VII. Information Retrieval: Dictionary Representations and Cluster Evaluation

P. C. Leech and R. C. Matlack, Jr.

Abstract

Representations of the Cranfield collection by both a thesaurus dictionary and a null, or word-stem dictionary, are subjected to Rocchio's clustering algorithm. The clusters generated are then evaluated using 42 search requests, and the results are compared to find optimum cluster sizes for both null and thesaurus dictionary representations of the collection. The thesaurus dictionary produces the highest recall and precision when the cluster size is approximately 5% of the collection size. The word stem dictionary produces comparable results only if the cluster size is increased to approximately 10% of the collection size. Deletion of low-weighted concepts from the cluster centroids of either dictionary representation does not affect appreciably the recall and precision values obtained.

1. Introduction

The prolific volume of literature published on a daily basis throughout the world makes it impossible for an individual to be aware of all innovations in his particular field of interest. The extensive amount of time involved in manual library searches necessitates automated document retrieval systems that are capable of retrieving information pertaining to a specific topic in a matter of minutes or, hopefully, even seconds. There are two basic problems, however, inherent to the design of such a system.

First, if the document collection is extremely large, comparison between the search request and each document in the collection is not practical because of the time taken to perform the search operations. Consequently, some means for categorizing, or "clustering", documents according to subject matter must be employed in an effort to maximize search efficiency. In the clustering procedures, such as that suggested by Rocchio [7], parameters are specified to control the number and size of clusters to be generated. These clusters are formed so that the similarity between the documents within a given group is much greater than that between documents contained in different clusters. The retrieval process then consists of a two-level search. The clusters are examined first to determine subject similarity, and a search is then made of the individual documents in those clusters selected on the basis of their relevance to the search request.

Second, numerous types of dictionaries may be used to assign word, or concept identification to the documents in the collection. [3] Two such dictionaries are the null (word stem) dictionary and the thesaurus dictionary.

The null dictionary consists of a set of word stems constructed from the words included in a typical document collection. Each distinct word stem is assigned a unique concept number, and retrieval is consequently based upon word-stem matching alone.

The thesaurus dictionary groups words of similar meaning into sub-groups or classes and assigns a concept number to the class as a whole rather than to the individual words or word stems. The thesaurus dictionary

recognizes synonyms and may therefore retrieve some documents which cannot be easily obtained by a word matching procedure alone. [6]

Application of both the null dictionary and the thesaurus dictionary to the same document collection results in two distinct representations of the collection. It is expected, then, that these two representations will require different input parameters to the clustering procedure in order to achieve maximum retrieval of relevant documents. It is of paramount concern to determine which parameters should be applied to each representation in order to maximize search efficiency and at the same time minimize the loss of relevant documents retrieved from the search.

The purpose of this report is to examine the above-mentioned problems by comparing the results of Rocchio's clustering procedure when applied to a document collection represented by both a null dictionary and a thesaurus dictionary.

2. Rocchio's Clustering Procedure

In an effort to jointly maximize search efficiency and minimize the loss of relevant documents retrieved in the search, Rocchio stipulates the following input parameters to his algorithm:

(1) the number of clusters desired;

(2) lower and upper bounds on the number of documents (represented by n-dimensional property vectors) to be included in a cluster; and

(3) a lower bound on the correlation (a similarity measure) between a document and a classification vector, below which a document will not be placed in a cluster.

The clustering algorithm proceeds as follows: an unclustered document is selected as a possible cluster center. All of the other unclustered documents are correlated with it, and the document is subjected to a density test to see if a cluster should be formed around it. This test specifies that more than N_1 documents should have correlations higher than p_1 with the document in question, and that more than N_2 documents should have correlations higher than p_2. This test ensures that documents on the edge of large groups do not become cluster centers. If the document passes the density test, a cut-off correlation, p_{min}, is determined from the cluster size limits and the distribution of correlation values. Documents with a correlation above p_{min} are placed above the cut-off; and if the correlations fall below p_{min} before the cluster size limit is exceeded, the cut-off is chosen at the greatest correlation difference between adjacent documents.

A classification vector is then formed by taking the centroid of all the document vectors presently included in the cluster. This centroid vector is matched against the entire collection, and the cut-off parameters for cluster size are reapplied to create an altered cluster.

As a result of this process, some documents may appear in more than one cluster; and some which were in a cluster when the centroid was originally formed may not remain in any cluster. These documents, as well as those which failed the density test, are termed "loose", and those within the cluster are termed "clustered".

This entire procedure is repeated with all unclustered documents, the first pass terminating when all items are either clustered or loose.

At this time, no guarantee exists that the established minimum number of clusters has been formed. Consequently, some documents which failed the density test in the first pass are chosen as cluster centers. If the number of clusters formed in the first pass is too high, the density test may be made stricter and the first pass repeated.

At the end of the clustering process, there still may be some loose documents. Although these documents correlate poorly with any of the centroid vectors, they may be assigned to those clusters with which they correlate most highly.

Rocchio explains [6,7] that if k centroids are found with a request correlation exceeding a chosen threshold value, and if each cluster contains an average of N/x documents, the total number of comparisons made with the search request for a two-level search will be

$$N_c = x + kN/x.$$

If the input parameters to the clustering procedure produce a number of clusters equal to $\sqrt{kN}$, the total number of comparisons, and hence the search time required, will be reduced to a minimum.

3. The Experiment

The basis for this experiment is the SMART automatic document retrieval system [1,2]. In this system, the documents and search requests are represented by n-dimensional property vectors. Each concept within a vector is weighted according to its frequency of occurrence in the document. Documents are retrieved by comparing the concepts contained in the document vector with those contained in the search request vector.

The experimental environment consists of two forms of the Cranfield collection: the Cranfield-Null (NULL) and the Cranfield-Thesaurus (THES). The NULL is produced by a word-stem identification process and consists of 200 documents relating to aeronautical engineering with 42 associated queries. The THES is produced by a synonym dictionary look-up procedure and consists of the same 200 documents and 42 queries.

These two document collections are then subjected to Rocchio's clustering algorithm. The input parameters controlling cluster size and number of clusters formed are varied. Concepts below a specified minimum weight are deleted from the centroid vectors in some cases. Results of the several runs are compared in terms of precision and recall[1], time requirements, etc.

Each computer run consists of cluster generation followed by a two-level search for each of the 42 queries. The cosine correlation function[2] is utilized to determine similarity among documents as well as between each document and the search requests. Only the three clusters which correlate most highly with the search requests are retained for evaluation. A comparison is also made between each search request and every document in the collection so that a two-level search may be compared with its corresponding full search.

---

[1] $\text{precision} = \dfrac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved}}$

$\text{recall} = \dfrac{\text{number of relevant documents retrieved}}{\text{number of relevant documents in the collection}}$

[2] Let d_1 and d_2 represent two document vectors. Then the cosine correlation function, S, is given by:

$$S_{d_1 d_2} = \frac{d_1 \cdot d_2}{|d_1| \quad |d_2|}$$

4. Results and Evaluation

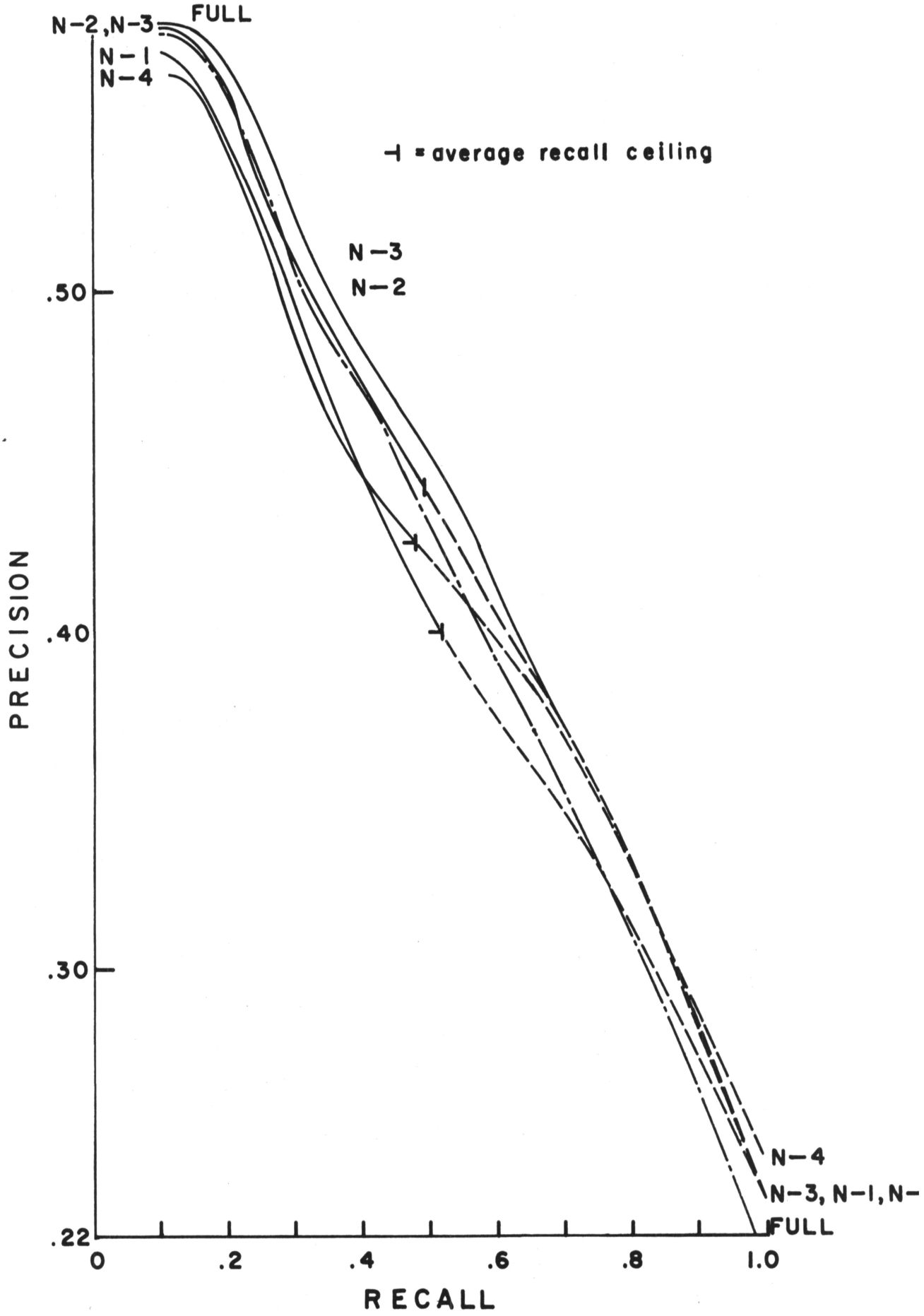
Eight computer runs were made: four on the NULL collection and four on the THES collection. The results of the eight runs are summarized in Table 1 and in the recall-precision graphs, Figs. 1-4. Although results from the evaluation of one, two, and three clusters are presented in Table 1, only the results from the evaluation of two clusters are presented in Figs. 1-4 for the sake of clarity. The results from two clusters were chosen because they produce a reasonable recall ceiling[1] of ~.5 and because they do not otherwise differ appreciably from the results of examining one or three clusters.

A) The Null Dictionary

An examination of Fig. 1 reveals that the maximum performance in terms of precision and recall is achieved with the null dictionary when the documents are grouped into large clusters. As the cluster size is diminished, a smaller number of relevant documents is retrievable when the precision value is kept constant. A possible explanation for this observation is that smaller clusters are represented in the SMART system by correspondingly smaller centroid vectors which become too specific, or too confined, with the use of a null dictionary. Documents are essentially clustered by a word-stem matching process, and this may sometimes lead to erroneous results. For example, the Cranfield collection contains both the terms "compressor" and "compressible flow". The two terms are entirely different in meaning, yet they have the same word-stem, so that documents

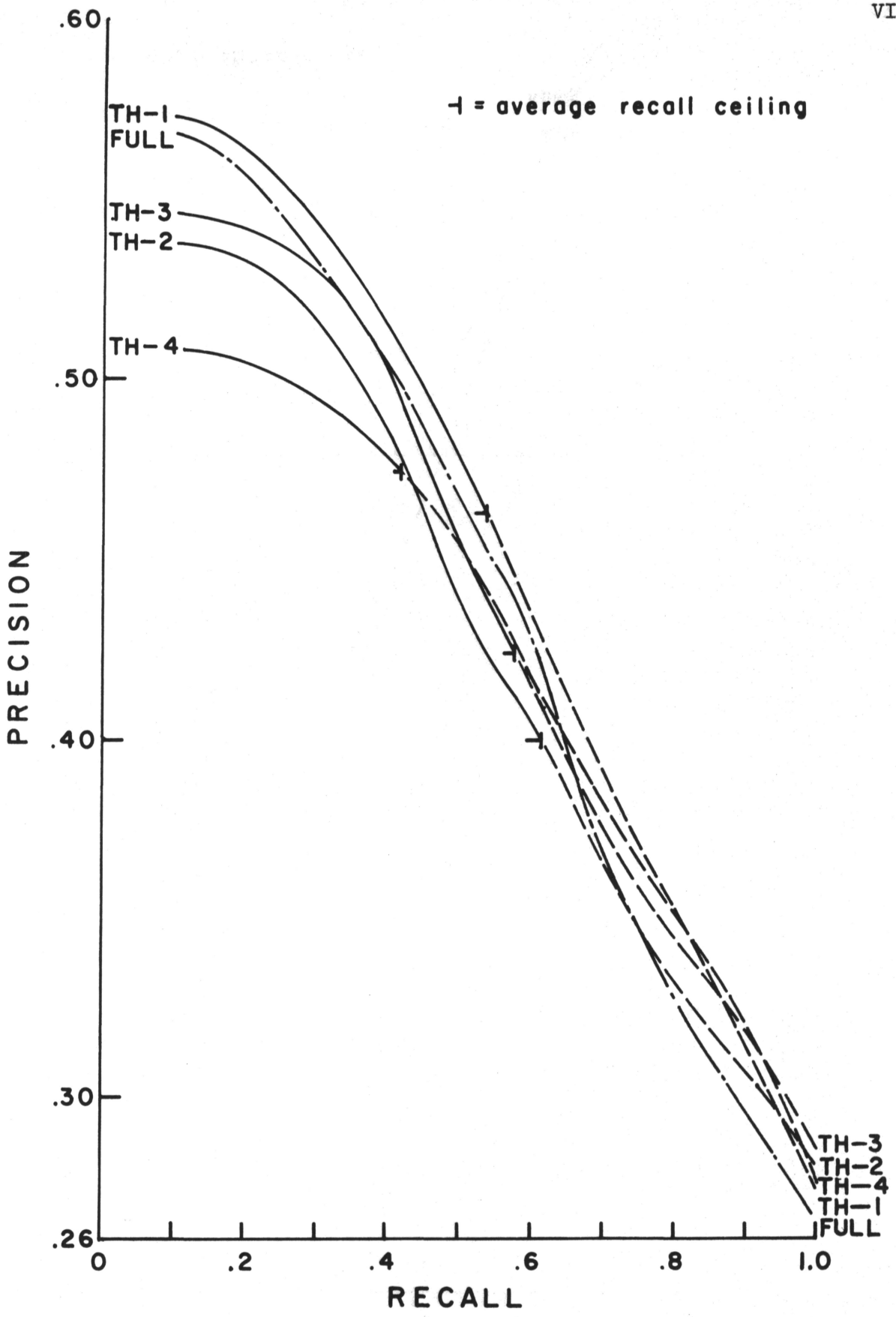
---

[1] recall ceiling = $\frac{\text{no. relevant documents able to be retrieved by searching the cluster(s) only}}{\text{no. relevant documents in the entire collection}}$



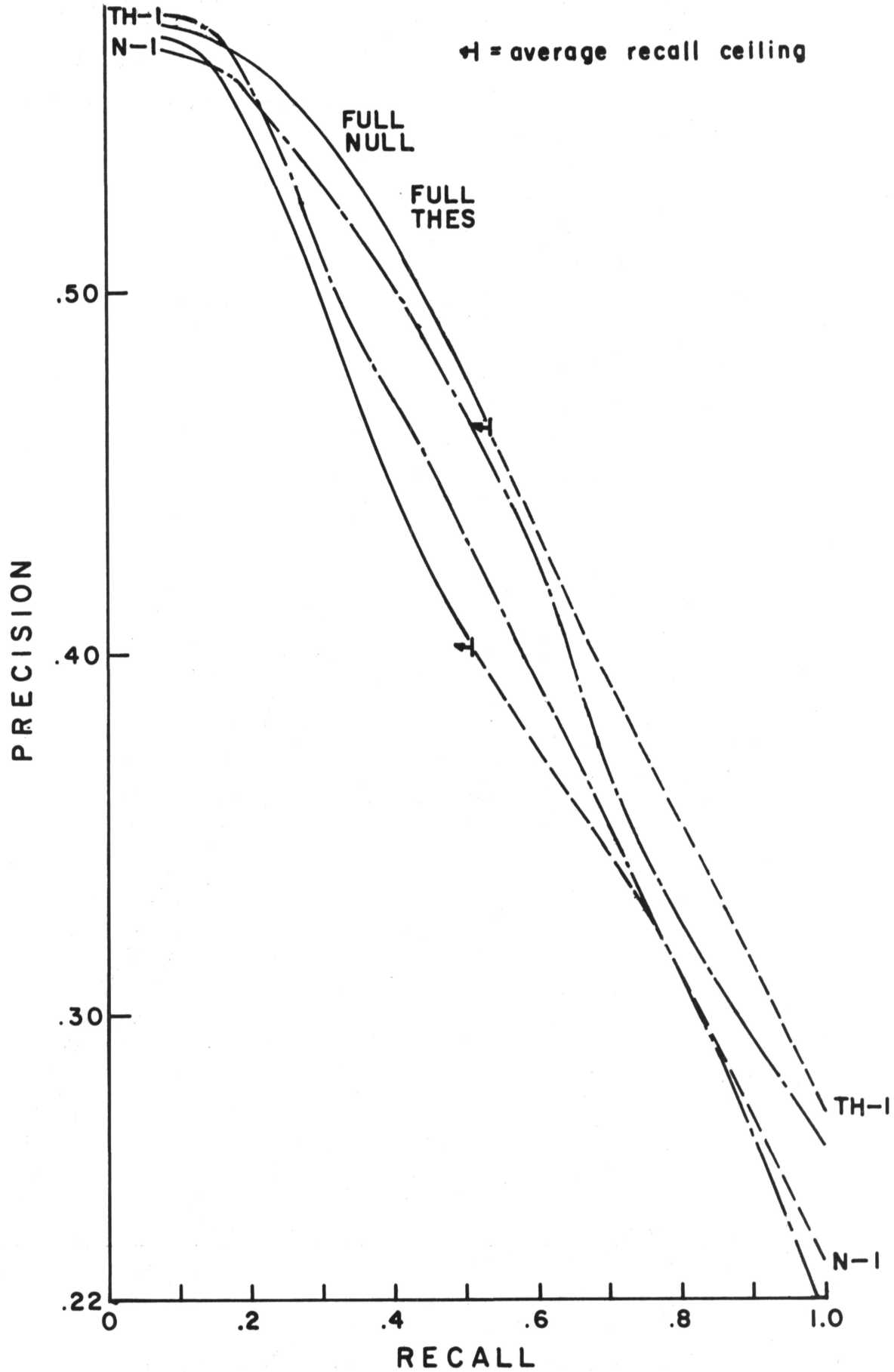
Recall-Precision Graph for NULL Dictionary

Fig. 1



Recall-Precision Graph for Thesaurus Dictionary

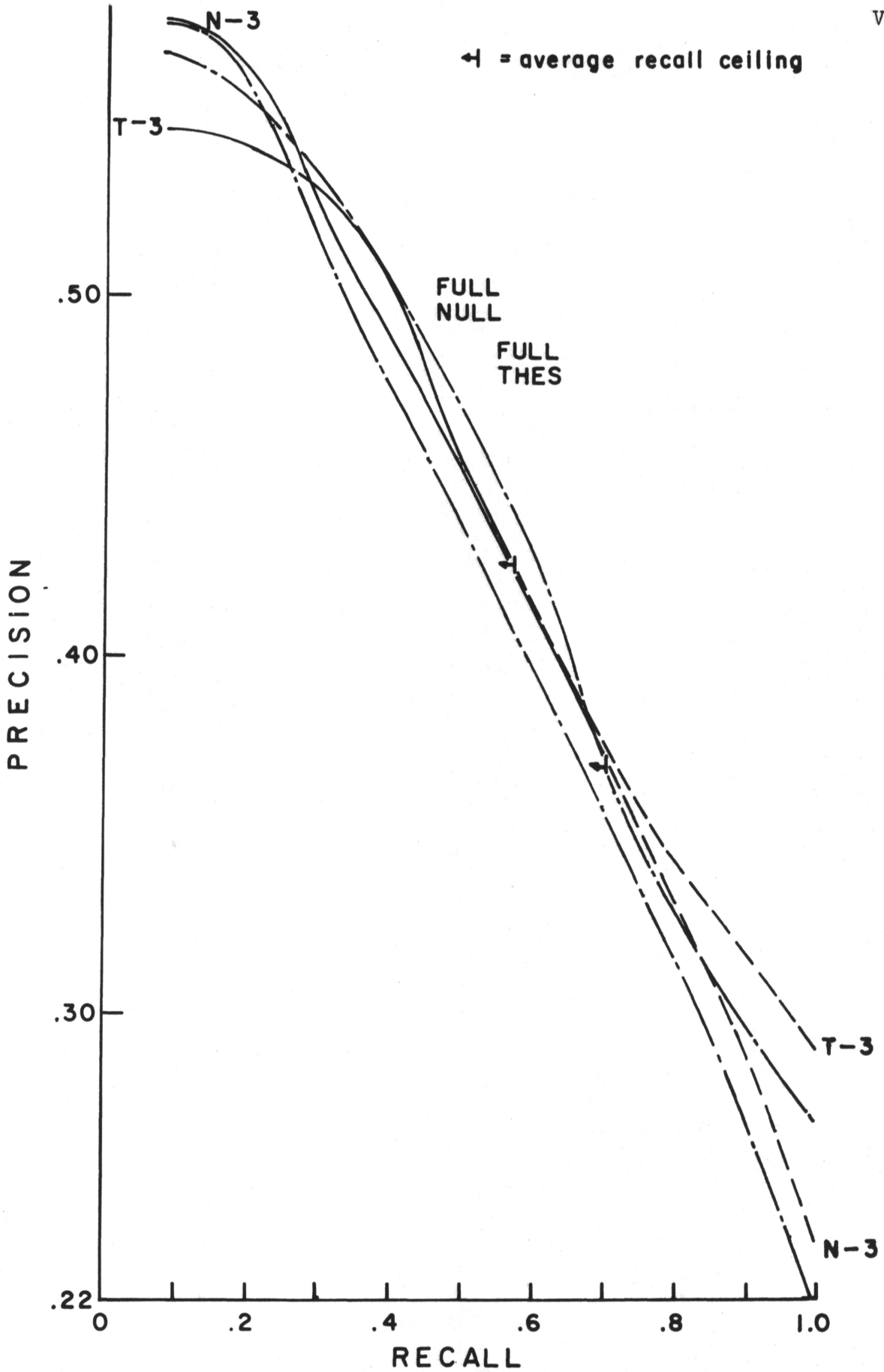
Fig. 2



Recall-Precision Graph for NULL-1 vs. THESAURUS-1

(small clusters, no concepts deleted)

Fig. 3



Recall-Precision Graph for NULL-3 vs. THESAURUS-3

(large clusters, concepts deleted)

Fig. 4

dealing with compressors and compressible flow could be incorrectly grouped together. Conversely, documents dealing with topics that are similar, but which have different word stems, will probably not be included in the same cluster. These disadvantages of the null dictionary are accentuated as the cluster size decreases, since the relevant documents to any given query are apt to be placed into one of several clusters which may or may not correlate highly with the given query.

The deletion of low-weighted concepts from the centroid vectors of the null dictionary appears to offer two important advantages. First, the centroid vectors occupy less space in the computer memory. This advantage applies particularly to centroid vectors of large clusters. In fact, one run was made with the null dictionary in which large clusters were to be generated, with no concepts deleted. The run resulted in 670 concepts associated with the first centroid vector, which exceeded the storage capacity of the computer for the vector.

The second advantage of low-weighted concept deletion is that, other factors being equal, the computer requires less time to match the query with the centroid vectors (since the centroid vectors are shorter), while the recall-precision performance is relatively unaffected. These results are illustrated in Table 1 and in Fig. 1 of Runs 1, 2, and 4 of the null dictionary. Run 1, for which no concepts were deleted, required more time than either of the other two runs, for which low-weighted concepts were deleted. In addition, the recall-precision performance of Run 4 is only slightly lower than that of Run 1, while the performance of Run 3 exceeds that of Run 1 considerably, and even compares favorably to the recall and precision attained from a search of the full collection.

B) The Thesaurus Dictionary

In contrast to the null dictionary results, Fig. 2 reveals that the thesaurus dictionary performs most effectively when documents are grouped into fairly small clusters. Run 1, which generated a cluster size of 9.8 documents per cluster, exhibits maximum recall-precision performance, superior to Runs 2 and 3 in which larger clusters were generated. However, Run 4, which generated a cluster size of only 6.6 documents per cluster, performs quite poorly in terms of recall and precision. One concludes that an optimum cluster size for the thesaurus dictionary is approximately 5% of the entire collection size.

That the thesaurus dictionary performs more favorably on small clusters than on large clusters is understandable. Unlike the null dictionary, words are assigned concept numbers according to their meaning, thus eliminating the possibility of faulty word groupings due to similar word stems. Also, each concept number is more general in nature, representing a group of synonyms rather than an isolated word stem. These characteristics of a thesaurus dictionary enable documents which are truly similar in content to be grouped effectively into small clusters. As the cluster size increases, however, the clusters must necessarily include documents which are not as similar in content, and the resulting centroid vectors will not be as indicative of the document content within the clusters. In order to retrieve the same number of relevant documents from a collection with large clusters, a proportionately higher number of irrelevant documents will be retrieved also, thus causing the recall-precision performance of large clusters to suffer.

The deletion of low-weighted concepts from the centroid vectors of the thesaurus dictionary appears to be advantageous, as it was with the null dictionary. A comparison of Runs 2 and 3 in Table 1 and Figure 2 indicates that although the deletion of low-weighted concepts causes no saving in evaluation time, the recall-precision performance of Run 3, in which the concepts were deleted, is superior to that of Run 2. That there is no saving in evaluation time is attributed to the fact that most of the weights of the concepts in the centroid vectors of a thesaurus dictionary are normally high, since the weight of each concept number represents the number of times that any of several synonyms appears. If concept numbers of greater weight were deleted, perhaps a time saving would be achieved, but one would then run the risk of losing some of the information contained in the centroid vectors, thus causing a loss in precision and recall.

C) Comparison of the Null and Thesaurus Dictionaries

Fig. 3 is a plot of the recall-precision curves from both a NULL run and a THES run, each of which generates small clusters with no concepts deleted from the centroid vectors. In addition, the recall-precision curve resulting from a full search is plotted for both collections. The two curves resulting from a full search indicate that the thesaurus dictionary produces better results than does the null dictionary for all but very low recall values. The null dictionary will produce a higher precision for low recall values because the possibility of false retrievals due to questionable synonym groupings is excluded [6]. However, as more documents are retrieved, the synonym groupings of the thesaurus dictionary become more effective, and relevant documents are able to be retrieved which are

not obtained from a word-stem matching process alone.

A comparison of the results for the two dictionaries after clustering supports the previous observation that small clusters produce better results when a thesaurus dictionary is used. A specific example of the superior performance of the thesaurus dictionary is shown in Table 2a. Seven documents are considered relevant to query No. 24. A search of the two most highly correlated clusters from the null dictionary retrieves only one of the seven relevant documents, resulting in low values of both precision and recall. The thesaurus dictionary, on the other hand, retrieves all seven relevant documents, producing higher values of both precision and recall.

The thesaurus dictionary is desirable for other reasons when small clusters are formed. A search of two clusters from the thesaurus collection yields an average recall ceiling of 0.5341, compared to 0.5180 for the null dictionary. This is attained by searching a slightly smaller percentage of the collection and taking much less time to do it (16 minutes for the thesaurus collection, 22 minutes for the null collection). The time difference is again due to the fact that both the documents and the centroid vectors are represented by fewer concept numbers when a thesaurus dictionary is used, so that the matching of the query with the centroid vectors and their associated documents is more readily accomplished.

The results of the null and thesaurus dictionaries for large clusters are shown in Fig. 4. As explained previously, large clusters tend to favor a null dictionary, and Fig. 4 reveals that the null dictionary yields better precision values for recall of less than 0.28. Table 2b illustrates this fact in reference to a search for documents considered relevant to

query No. 32. The null dictionary retrieves three of the four relevant documents if two clusters are searched, whereas the thesaurus dictionary retrieves only two, causing a decrease in the values of both precision and recall.

The clusters formed from the null dictionary also produce a higher recall ceiling. This is to be expected, as the thesaurus dictionary clusters are slightly smaller and are apt to exclude some of the relevant documents which may be included in the larger clusters of the null dictionary.

There are two serious drawbacks, however, to the use of the null dictionary even when large clusters are formed. If a recall of greater than 0.28 is desired by the user, Fig. 4 shows that the thesaurus dictionary will produce better precision values than will the null. In addition, the null dictionary requires more time to match the queries with the cluster centroids and their associated documents.

5. Conclusions

The optimum conditions for an effective cluster search are that a maximum number of relevant documents be retrieved (high recall), a minimum number of irrelevant documents be retrieved (high precision), and that the query-document matching procedure be as efficient as possible. When all of these factors are considered, the results of this report indicate that a generation of small clusters[1] from a thesaurus dictionary is the most desirable. Under these conditions, a search of the two most highly correlated clusters produces recall and precision values which equal or surpass

---

[1] cluster size approximately 5% of the collection size

those obtained from a search of the entire collection, whether the collection be represented by a null or a thesaurus dictionary. In addition, less than 10% of the entire collection is scanned, thereby reducing the time element of the query-document matching process considerably. In order to produce comparable values of precision and recall from a null dictionary, almost 20% of the entire collection must be scanned, and even then the precision suffers for any recall greater than 18%.

The results presented in this report are based upon limited amount of data, and should not be considered wholly conclusive. A continuation of this study, requiring further investigation of optimum cluster size for larger collections is essential, to determine the validity of the results presented.

References

[1] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System, Communications of ACM, Vol. 8, No. 6, June 1965.

[2] M. E. Lesk, The SMART System--Typical Processing Sequences, Report ISR-8 to the National Science Foundation, Section I, Harvard Computation Laboratory, December, 1961

[3] G. Salton and M. E. Lesk, Information Analysis and Dictionary Construction, Report ISR-11 to the National Science Foundation, June, 1966.

[4] J. D. Broffitt, et.al., On Some Clustering Techniques for Information Retrieval, Report ISR-11 to the National Science Foundation, June, 1966.

[5] G. Salton, et. al., Information Storage and Retrieval, Report ISR-9 to the National Science Foundation, Harvard Computational Laboratory, August, 1965.

[6] G. Salton, Automatic Information Retrieval, Manuscript, Cornell University, 1966.

[7] J. J. Rocchio, Harvard University Doctoral Thesis, Report ISR-10 to the National Science Foundation, April, 1966.

RUN	PARAMETERS		CLUSTER AND EVALUATION CHARACTERISTICS									TIME TO:	
	Concepts Deleted	#docs in cluster max, min	number clusters	mean no. docs per cluster	Per Cent Scanned 1 clust.	2 clust.	3 clust.	Ave. Recall Ceiling* 1 clust.	2 clust.	3 clust	generate	evaluate
THES-1	-	5,15	26	9.8	4.9	9.6	14.5	.3547	.5341	.6011	48 min	16 min
THES-2	-	10,20	18	15.7	7.7	15.5	23.7	.4249	.6188	.6675	36 min	16 min
THES-3	≦ 12	10,20	18	16.1	7.8	15.8	23.7	.4350	.5781	.6429	36 min	16 min
THES-4	≦ 12	5,10	30	6.6	4.5	8.6	12.8	.2993	.4556	.5327	50 min	15 min
NULL-1	-	5,15	23	10.1	5.2	10.6	15.9	.3769	.5180	.6227	64 min	22 min
NULL-2	≦ 1% t. wt.	5,15	33	7.1	4.0	7.3	11.1	.3647	.4931	.5971	65 min	20 min
NULL-3	≦ 12	10,20	14	18.1	9.4	19.7	29.3	.4338	.7042	.7638	45 min	18 min
NULL-4	≦ 12	5,10	26	9.6	4.2	9.1	13.3	.3211	.4822	.5579	64 min	18 min

* recall ceiling is defined as the fraction of of relevant documents able to be retrieved by searching the cluster(s) only

Comparison of Thesaurus Clusters with Stem Clusters

Table 1

a. Query 24: Has anyone derived simplified pump design equations . . ?

Relevant documents: 142, 143, 144, 145, 146, 147, 148

NULL-1		THES-1	
Rank	Document	Rank	Document
1R	145	1R	145
2	167	2R	146
3	85	3R	144
4	164	4	34
5	68	5R	142
6	66	6	38
7	14	7	35
8	87	8	37
9	102	9	14
10	165	10R	147
11	105	11	101
12	25	12R	143
13	26	13R	148
.	.	.	.
.	.	.	.
23	60	23	70
Precision	.0435	Precision	.3043
Recall	.1429	Recall	1.0000

b. Query 32: What is the solution of the Blasius problem with three point boundary conditions . . ?

Relevant Documents: 5, 6, 7, 26

NULL-3		THES-3	
Rank	Document	Rank	Document
1R	5	1R	5
2R	7	2	16
3R	6	3	180
4	139	4R	7
.	.	.	.
.	.	.	.
20	18	20	72
Precision	.1500	Precision	.1000
Recall	.7500	Recall	.5000

* R labels indicate relevance to the given query

Specific Examples of Document Retrieval from a Two-Cluster Search

Table 2