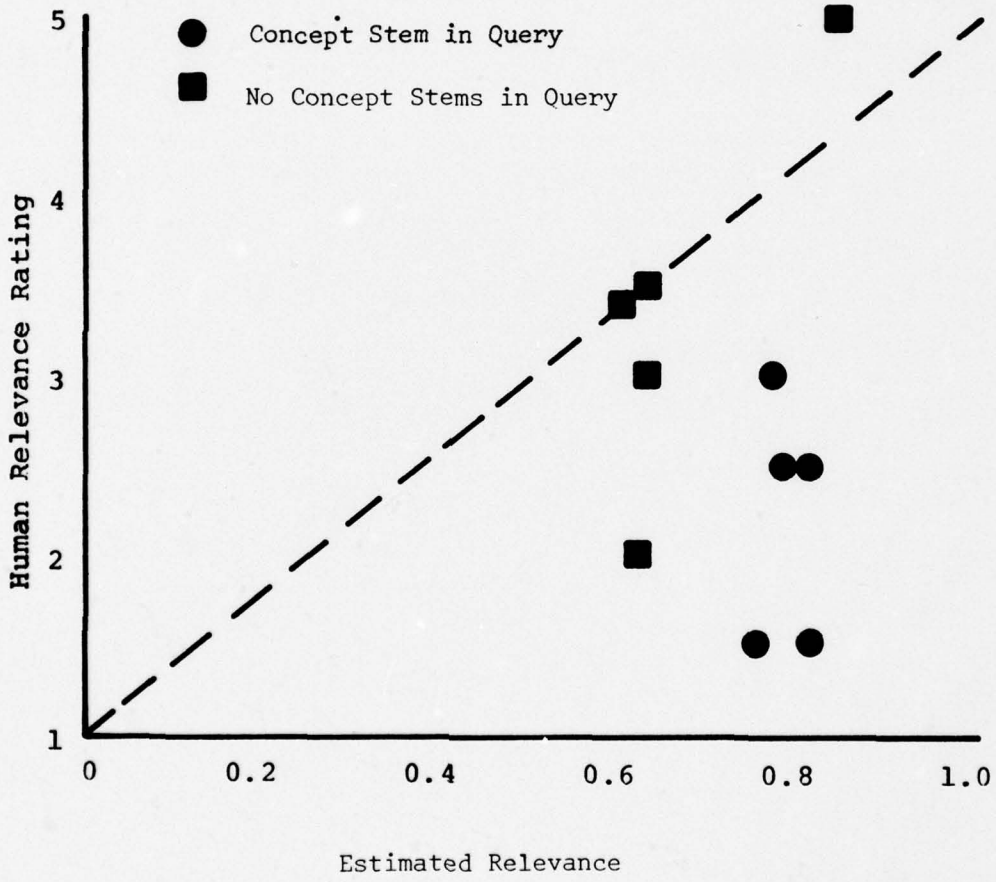


Figure 2-5 Effect of Concept (Cluster Center) Stems in a Query on Retrieval Performance

SECTION 3

ADDITIONAL TESTING OF ASSOCIATIVE RETRIEVAL TECHNOLOGY

The above study demonstrated that the nonclustering system was performing significantly better than the system which attempted to cluster associated stems. However, the absolute level of retrieval performance was less than extraordinary. Questions remained as to the influence of the identity associations and of the differential density of the document space (i.e., frequent or infrequent topic domains) on retrieval effectiveness.

In addition, serious problems remained in the latest implemented version of the RADCOL system. Some obscure coding errors remained from previous versions of the system which, for example, caused the system to occasionally cycle infinitely during a retrieval operation. Moreover, idiosyncratic code which functioned correctly in the originally implemented system, introduced errors and irratic behavior into the system when modifications to either the code or the data base were attempted. It was thus decided that system modules of questionable reliability would be rewritten in FORTRAN to insure correct and tractable operation of the system. (FORTRAN was an acceptable language for the reimplementation since much of the bit manipulation, presumed to be required by the JOVIAL routines, was found to be unnecessary.)

For a data base of 4000 abstracts, it is difficult to assess the relevance of every abstract to each query submitted to the system. It was decided that to obtain more reliable estimates of retrieval performance with

the modified systems, it would be desirable to evaluate objectively the topical structure of the data base. This analysis is discussed in the next section.

3.1. OBJECTIVE ANALYSIS OF THE DATA BASE

In order to evaluate the adequacy of retrievals generated by a particular query, it is necessary to assess the size of the subset of potentially relevant abstracts. Fortunately, information existed in the data base that allowed the conceptual structure of the data base to be objectively analyzed, independent of the exact content of the abstracts. The CIRC abstracts used in this study contained topic tags, sets of standard descriptor words and phrases. By tabulating the occurrences of the topic tags, it was possible to assess the relative frequency of various topics within the data base.

Figure 3-1 shows the distribution of the 22,000 occurrences of topic tags over the 4000 abstracts. As can be seen in this figure, many of the topic tags occur with only a single abstract. Most of these topics refer to very specific or atypical information, such as model numbers of particular pieces of equipment (e.g., 2009 Starlight Scope, AB47J Helicopter, and Norepinephrine).

At the other end of the distribution, several topic tags are found to occur with significant subsets of abstracts. The 50 most frequent topic tags, shown in Table 3-1, occur with 28 or more abstracts. For the purposes

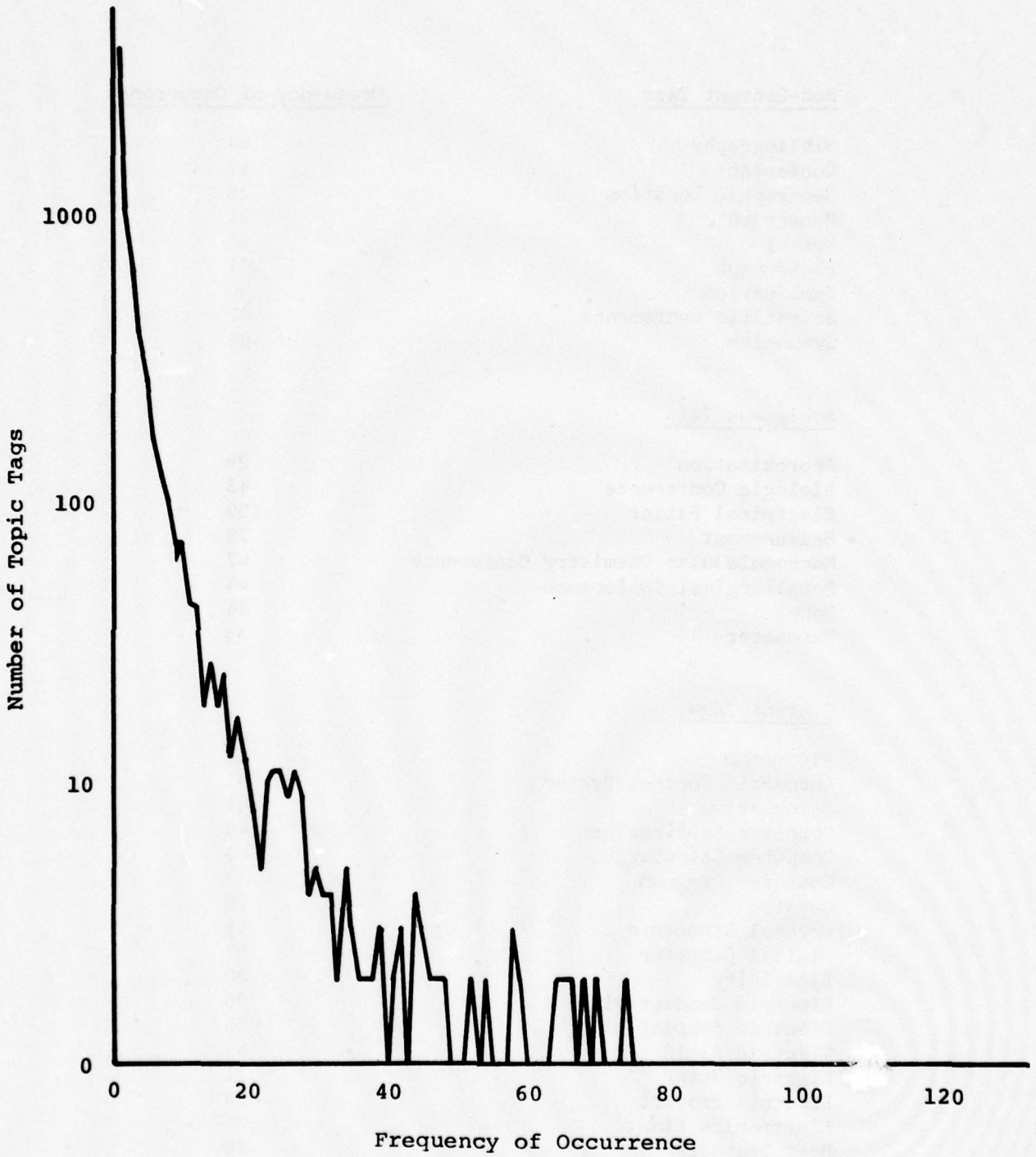


Figure 3-1 Distribution of Topic Tags
Over 4000 Abstracts

Non-Content Tags	Frequency of Occurrence
Bibliography	64
Conference	67
Geographic Location	28
Monograph	31
Patent	33
Photograph	201
Publication	30
Scientific Conference	32
Symposium	63
Ambiguous Tags	
Approximation	29
Biologic Conference	43
Electrical Patent	129
Measurement	38
Macromolecular Chemistry Conference	47
Metallurgical Conference	44
Model	38
Parameter	33
Content Tags	
Algorithm	69
Automatic Control System	41
Calculation	53
Computer Application	46
Computer Calculation	43
Computer Program	37
Crystal	29
Crystal Structure	28
Digital Computer	57
Elasticity	30
Electric Conductivity	30
Electric Current	35
Electric Field	45
Electric Motor	31
Electric Property	34
Electronics Plant	34
Heat Transfer	29
Hydrodynamics	28
Hydrogen	33

Table 3-1 50 Most Frequent Topic Tags

Content Tags (cont'd)	Frequency of Occurrence
Industrial Plant	41
Magnetic Field	65
Mathematical Expression	73
Mathematical Method	141
Mathematical Model	58
Mathematics	43
Oscillation	51
Rock	31
Semiconductor Device	33
Semiconductor Single Crystal	36
Silicon	29
Single Crystal	44
Temperature Dependence	57
Water	40

Table 3-1 (cont'd)

of analyzing the content of the data base, these topic tags can be classified into three groups corresponding to their relevance to the substantive content of the abstracts. The "Noncontent Tags" tend to describe the type of information to be found in a document, such as Photograph, Bibliography, or Conference. The "Ambiguous Tags" refer to very general topics such as Metallurgical Conference or Model. Although these tags might contain more information about the subject matter of the abstract than do the "Noncontent Tags", the amount of additional information is not sufficient to accurately describe the content of the abstracts.

The last group, "Content Tags", refer to substantive topics within the data base, such as Digital Computer, Crystal Structure, or Mathematical Model. These tags can be used to obtain the subsets of potentially relevant abstracts within these topic areas. In addition, they can be used to assess the major content areas of the data base. For example (from Figure 3-2), the area of digital computers appears to be well represented in the data base (Automatic Control System, Computer Application, Computer Calculation, Computer Program, Digital Computer) as does the subject of crystals (Crystal, Crystal Structure, Electric Conductivity, Semiconductor Device, Semiconductor Single Crystal, Silicon, Single Crystal). (Note that these topic tags are only the most frequent, occurring with a minimum of 28 abstracts. In addition, there are other related topic tags which occur with much less frequency.) An analysis of the subjects covered by the Content Tags indicate that the data base is clearly not homogeneous, but rather, certain substantive domains tend to dominate. Equipped with this information, it is thus possible to

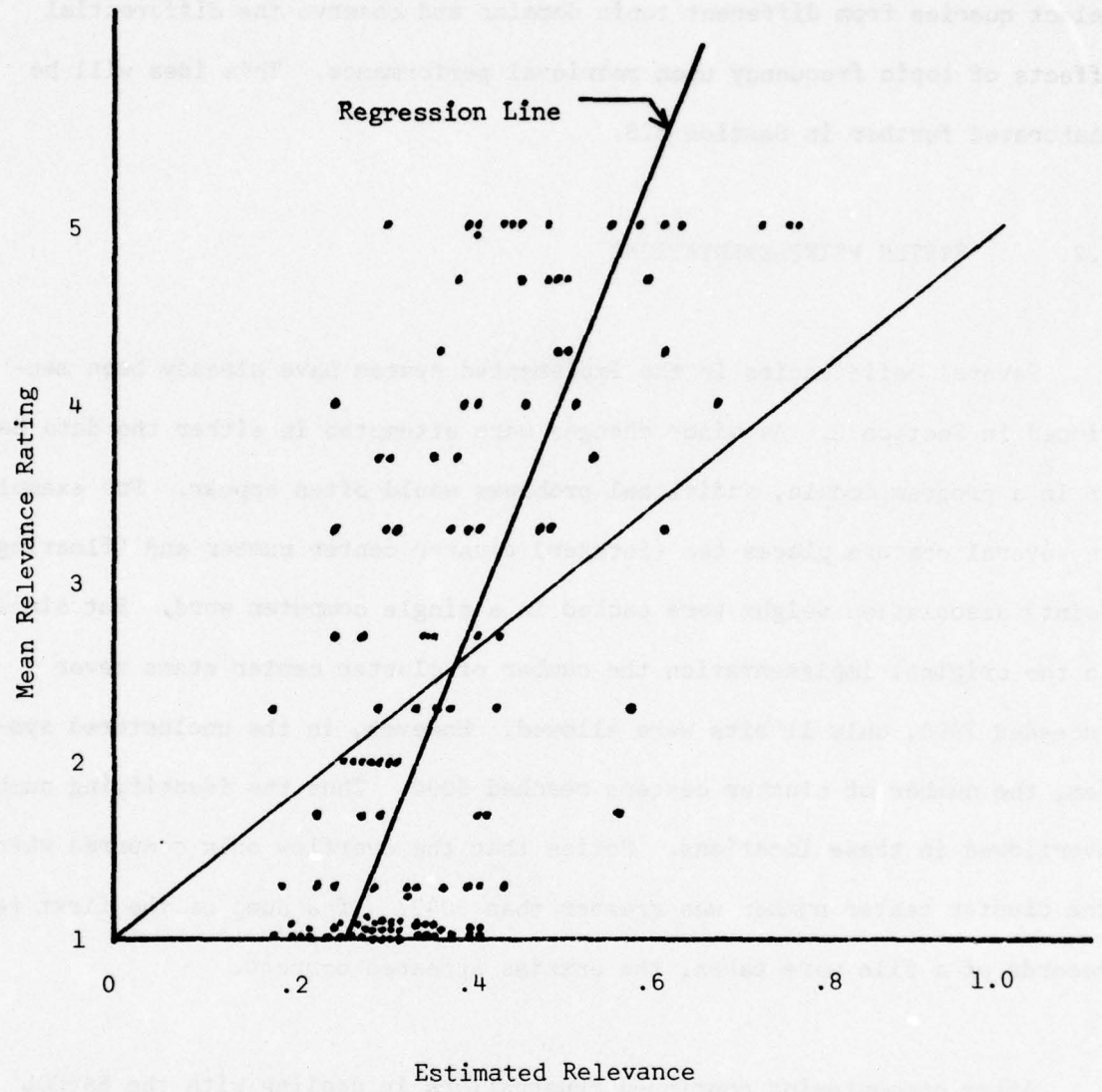


Figure 3-2 Retrieval Performance for RADCOL
Version A (Identity Association ≡ 1.0)
(r(118) = 0.67)

select queries from different topic domains and observe the differential effects of topic frequency upon retrieval performance. This idea will be elaborated further in Section 3.3.

3.2. SYSTEM REIMPLEMENTATIONS

Several deficiencies in the implemented system have already been mentioned in Section 2. As minor changes were attempted in either the data base or in a program module, additional problems would often appear. For example, in several obscure places the (integer) cluster center number and (floating point) association weight were packed in a single computer word. But since in the original implementation the number of cluster center stems never exceeded 2000, only 11 bits were allowed. However, in the unclustered system, the number of cluster centers reached 5000. Thus the identifying number overflowed in these locations. Notice that the overflow only occurred where the cluster center number was greater than 2047. If a dump of the first few records of a file were taken, the entries appeared correct.

After encountering continued frustrations in dealing with the RADCOL implementation, it was decided that the only way to effectively perform additional tests on the system in a reliable manner would be by first rewriting the troublesome system modules in a reliable and tractable manner. Programs prior to the cluster center generation program were functioning correctly, and thus were not modified. The main data base generation programs, including the dictionary program, the concept vector generation program,

and the inverted list building program were all rewritten in FORTRAN, as were many additional diagnostic and utility routines. The entire on-line system was replaced by a batch retrieval program. The loss of the interactive capabilities of the original on-line system were more than compensated by the efficiency and reliability of its replacement.

The results of the first tests implicated the identity associations in the dictionary as one cause of the suboptimal performance. The conversion of the dictionary generation program from JOVIAL to FORTRAN greatly simplified the task of varying the identity associations for the additional study. Three complete versions of RADCOL (without clustering) were generated. Version A retained the identity associations at 1.0, but was regenerated with the new program modules. In version B, the identity associations were arbitrarily defined as 0.5. (The motivation for this system was that the weights of the identity associations would be more comparable to the true associations in the data base.) In version C, the identity associations were eliminated from the dictionary. This last RADCOL version was thus a true associative retrieval system.

3.3. EVALUATION OF THE SYSTEMS

The knowledge of the topical structure of the data base, discussed in Section 3.1., enabled target abstracts to be selected from particular topic areas. Two extreme sets of topic tags were selected for the study. The first set consisted of tags which occurred frequently in the 4000 abstracts.

These tags designated subject areas for which there were a sizeable number of abstracts. The other set of topic tags corresponded to subject areas which were very infrequent in the data base. For the frequent topic group, a set of ten abstracts was designated as target abstracts for retrieval purposes. For the infrequent group, a set of 14 abstracts was selected. The infrequent abstracts were chosen subject to the additional restriction that each of the topic tags associated with an abstract from this group must be infrequent. That is, an abstract was classified as belonging to the frequent topic group if any one of its topic tags occurred frequently in the data base. An abstract was considered as infrequent only if all of its associated topic tags were infrequent in the data base.

For each abstract, a query was generated for which the abstract would contain the required information. No attempt was made to extract words from the abstracts for use in the queries, although there was some overlap in the words used in the queries and abstracts. The set of 24 queries, 10 frequent and 14 infrequent, is contained in Appendix B.

The set of queries was input to each of the RADCOL versions in turn. For each query and each system version, the top 64 abstracts were retrieved. The rank order of the target abstract was recorded for each query/version combination, and the top 5 retrieved abstracts for each version were tabulated for each query. That is, every query had associated with it between 5 and 15 abstracts, depending upon the overlap in the abstracts retrieved from each RADCOL version.

In studies where ratings are used to derive a scale, the extent of the rating scale is subjectively equated with the range of the items evaluated. The lowest value on the scale will be paired with the minimal item in the set, and the highest value with the maximal item. Thus if the set of items used in a study represents only a portion of the actual range, the rating scale will either underestimate the low end, or overestimate the high end of the range encountered, depending on which end of the distribution is truncated. In the study discussed in Section 2, every abstract used in the ratings had been retrieved by the query with which it was paired. Thus completely unrelated abstract-query pairs were not included in the study. As a result, it is reasonable to expect that some of the lowest ratings were actually underestimates of the true relevance rating. In order to avoid this problem in the present study, one or more abstracts were randomly paired with each query. The range of abstracts then seen by the raters will extend from irrelevant to high relevant. These irrelevant abstracts were included to insure that the rating scale corresponded to the entire range of relevance. They were not, however, included in the subsequent analysis.

In all, there were 231 pairs of abstracts and queries. Three raters saw every query, one at a time. For each query, they saw the corresponding set of abstracts, individually and in a random order. The raters' task was to read the query and abstract, and then rate on a scale from 1 to 5 the relevance of the abstract, given the query.

3.4. RESULTS AND DISCUSSION

To evaluate the reliability of the human relevance ratings, the responses for the three raters were correlated across every query-abstract pair. The product-moment correlations among the judges were 0.83, 0.84, and 0.91. With 229 degrees of freedom, it is clear that each rater was using an essentially identical definition of the concept of relevance. (The consistency in the rating results were even more impressive, given the diverse backgrounds of the judges: a programmer, a mechanical engineer, and an experimental psychologist.)

The ratings were averaged across the three judges to arrive at an estimate of relevance between a query and each of the abstracts retrieved by that query. As an additional check on the reliability of the subjective ratings, the mean ratings given to the target abstract-query pairs were evaluated. Each of the 24 pairs had a mean relevance rating of at least 4.67 on the 5 point scale. In addition, every abstract which was randomly paired with a query for the rating procedure was given a rating of 1.0. These results supported the use of the rating paradigm as a reliable and valid evaluation procedure.

Both the rated and estimated relevance were tabulated within each RADCOL version for the first five abstracts retrieved by each of the 24 queries, as well as for the target abstracts. Scatter plots of the rated and estimated relevance of these retrieved abstracts are shown in Figures 3-2, 3-3, and 3-4

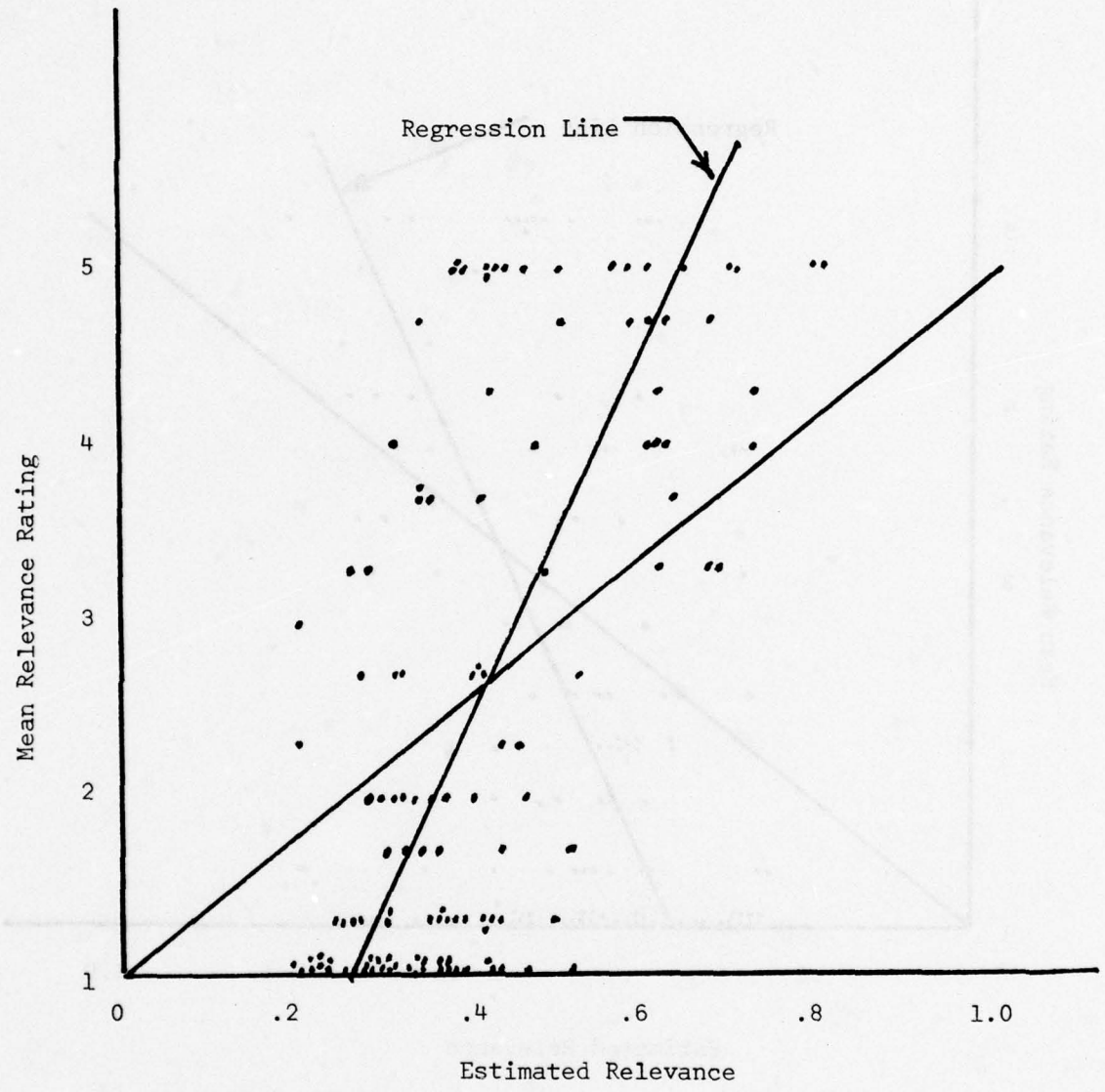


Figure 3-3 Retrieval Performance for RADCOL Version B
(Identity Association ≡ 0.5)
(r(118) = 0.66)

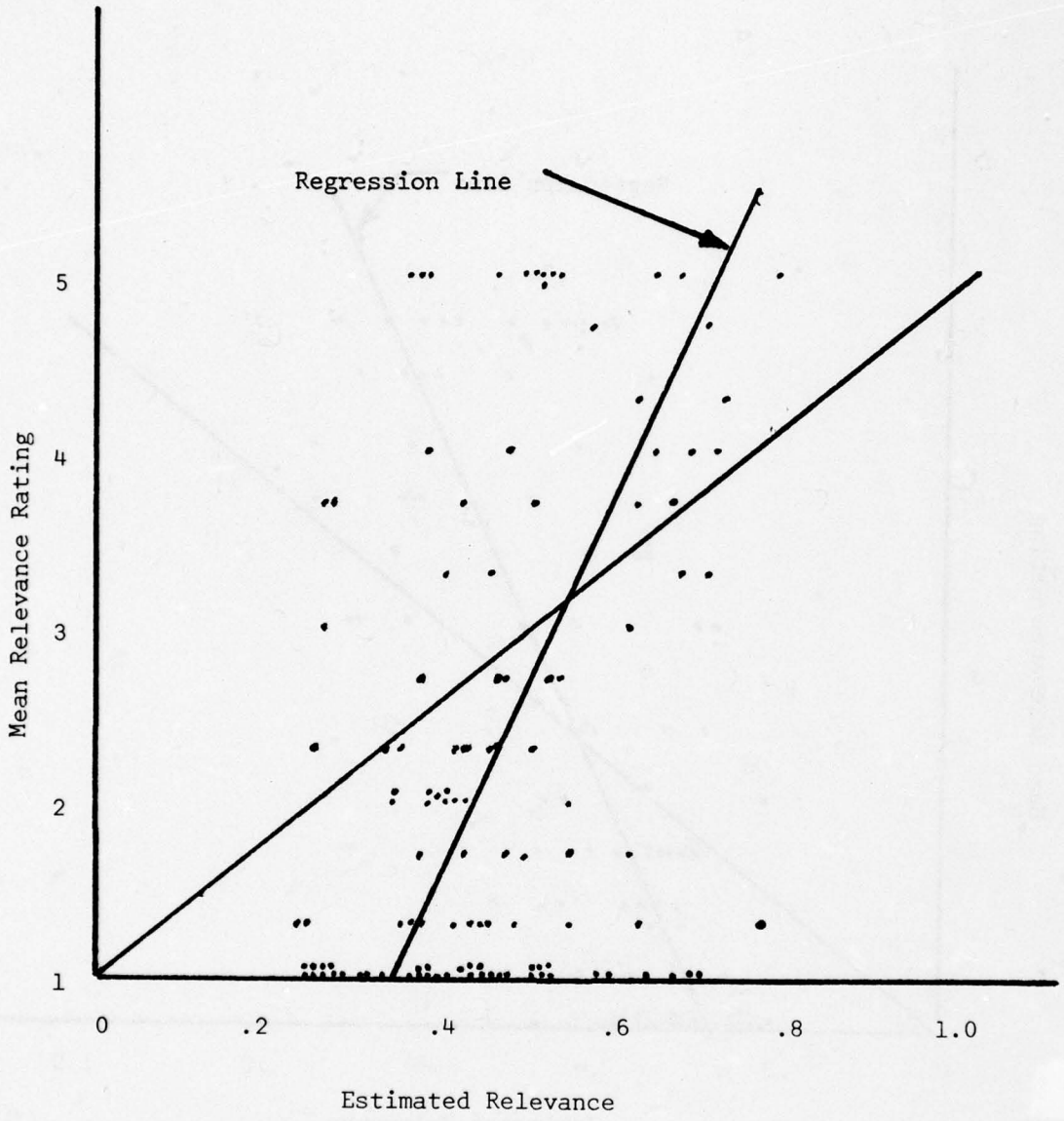


Figure 3-4 Retrieval Performance for RADCOL Version C
(Identity Association ≡ 0.0)
(r(118) = 0.36)

for RADCOL versions A, B, and C, respectively. The corresponding product-moment correlation between the two relevance measures was 0.67, 0.66, and 0.36 for versions A, B, and C. The difference in correlation value between version C (no identity associations) and the other two versions (identity associations $\equiv$ 1.0 and 0.5) was highly significant ($z = 3.17$, $p < .001$).

The difference in the correlations was surprising. The implication is that the RADCOL version without any identity associations (i.e., the purely associative system) performed relatively poorly. In examining the three scatter plots, a discrepancy was noted in the distribution of points which may have contributed to the poor performance of this version. If the irrelevant abstract-query pairs are examined (i.e., those pairs receiving a mean rating of at most 2.0), version A retrieved only one abstract with an estimated relevance score of 0.50 or higher. Similarly, version B retrieved only two irrelevant abstracts in this category. Version C, the pure associative system, retrieved 18 abstracts in this area. It appears that this last version is overestimating the relevance of unrelated abstracts.

One possible cause for the overestimation of relevance in version C is the particular definition of an association used in the RADCOL system. The association between two stems in this system is defined as the overlap or co-occurrence of the two stems across the entire data base. The greater the overlap in the occurrence of the two stems, the greater the association between them. This relation between the overlap of words (stems) and their semantic relatedness was empirically investigated by Rubenstein and Goodenough [6]. They found that synonymy and overlap were clearly related, but only if

the amount of overlap of the two words was relatively great. At moderate or low levels of overlap, there was no relation between the amount of overlap and the (subjectively determined) semantic relation between the two words. In the RADCOL system, low or moderate levels of overlap between two stems may thus result in a spuriously high association value being assigned to the stem pair. The implication is that these spurious associations will artificially inflate the estimated relevance of an abstract for a query, especially when there are no other common associations between the query and abstract. The result would be the discrepant points in Figure 3-4.

With RADCOL versions A and B, the effect of these artifactual associations would be greatly diminished. In these two versions, the magnitude of the identity associations will be larger than these spurious associations (often by a factor of 2 or 3). The larger association weights of the identity associations will tend to dominate the estimation of relevance, and thus the influence of these irrelevant associations will be minimal. (Implications of this hypothesis for associative retrieval systems will be discussed further in Section 4).

An unexpected finding in examining the performance of version A was that there was a marginally significant difference in the correlation values between system 4 (see Figure 2-4) and version A ($z = 1.634$, $p \simeq .05$). Conceptually, these two systems are identical. The improved performance of version A, however, can be attributed to the corrections to the software discussed in Section 3.2.

Another approach to the evaluation of the three RADCOL versions is by assessing directly the relevance of the abstracts retrieved. If one of the systems is performing better than another, then the mean (rated) relevance of its retrieved abstracts should be greater. Differences in overall abstract relevance were assessed by a one-way analysis of variance (ANOVA). Across the set of 24 queries, there were no significant differences between the mean relevance of the top five abstracts retrieved by each version ($F(2,69) < 1$). The ANOVA table is presented in Table 3-2. (The analysis of variance was repeated in a two-way design, where the second factor represented a query topic as either frequent or infrequent. The only significant effect was for the query factor. The difference was not surprising since one would expect that there are more relevant abstracts to be retrieved from the frequent topic classes). This analysis indicates that overall, the three systems were retrieving abstracts of comparable relevance, even though estimated relevance was a poorer predictor of true abstract relevance for version C.

An important test of an information retrieval system is that the system should be able to retrieve the abstract which was used to generate a test query. But in a retrieval system such as RADCOL where the abstracts are retrieved according to their estimated relevance to the query, the target abstract should not only be retrieved, but should consistently be the first abstract retrieved. The results of the retrieval of the target abstracts is shown in Table 3-3 for the three RADCOL versions. Although the three versions differ slightly in the retrieval order of the target abstracts, overall the results are impressive. For version A, 16 of the 24 target abstracts were

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Squares	F
Version	2	.41	.20	<1 (not significant)
Query/Version	69	39.85	.58	-
Total	71	40.26		

Table 3-2 Analysis of Variance Table for Evaluating the Overall Relevance of Retrieved Abstracts

	Version A		Version B		Version C	
Query #	Retrieved Rank	Estimated Relevance	Retrieved Rank	Estimated Relevance	Retrieved Rank	Estimated Relevance
1	1	.41	1	.39	3	.36
2	1	.64	1	.70	1	.67
3	1	.73	2	.70	7	.62
4*	>64	-	>64	-	>64	-
5	1	.44	5	.44	6	.39
6	1	.31	1	.38	1	.51
7	3	.39	3	.34	19	.24
8	1	.49	1	.50	4	.53
9*	>64	-	>64	-	>64	-
10*	15	.29	9	.37	2	.51
Number of Target Abstracts Retrieved First	6		4		2	
11	1	.76	1	.79	1	.71
12*	46	.21	>64	-	>64	-
13	1	.46	1	.42	5	.36
14	1	.62	1	.64	1	.64

Table 3-3 Retrieval Performance for the Target Abstracts

15	1	.56	1	.56	1	.53
16	1	.59	1	.60	1	.51
17	1	.59	1	.58	1	.46
18	1	.45	1	.46	1	.38
19*	12	.21	6	.22	> 64	-
20	1	.77	1	.79	1	.78
21	1	.41	1	.42	1	.49
22	1	.49	1	.50	18	.50
23	4	.46	1	.58	4	.59
24*	13	.44	50	.44	> 64	-
Number of Target Abstracts Retrieved First	10		11		8	

* Queries for which the target abstract was not one of the first five abstracts retrieved by Version A.

Table 3-3 Continued

retrieved first: 18 within the first 5. For versions B and C, the corresponding number of target abstracts retrieved first was 15 and 10.

Two important comments are in order here concerning the information in Table 3-3. First, there is a clear difference with reference to whether a target abstract vector is located in a dense or sparse area of the vector space. When an abstract is located in a sparse area of the space, then as long as the query vector is anywhere near the abstract, it will not only be retrieved, but its rank will typically be one. But if that target abstract vector were in a dense area of the space, the location of the query vector would be more critical (in terms of target abstract retrieval). Here the target abstract vector may be quite near the query vector, but so may several other relevant abstract vectors. Thus the target abstract might be retrieved second or third or ... . If it were in a sparse area, it would certainly be retrieved first. This situation is illustrated in Table 3-4, where the first five retrieved abstracts from version A are listed for a frequent and infrequent topic query. With the frequent query there is a gradual diminution of estimated (and rated) relevance for each subsequent abstract retrieved. With the infrequent topic query, there is a difference of a factor of two between the estimated relevance of the first and second abstracts retrieved. The problem is not with the evaluative function, for the estimated relevance will have the same interpretation in both situations. The problem is that with a retrieval system in which the abstracts are retrieved in order of their estimated relevance, the output of the abstracts on an ordinal scale distorts the estimated relevance, which is measured on an interval scale. This ranking of the retrieved abstracts may affect the perceived operation of

Frequent Topic Query: Electron Streams Formed by Electrical and Magnetic Fields

Retrieval Order	Estimated Relevance	Rated Relevance
1	.73	5.0(Target Abstract)
2	.68	4.0
3	.62	3.33
4	.58	2.33
5	.57	1.67

Infrequent Topic Query: Serotonin Levels as Estimates of Radio-resistance and Magnetic Fields

Retrieval Order	Estimated Relevance	Rated Relevance
1	.59	5.00(Target Abstract)
2	.23	1.67
3	.22	1.00
4	.20	1.00
5	.20	1.00

Table 3-4 Abstract Retrievals from Dense and Sparce Portions of the Abstract Space

the system, for the attention paid by the user of the system to the list of abstracts retrieved will often be inversly related to the rank order of the abstracts.

Secondly, referring back to Table 3-3, there were several queries for which the target abstract was either retrieved far down the list, or was not retrieved at all. Since version A appeared to have performed the best in reference to retrieving target abstracts, version A's failures were analyzed in an attempt to locate the causes of the suboptimal performance. One problem was traceable to a system malfunction, and one to the system user. Other difficulties were more conceptual in nature.

For query 4 (MEASUREMENT *INSTRUMENTS* FOR ELECTRICAL COMMUNICATION), the target abstract was not retrieved because several of the important words in the abstract were not stemmed correctly. As a result, the incorrect stems were not recognized as content stems by the system, and the concept vector was not a true representation of the abstract. The stemming problem appeared to have been a (hardware/software) malfunction, not a design or implementation problem in the stemming routine.

The problem with query 24 (SOVIET AGREEMENTS ON SCIENTIFIC AND TECHNICAL COOPERATION) was not with the system, but with the user of the system. This query did retrieve relevant abstracts, as reflected in the fact that the mean relevance rating for the first five abstracts was 4.0. This query, however, referred to just too broad a topic for a data base containing abstracts

solely in the area of foreign technology. The implication of the query is that a wide range of information is needed, and that is just what the system supplied.

The other problems were more central to the functioning of the system. For example, in query 10 (AUTOMATIC PROGRAMMING, COMPOSING, AND PRINTING MACHINE) the word composing was stemmed to compos. But in abstracts, the word composite is also stemmed to compos. The result is that abstracts dealing with carbon compounds and other composite materials were retrieved along with abstracts whose domain was the printing process. In addition, there appeared to be an emphasis on this stem in the retrieval operation. The other stems in the query did not influence the selection of abstracts as much as the stem compos.

As in query 10, the common cause of the remaining retrieval difficulties was that one portion of a query was exerting undue influence on the selection of abstracts to be retrieved. For example, query 12 (SOVIET TRACTOR AND SCRAPER PRODUCTION) retrieved abstracts covering the Soviet Union, independent of whether heavy equipment was discussed. This varying influence of query words was not limited to substantive topics. Query 19 (INCREASED RADIOACTIVE IMMUNITY BY POLYSACCHARIDE INJECTION) retrieved several abstracts which discussed increases, independent of whether radioactive immunity was the topic of the increase.

The problem of certain stems exerting influence upon a retrieval operation appeared to be correlated with the generality of the stem in the data

base. These stems had two statistical features in common: the frequency of occurrence of the stems across the data base was high, and the Dennis measure for each of the stems was relatively low. That is, these stems occurred often in abstracts, but they did not discriminate very well among the abstracts. When the ratio of the Dennis measure to frequency of occurrence was calculated for a set of stems, it was found to vary inversly with the (subjective) generality of a stem.

This ratio, which will be defined as a stem's "importance index," can be used to normalize the influence of stems in queries by multiplying the weight of a query stem (nominally 1.0) by the "importance index". In a pilot test this procedure was approximated by scaling the "importance index" to integer values and repeating each query word the indicated number of times. (The integer values ranged from 1 to 10.) When queries 12 and 19 were modified in this manner and input to version A, the retrieval of their respective target abstracts improved markedly. For query 12, the estimated relevance increased from 0.21 to 0.38, and the rank of the target abstract increased from 46 to 1. Similarly with query 19, the estimated relevance of the target abstract increased from 0.21 to 0.23, and the rank increased from 21 to 1.

This informal investigation implied that through a simple weighting procedure, it might be possible to further improve the retrieval effectiveness of an associative retrieval system without feedback or complex user interaction. Salton [7] has also discussed a similar procedure for normalizing the density of a document space. Although comprehensive and statistically reliable testing is required in order to verify the influence of the

"importance index" weighting, it does appear premature to accept the results in this report as indicative of the optimal performance of an associative retrieval system.

SECTION 4

CONCLUSIONS AND RECOMMENDATIONS

4.1. SYSTEM EVALUATION -- SOFTWARE LEVEL

Many performance aspects, such as reliability, accuracy, and efficiency enter into the evaluation of a software system. In reference to RADCOL as originally implemented, several factors, some of which were beyond the control of the system designers and programmers, resulted in a system which was less than effective. Evidence of the use of modern programming practices was lacking. For example, several of the program modules were written in such a way that it was more expedient to completely rewrite them than to spend the effort to debug and/or modify them. Other examples of suboptimal performance included files that were sorted twice (on the same key) without any intervening reordering of the records.

A discussion of other portions of the software, such as the on-line system and the user interface could continue, but would be fruitless. By the time the evaluation was completed, there had been so many necessary changes made to the original RADCOL system that RADCOL was no longer the subject of the evaluation. It was clear that the original RADCOL system was at best marginally effective. Moreover, it appeared that it could have been expanded to handle larger data bases only in an inefficient and ad hoc manner. What started as an assessment of the RADCOL system became, by necessity, an assessment of associative retrieval methodology.

4.2. SYSTEM EVALUATION -- CONCEPTUAL LEVEL

The results of the second experiment showed that overall, version A was performing slightly better than version B, and significantly better than version C. But version C is the purely associative retrieval system. Version A includes the identity associations in the dictionary with a weight of 1.0. The effect of these identity associations on a typical query concept vector is shown in Table 4-1.

Assume for the moment that the query vector in Table 4-1 is being evaluated against an abstract whose vector is identical to the query vector. The dot product between the two vectors will be the (normalized) sum of the squared weights shown in the third column. Note that the terms from the identity associations are between one and two orders of magnitude greater than the terms from the computed associations. The result is that the associations are an insignificant factor in the calculation of the dot product. That is, the associations could have been eliminated and the resulting dot product would have varied only slightly. If the abstract vector varied somewhat from the query vector, the dot product would be reduced, but the degree of similarity would still be heavily influenced by the identity associations.

If a query and an abstract vector have only the nonidentity associations in common, the resulting dot product will be greatly reduced. But more importantly, many of these associations will be spurious as discussed in

Query: Orientation of Metal Oxide Crystals

Query Vector

Stem Number	Weight	Weight^2
677	.169	.028
981*	1.000	1.000
1409	.137	.019
1727	.168	.028
1806	.252	.064
1954	.114	.013
2271	.179	.032
2518	.270	.073
2628*	1.000	1.000
2996	.241	.058
3029	.121	.015
3043*	1.000	1.000
3077*	1.000	1.000
3161	.132	.017
3959	.218	.048
4205	.121	.015

* Query Stems

Table 4-1 Query Concept Vector From Version A.

section 3.4. Thus abstracts and queries which do not share some of the same content stems will have a very low dot product, but even then the estimated relevance will be overstated. This situation is illustrated in Figures 3-2, 3-3, and 3-4, where virtually none of the abstracts has an estimated relevance value less than 0.2.

But the version where the associations were not influential is also the version where overall performance was the best. The conclusion must then be that the statistical associations (defined as the co-occurrence of stems across the data base) contribute little, if anything, to the retrieval performance. The important feature of the system appears to be the evaluative function, the mechanism by which the abstracts are rated for appropriateness to the queries. That is, the abstracts are not either included or excluded from a retrieval set, but are rated along a continuum in terms of estimated relevance, the distance between the query and abstract in a multidimensional space.

As an informal test of this hypothesis, another version of RADCOL was generated where the dictionary contained only the identity associations. When the 24 queries were submitted to this system, the results were essentially equivalent to the results from version A. Sixteen of the target abstracts were retrieved first; 18 were within the first five abstracts retrieved. In addition, almost every abstract that was one of the first five retrieved in version A was also retrieved high on the lists in the nonassociative version. More importantly, of the three abstracts that were retrieved high from Version A but not at all from the nonassociative version, each had a mean relevance rating of 1.0. The implication is that these abstracts had been retrieved

by version A as a result of strictly spurious associations between the query and these abstracts.

One of the arguments for associative retrieval systems is that documents can be retrieved which are conceptually similar but do not contain the same words. This may be a valid point if the retrievals were made on the basis of document titles, or short descriptive phrases. But in examining the abstracts from the data base used in this study, it appeared that a relevant abstract did typically contain one or more query words. There was no evidence of an abstract in the data base that was conceptually similar to another abstract or to a query but did not contain at least some of the same content words. There was one query (number 10) where version C did perform better than version A. But upon close inspection, the reason was that the query contained a stem whose "importance index" was very low. (This query was discussed in section 3.4.). For version C, the problem stem was not included in the query vector. Thus the improved performance of the associative system in this instance was due to the exclusion of a stem from the vector rather than due to any feature of associative retrieval. (When this stem was manually removed from the query, version A then performed as well as version C.)

It should not be concluded that the associative aspects of an associative retrieval system are ineffective. Perhaps the problem lies with the definition of associations. As mentioned previously, Rubenstein and Goodenough [6] have shown that a statistical definition of associations is only valid for relatively high overlaps of words. The diversity of the CIRC data base would certainly cast doubts on the amount of overlap possible between stems in the

abstracts. Perhaps more importantly, Deese [8] and Jenkins and Saporta [9] have shown that semantically meaningful associations can be either syntagmatic or paradigmatic, depending upon the parts of speech of the particular words. (Syntagmatic associations are typically sequential as in green thumb or play ball. Paradigmatic associations are those in which the two words fit a common grammatic paradigm such as hot cold.) Their research implies that the definition of associations used in the RADCOL system is perhaps too simplistic to be useful as a general relationship.

In addition, a meaningful association between stems in an abstract may be influenced by the proximity of the stems, as well as by paradigmatic and syntagmatic constraints. Two contiguous stems in a sentence may be much more associatively related than the first and last stems of an abstract. But the statistical definition of association in RADCOL is not influenced by the proximity of the stems in the abstract. The result is certainly additional spurious associations between conceptually independent stems. Given these and other problems of defining associations in a purely statistical manner, it is somewhat suprising that version C performed as well as it did. One reason is that when a particular stem appeared in both a query and an abstract, the same set of associations were mapped onto the two concept vectors. The comparison between the two vectors were then similar, if not identical, to a comparison of the actual stems in the query and abstract.

But even without the associative aspects, a system based on version A appears significantly superior to a Boolean type of retrieval system. One reason is the manner in which the queries and abstracts are represented as

vectors in a multi-dimensional space. With a Boolean system, the query defines a retrieval set on the universe of abstracts in the data base. A precise query might result in a very small or even null retrieval set. A general query might define a significant portion of the data base as the retrieval set. With a retrieval system where the queries and abstracts are represented in a continuous space, the specificity of a query takes on a different importance. Since the query is a vector in the space, a very precise query is desirable, for it will locate the query more precisely in the space. The query can be as long as an abstract and still retrieve relevant abstracts. (In fact one of the features of the RADCOL system was that an abstract can be defined as a query in order to retrieve similar abstracts from the data base.)

It is important to reiterate that the system evaluated in the second experiment performed very well. Relevant abstracts were retrieved and irrelevant ones were rejected sufficiently to obtain a correlation of 0.67 between rated and estimated relevance. The point is not that the system contains deficiencies, for as an experimental system it was quite successful. Rather, the conclusion is that the associative part of the associative retrieval system evaluated in this report was not contributing to the retrieval performance. It greatly increased storage requirements and processing time without improving the retrieval performance over an essentially nonassociative system with an evaluative function (i.e., version A).

4.3. RECOMMENDATIONS

The evaluation which has been described in this report leads to the following conclusions:

1. The system that evolved during the course of this study was an effective retrieval mechanism.

2. The statistical aspects of the system were shown to be unnecessary in order to obtain a reasonable level of retrieval performance. That is, the statistical measures used in this study did not appear to add significantly to the retrieval operation.

The following recommendations are based on the above conclusions:

1. The results of this evaluation have shown that a linguistic analysis of abstract contents was not adequately approximated in this study using purely statistical measures. The implication is that more conceptual and empirical work in the area of linguistic relationships will be necessary in order to develop more intelligent information retrieval systems.

2. But, perhaps a cautionary note is in order here. For the past ten years, Dr. Gerard Salton and his colleagues at Cornell University have been publishing studies indicating that statistical associations

are an effective mechanism for information retrieval systems. In addition, there are other systems which use statistical associations in different ways. For example, the Message Extraction Through Estimated Relevance (METER) system (see RADC-TR-77-89) uses statistical associations in a small, dynamic I&W environment to evaluate the relations among messages. RADCOL, on the other hand, is concerned only with the relation between the abstracts and a query in a large, static data base, where each abstract is an independent element of the system. The differences between two systems such as RADCOL and METER (in terms of data base size and type, system function, and operational environment) are large enough that it is unclear at this moment how the results of this study can be generalized.

3. Because of the superior performance of the system using only 1.0 correlations, it is possible to design a relatively simple retrieval system for use with Science and Technology (S&T) abstracts, where the abstracts and queries are directly represented in a multidimensional space, rather than indirectly through their (statistical) associations. Such a system would require significantly less storage and less start-up time, and thus would not only perform better than a Boolean retrieval system (due to the estimated relevance function), but would be cost competitive as well. The direct retrieval system should initially be implemented as an experimental system in order to examine its performance characteristics on larger data bases, and its compatibility with Boolean systems such as CIRC/ CIRCOL.

4.4. AREAS FOR FUTURE DEVELOPMENT

On the basis of our evaluation of the RADCOL system, we believe that additional research should be undertaken to develop a more effective associative retrieval system. We believe that is is important to develop this system for the following reasons:

1. The results of the RADCOL version A (where the influence of the associations was minimal) are encouraging enough to warrant further experimental development. We believe that we have shown that a system where the queries and abstracts are represented in a vector space, and where a metric for estimated relevance is employed, is effective in locating abstracts which are relevant to natural language queries.

2. Much more is known about word associations than is currently used in the RADCOL system. By developing algorithms based on linguistic knowledge, it should be possible to produce a more effective information retrieval system.

3. There is a clear need for further experimentation in the development of more effective retrieval techniques for Air Force and other governmental requirements. RADC has traditionally been among the leaders in the development of advanced retrieval techniques.

The following areas of development are suggested. The emphasis is on insuring that each component of the system is logically appropriate from a linguistic perspective.

1. Since the selection of content stems is of fundamental importance in an associative retrieval system, several aspects of stem selection should be investigated. Some of the factors include:

   a. The current stemming algorithm should be critically evaluated and other, more linguistically acceptable, stemming procedures should be examined for possible use in the system.

   b. The Dennis measure and other discrimination measures should be analytically and/or empirically evaluated. Although the Dennis measure was used in the RADCOL system, there does not seem to have been any attempt in the literature to validate this measure in either an analytic or an empirical way. It is quite possible that the use of the Dennis measure for stem selection contributed to the ineffectiveness of the associations.

   c. The "importance index" should be investigated as a means for normalizing the weighting of content stems in both the abstracts and queries. It is anticipated that using this index with the abstract vectors will transform the vector space so that the distribution of abstracts in the space will be more uniform.

d. The distribution of query terms should be investigated. It may be as important for queries to discriminate among the concept vectors as it is for the stems to discriminate among the abstracts.

2. The definition of term association, the dot product over the set of abstracts, should be critically examined along with other, more linguistically appropriate definitions.

3. The definition of query vector and document vector similarity, the dot product, should also be critically examined and other possible approaches evaluated.

4. The relevance feedback feature should be implemented and tested. Relevance feedback would allow the user to dynamically modify a query by specifying relevant and irrelevant abstracts, rather than explicitly specifying additional words (stems) to add or delete from the query.

5. A comprehensive user interface should be designed and implemented. The primary function of an information retrieval system is the facilitation of information accession by a user. But for many Boolean systems and for the RADCOL system, the user interfaces are more of a logical challenge than a natural extension of the user's cognitive abilities. A well-designed user interface will contribute to the overall efficiency of the retrieval system.

6. The performance of the system should be assessed for larger data bases (e.g., 100k documents) to detect possible problems (as in accuracy, storage, or timing considerations) resulting from the processing of large numbers of documents. Also the possibility of the use of data base partitioning should be investigated.

7. The compatibility of the associative retrieval system with a Boolean system such as CIRC/CIRCOL should be investigated.

The preceding list represents only a selection of the research projects in which an associative retrieval system might be expected to play a significant role. However, before more sophisticated and more accurate mechanisms can be developed additional research is needed in the areas of natural language analysis and information retrieval technology. The conclusion of this report was that the global statistical associations used by the RADCOL system were ineffective in terms of retrieval performance. This is not to say that more linguistically acceptable associative mechanisms are not useful for building more accurate information retrieval systems. Their actual effectiveness can only be evaluated through continued research.

REFERENCES

1. Dennis, Sally F., The Design and Testing of a Fully Automatic Index-Searching System for Documents Consisting of Expository Text, in G. Schecter (Ed.), Information Retrieval -- A Critical View, Thompson Book Co., 1967.

2. Black, Max, Vagueness, Philosophy of Science, 1937, Vol. 4, pp. 427 - 455.

3. Black, Max, Reasoning with Loose Concepts, Dialog, 1963, Vol. 2, pp. 1 - 12.

4. Hersh, Harry M. & Caramazza, Alfonso, A Fuzzy Set Approach to Modifiers and Vagueness in Natural Language, Journal of Experimental Psychology: General, 1976, Vol. 105, pp. 254 - 276.

5. Torgerson, Warren S., Theory and Methods of Scaling, New York: Wiley, 1958.

6. Rubenstein, Herbert & Goodenough, John B., Contextual Correlates of Synonymy, Communications of the ACM, 1965, Vol. 8, pp. 627 - 633.

7. Salton, G., Wong, A., and Yang, C.S., A Vector Space Model for Automatic Indexing, Communications of the ACM, 1975, Vol. 18, pp. 613 - 620.

8. Deese, James, Form Class and the Determinants of Association, Journal of Verbal Learning and Verbal Behavior, 1962, Vol. 1, pp. 79 - 84.

Deese, James, The Structure of Associations in Language and Thought, Baltimore, The Johns Hopkins Press, 1965.

9. Saporta, S., in Jenkins, J.J. (Ed.). Associative Processes in Verbal Behavior: Report of Minnesota Conference. Minneapolis: University of Minnesota, 1959.

APPENDIX A

QUERIES USED IN INITIAL RADCOL EVALUATION

The following is the set of queries used in the testing described in Section 2.

	QUERY	TARGET ABSTRACT
1.	X RAY RAMAN IR SPECTRA PCL4MCL6 ASCL3	2446
2.	DEFOCUSSED GALILEAN SYSTEM MIRRORS	2468
3.	DIFFERENTIATING CIRCUITS INTERVAL COUNTERS	2488
4.	DISPERSION CRYSTALS MONOCLINIC LATTICE	2632
5.	QUASI ABRUPT P-N JUNCTIONS	2676
6.	HUNGARIAN ROUMANIAN HIGH POWER TRANSMISSION LINE	2688
7.	FLOW DENSITOMETER FOR ALKYL SULFATES	2727
8.	SEMICONDUCTOR DEVICES FOR RECTIFICATION	2815
9.	COMPUTER PROGRAMS FOR STEAM POWER STATION CONTROL	3023
10.	REGULATIONS FOR TRANSPORTATION OF RADIOACTIVE MATERIALS	3309
11.	BAND DIAGRAM CONFIGURATION OF P-N JUNCTION IN THERMAL EQUILIBRIUM	3544
12.	BETA EMITTING ISOTOPES SR-90, Y-90	3688
13.	TYPHOID FEVER IMMUNIZATION VACCINES	3776
14.	HEAT CONDUCTION EQUATIONS FOR SUDDEN HEATING	3886
15.	FLOW FUNCTION CALCULATIONS FOR SUPERSONIC JETS	4018
16.	GOLDHABER EFFECT	4082
17.	DETERMINATION OF STRESSES IN DANGEROUS CROSS SECTIONS OF PARTS OF COMPLEX SHAPE BY THE METHOD OF PLANE SECTIONS	4148

APPENDIX B

QUERIES USED IN SECOND RADCOL EVALUATION

The following set of queries, drawn from specific areas of the abstract space, were used in the testing described in Section 3.

	QUERY	TARGET ABSTRACT
FREQUENT TOPIC QUERIES		
1.	LARGE SCREEN COMPUTER DISPLAYS	11
2.	PROGRESS IN SOVIET PUNCH CARD TECHNOLOGY	49
3.	ELECTRON STREAMS FORMED BY ELECTRICAL AND MAGNETIC FIELDS	197
4.	MEASUREMENT INSTRUMENTS IN ELECTRICAL COMMUNICATIONS	3354
5.	FREE SURFACE OF SINGLE CRYSTAL GE FILM PREPARED BY VACUUM EVAPORATION	3003
6.	LINEAR MODELS OF CRYSTAL OPTICAL AXIS STRUCTURE INTERFACE	3004
7.	ORIENTATION OF METAL OXIDE CRYSTALS	32
8.	VALIDATING MODELS OF SURFACE ELECTRONIC STRUCTURE BY LIGHT ABSORPTION AND PHOTOCONDUCTANCE	3041
9.	SOVIET AUTOMATIC CONTROL SYSTEM FOR EMERGENCY FIRST AID MEDICAL TREATMENT	4994
10.	AUTOMATIC PROGRAMMING, COMPOSING, AND PRINTING MACHINE	2085
INFREQUENT TOPIC QUERIES		
11.	SELFALIGNING BALL BEARINGS	65
12.	SOVIET TRACTOR AND SCRAPER PRODUCTION	92

MISSION
of
Rome Air Development Center

RADC plans and conducts research, exploratory and advanced development programs in command, control, and communications (C^3) activities, and in the C^3 areas of information sciences and intelligence. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.