VI. An Evaluation of Rocchio's Clustering Algorithm

Robert T. Grauer and Michel Messier

Abstract

This report evaluates the success of a clustering algorithm developed by J. J. Rocchio. It contains a description of the algorithm which was obtained from Information Storage and Retrieval Report ISR-10. [1] Also included is a list of the parameters required by Rocchio's algorithm and a discussion of problems related to it. Results of sixteen computer runs are presented in tabular form. These runs are subdivided into smaller groups (according to the values of the input parameters) for the purpose of detailed analysis. Complete discussions are given for six of these groups. Precision-recall plots are included for each of the above mentioned studies.

1. Introduction

The basic classification problem consists in subdividing a given set of documents into a reasonable number of smaller sets in such a way that the documents within each subgroup are sufficiently similar to justify ignoring the individual differences between them. That is, one seeks to be able to represent each individual document of a subgroup or cluster by a typical document of that cluster. Document clusters are usually formed to facilitate the matching process between search requests and the given document collection. One seeks to maximize search efficiency and simultaneously minimize the loss of relevant documents retrieved in the search. The objective of this project is to evaluate the success of a clustering algorithm developed by J. J. Rocchio through use of a program written by Robert Williamson.

2. A Description of Rocchio's Algorithm

In Rocchio's algorithm all items are first considered as unclustered and from here they pass into one of two states — clustered or loose. An unclustered item is selected as a possible cluster center and subjected to a density test which requires that at least N_1 documents have a correlation of at least p_1 , and that at least N_2 documents have a correlation of at least p_2 with the document in question. This insures that items on the edge of large groups are not centers of groups, nor are annular regions accepted as clusters.

If a given document fails to pass the density test, it is considered loose, and another unclustered document is considered as a potential center. If a document does pass the density test, a cut-off correlation p_{min} is determined as a function of the category size limits and the correlation distribution. Documents with a correlation above p_{min} are automatically placed above the cut-off. If correlations fall below p_{min} before the 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 items belonging to the cluster at this time. The centroid is matched against the entire collection and the cut-off parameters on category size are reapplied to create an altered cluster. Three things may happen:

1) The correlation coefficient of the item may exceed the cut-off value, but may at the same time be below p_{min} . Such a document is marked loose to prevent its subsequent choice as a possible cluster center.

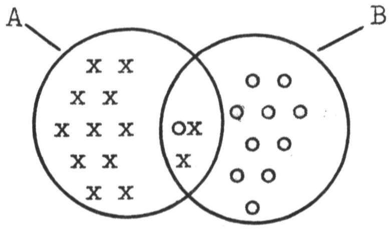
2) Documents with a correlation coefficient below the cut-off are unclustered.

3) Items above p_{min} are clustered.

This process is repeated with all unclustered items until every document is either clustered or loose. It is quite possible for a document to end up in two or more clusters since the centroid vector is matched against every document in the collection.

Since there exists no simple way of determining in advance how many categories will be formed in this manner, a second pass may be made which will alter the number of clusters formed by altering the density test parameters.

A third pass is used to cluster any documents which are still unclustered after passes one and two. One option is to place these documents in the cluster with whose centroid vector they correlate best. The third pass also redefines the centroid vectors and resulting clusters via a correlation partition routine which reduces the number of duplications. This routine seeks to eliminate the following type of situation illustrated in the diagram:



The two x's in the intersection correlate higher with centroid B but have been initially assigned to centroid A. The partition routine reassigns them to centroid B. Similarly for the o in the intersection.

The routine places all documents in the region of intersection into the appropriate centroid thus redefining the centroid vector. These new centroids are then matched against all documents in the collection to yield new clusters.

3. Problems to be Studied

In the present section, the required input parameters for the program are detailed, and several relevant questions relating to a clustering methodology are treated. Because of restrictions on available computer time, results are not available for all questions on the list. However, it is felt that in a complete report, every parameter mentioned here merits consideration, and any complete report on clustering requires answers to all questions. The authors feel that subsequent work should be undertaken to supplement the present report. The program requires specification of the following parameters:

1) Cluster generation starting point.
2) Minimum and maximum number of documents per cluster.
3) Lower bound on the correlation between an item and a cluster classification vector below which an item will not be placed in a given cluster.
4) Removal of low weight concepts from centroid vector.
5) Density test parameters.
6) Cut-off criteria.
7) Blending.
8) Correlation coefficient used.

Several other problems are also of interest. In evaluating the results of a cluster generation, the question arises of the number of clusters to be examined by the user. Examining all the documents in the three most relevant clusters will undoubtedly yield higher recall than examining only those in the single most relevant cluster, but where is the point of diminishing returns? Similarly are better results to be expected if one examines ten documents in the cluster which correlates most highly with the query or five documents in each of the two most highly correlated clusters?

There also arises the question of overlap between clusters. Are better results achieved with substantial duplication (the same document appearing in more than one cluster) or with a minimum of duplication?

Also of interest is the question of the cut-off criteria. Should cut-off occur at the greatest difference in correlation coefficients, or after the first difference between coefficients which exceeds p, or after the first n documents?

Documents clustered after the first pass are not considered in subsequent density tests. Because of this, changing the permutations of the documents may yield different clusters. If, however, all documents are considered in all density tests, the order in which documents are examined would have considerably less influence. This procedure would increase the number of clusters obtained, since each potential cluster center then has more available documents with which to apply the density test, thus resulting in higher probability of passing the test.

The program allows documents which are initially unclustered to be blended into the cluster with whose centroid vector they correlate best. However, the best correlation may be as low as .1000 in some cases. It is felt that these low correlated documents have a detrimental effect on precision. The effect of not blending all such documents, or not blending a percentage of such documents merits consideration.

After establishment of a set of "optimal" values for these parameters using the ADI collection, the same "optimal" values might be run with the Cranfield collection to see if optimal results are produced here as well; i.e. it is desired to determine whether an "optimal" set of parameters for one collection is also "optimal" for another collection.

It is felt that one set of parameters will be "optimal" for the user seeking high precision and an entirely different set "optimal" for the user seeking high recall.

4. Evaluation Scheme

Some fundamental definitions are required.

$$\text{Recall} = \frac{\text{No. of relevant documents retrieved}}{\text{No. of relevant documents in collection}}$$

$$\text{Precision} = \frac{\text{No. of relevant documents retrieved}}{\text{Total no. of documents retrieved}}$$

$$\text{User Percentage Scanned} = \frac{\text{No. of documents scanned by user}}{\text{Total no. of documents in collection}}$$

$$\text{Machine Percentage Scanned} = \frac{(\text{No. of documents scanned by machine + No. of clusters})}{\text{Total no. of documents in collection}}$$

Recall ceiling = the maximum recall that can be obtained by searching all the documents in a cluster

Several methods exist for any proposed evaluation. A precision-recall plot is useful because it shows which run yields the best results. It also affords a direct comparison with the results of a full search. It is deficient in that it does not show effects of the percentage of the collection scanned by the machine.

A plot of machine percentage of collection scanned versus recall ceiling is deficient in that it neglects effects of the user percentage of the collection scanned. It also observes the actual search results. For example, if run A yields a recall ceiling of .2500 and the machine scans 30% of the collection whereas run B yields a recall ceiling of .3500 and the machine scans 40% of the collection, it is unclear which run is actually better. This method is also deficient in that comparisons with a full search are meaningless here. A plot of user percentage of collection scanned versus recall ceiling poses similar difficulties. Ideally a three dimensional plot could be used, but the difficulties of construction eliminate it from consideration. An alternative is the computation of a "value" of the run according to:

$$\text{Value} = \alpha\,(\text{Recall Ceiling}) + \beta\,(\text{User Percentage}) + \gamma\,(\text{Machine Percentage})$$

where the values of α, β, and γ are arbitrarily determined. Different users would require different sets of coefficients. For example, a user who requires high recall and is unconcerned about the proportion of the collection to be scanned might use $\alpha = .8$ and $\beta = .2$ etc.

Since it is difficult to get a handle on suitable values of α, β, and γ, it is proposed to do an analogous type of evaluation in which three different sets of optimum parameters will be determined for different user and machine requirements:

> One set to yield maximum recall ceiling given that the percentage the user will scan is at most 20%, and that the machine is to scan at most 40%.
>
> A second set to minimize the percentage the user will scan (disregarding the percentage the machine is to scan) at a recall ceiling of at least .4000.

A third set to minimize the percentage the machine will scan (disregarding the percentage the user will scan) at a recall ceiling of at least .2500.

A further quantity to be examined when evaluating output is the time used by the computer for cluster generation.

Note that in the evaluation section of this report all precision-recall plots are given collectively at the conclusion of section B. These plots are broken at one, two, and three clusters. The graphs for two and three clusters could be extrapolated back to low values of recall, but such extrapolation would only clutter the plot, making its reading difficult.

A) Tabulation of Results

This section contains the complete tabulation of machine results for ADI runs one to sixteen inclusive. Six tables of data are presented in the following order:

Table 1	Cluster Parameters for runs 1 to 8.
Table 2	Cluster Parameters for runs 9 to 16.
Table 3	Cluster Characteristics for runs 1 to 8.
Table 4	Cluster Characteristics for runs 9 to 16.
Table 5	Evaluation parameters of user percentage and machine percentage for runs 1 to 16.
Table 6	Evaluation parameters of recall ceiling and machine time required for cluster generation for runs 1 to 16.

For the complete evaluation of these tables, the reader is referred to section B where detailed studies are made of selected groups of runs. The studies are listed below. The complete tables of results are included as follows:

Study 1 Many clusters with few documents per cluster versus few clusters with many documents per cluster.

Study 2 Consideration of entire collection in all density tests.

Study 3 Variation of density test parameters.

Study 4 Deletion of low weighted concepts from centroid vector.

Study 5 Redefinition of the centroid vector.

Study 6 Deletion of the maximum correlation partition routine.

Cluster Parameters	Run Number							
	1	2	3	4	5	6	7	8
1. Amt. of collection used in clustering	The entire collection of 82 documents. A constant parameter in all runs.							
2. First document used in clustering	Document No. 1. A constant parameter in all runs.							
3. Density Test	5 ≧ .25, 10 ≧ .15 A constant parameter in all runs.							
4. Minimum and maximum no. of docs/clus.	5,15	5,15	0,-	5,15	5,15	5,15	2,15	2,15
5. Correlation coefficient	Cosine coefficient.							
6. Difference in correlation to force break	Maximum difference between adjacent docs. A constant parameter in all runs.							
7. Is centroid vector redefined?	No. A constant parameter in all runs.							
8. Concepts deleted in centroid vector	–	Wt. ≦12	–	Wt ≦24	–	Wt ≦1%	–	–
9. Is max. correlation partioning used?	Yes	Yes	No	Yes	No	Yes	Yes	Yes
10. Are loose documents blended?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
11. Is entire collection used in all density tests?	No	No	No	No	No	No	No	Yes

Cluster Parameters for Runs 1 to 8 Inclusive

Table 1

Cluster Parameters	Run Number							
	9	10	11	12	13	14	15	16
1. Amt. of collection used in clustering	The entire collection of 82 documents. A constant parameter in all runs.							
2. First document used in clustering	Document No. 1. A constant parameter in all runs.							
3. Density test	5 ≧ .25,- 10 ≧ .15 Constant for runs 9-14						10 ≧ .25 20 ≧ .15	5 ≧ .40 10 ≧ .20
4. Minimum and maximum no. of docs/clus.	5,15	7,15	5,15	5,15	7,15	10,15	2,15	2,15
5. Correlation coefficient	Cosine coefficient							
6. Difference in correlation to force break	Maximum difference between adjacent docs. A constant parameter in all runs.							
7. Is centroid vector redefined?	No	No	No	Yes	No	No	No	No
8. Concepts deleted in centroid vector	–	–	Wt ≦36	–	–	–	–	–
9. Is max. correlation partioning used?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
10. Are loose documents blended?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
11. Is entire collection used in all density tests?	Yes	No	No	No	Yes	Yes	Yes	Yes

Cluster Parameters for Runs 9 to 16 Inclusive

Table 2

Cluster Characteristics	Run Number							
	1	2	3	4	5	6	7	8
1. Collection size	82	82	82	82	82	82	82	82
2. Number unclustered before blending	34	32	48	31	27	34	43	29
3. Number clustered before blending	48	50	34	51	55	48	39	53
4. Number duplications	5	8	0	5	8	4	3	28
5. Number clusters	7	8	34	7	7	7	14	28
6. Mean documents per cluster	12.4	11.3	2.4	12.4	12.9	12.3	6.1	3.9
7. Mean documents per cluster without blended documents	7.6	7.3	1.0	8.0	7.9	7.6	3.0	2.9

Cluster Characteristics for Runs 1 to 8

Table 3

Cluster Characteristics	Run Number							
	9	10	11	12	13	14	15	16
1. Collection Size	82	82	82	82	82	82	82	82
2. Number unclustered before blending	22	26	31	30	19	17	41	54
3. Number clustered before blending	60	56	51	52	63	65	41	28
4. Number duplications	57	6	3	6	34	51	18	4
5. Number clusters	17	7	7	7	11	10	19	8
6. Mean documents per cluster	8.2	12.5	12.1	12.6	10.5	13.7	5.3	10.7
7. Mean documents per cluster without blended documents	6.9	8.9	7.7	8.3	8.8	11.6	2.2	3.5

Cluster Characteristics for Runs 9 to 16

Table 4

Run No.	Machine Percentage			User Percentage		
	1	2	3	1	2	3
1	26.1	42.0	57.1	17.5	33.4	48.5
2	24.4	38.2	51.8	14.4	28.4	42.1
3	44.5	47.4	50.0	3.0	5.8	8.5
4	26.5	43.2	60.5	17.9	34.7	52.1
5	26.0	43.2	59.0	17.5	34.6	50.5
6	25.8	41.4	56.2	16.2	31.8	46.3
7	25.2	33.6	41.2	8.2	16.5	24.1
8	40.0	46.0	51.7	5.9	11.8	17.5
9	32.3	43.6	52.8	11.6	23.0	33.4
10	25.2	40.6	55.4	16.7	32.1	46.8
11	26.3	42.7	59.6	17.8	34.3	51.0
12	26.3	41.1	58.2	17.1	32.6	49.5
13	25.9	38.1	49.8	12.5	24.7	36.3
14	28.5	45.0	61.2	16.3	32.8	49.1
15	31.2	38.9	47.0	8.1	15.8	23.8
16	25.7	42.0	54.3	17.2	32.3	44.6

Evaluation Parameters of User Percentage
and Machine Percentage for Runs 1 to 16

Table 5

Run No.	Average Recall Ceiling			Machine Time (mins.secs)
	1	2	3	
1	.3117	.4656	.6434	6.52
2	.2991	.3955	.5733	7.15
3	.1573	.2568	.3192	14.37
4	.3431	.5035	.6692	7.01
5	.2595	.4880	.6517	6.28
6	.2990	.4727	.6259	7.20
7	.2007	.3725	.5129	10.58
8	.2097	.3817	.4305	14.56
9	.3079	.4546	.5920	9.02
10	.3163	.4546	.6358	6.38
11	.3085	.4912	.6317	6.41
12	.2466	.4871	.6571	7.55
13	.3565	.5033	.5825	8.23
14	.3843	.5405	.6074	8.30
15	.2237	.3384	.3833	15.17
16	.2941	.5194	.6223	11.38

Evaluation Parameters of Recall Ceiling
and Machine Time Required for Cluster
Generation for Runs 1 to 16

Table 6

B) Detailed Analysis

Study 1: To determine if superior results are obtained with many clusters of relatively few documents for fewer clusters with relatively many documents. The problem is attacked in run nos. 1, 3, 7, and 10 in which the input parameter denoting the minimum number of documents per cluster is varied. All other parameters are kept constant.

This particular study serves as an illustrative example of the method of evaluation employed in all other studies. The manner in which conclusions are reached for the three different user specifications is shown in detail together with relevant result data extracted from the complete tables found in the preceding section. For the other studies only the conclusions are shown.

Example: The data given below are used to determine the run which yields a maximum recall ceiling given that the percentage the user will scan is at most 20% and that the machine is to scan at most 40%.

Run No.	Machine Scanning			User Scanning		
	1	2	3	1	2	3
3	44.5%	47.4%	50.0%	3.0%	5.8%	8.5%
7	25.2%	33.6%	41.2%	8.2%	16.5%	24.1%
1	26.1%	42.0%	57.1%	17.5%	33.4%	48.5%
10	25.2%	40.6%	55.4%	16.7%	32.1%	46.8%

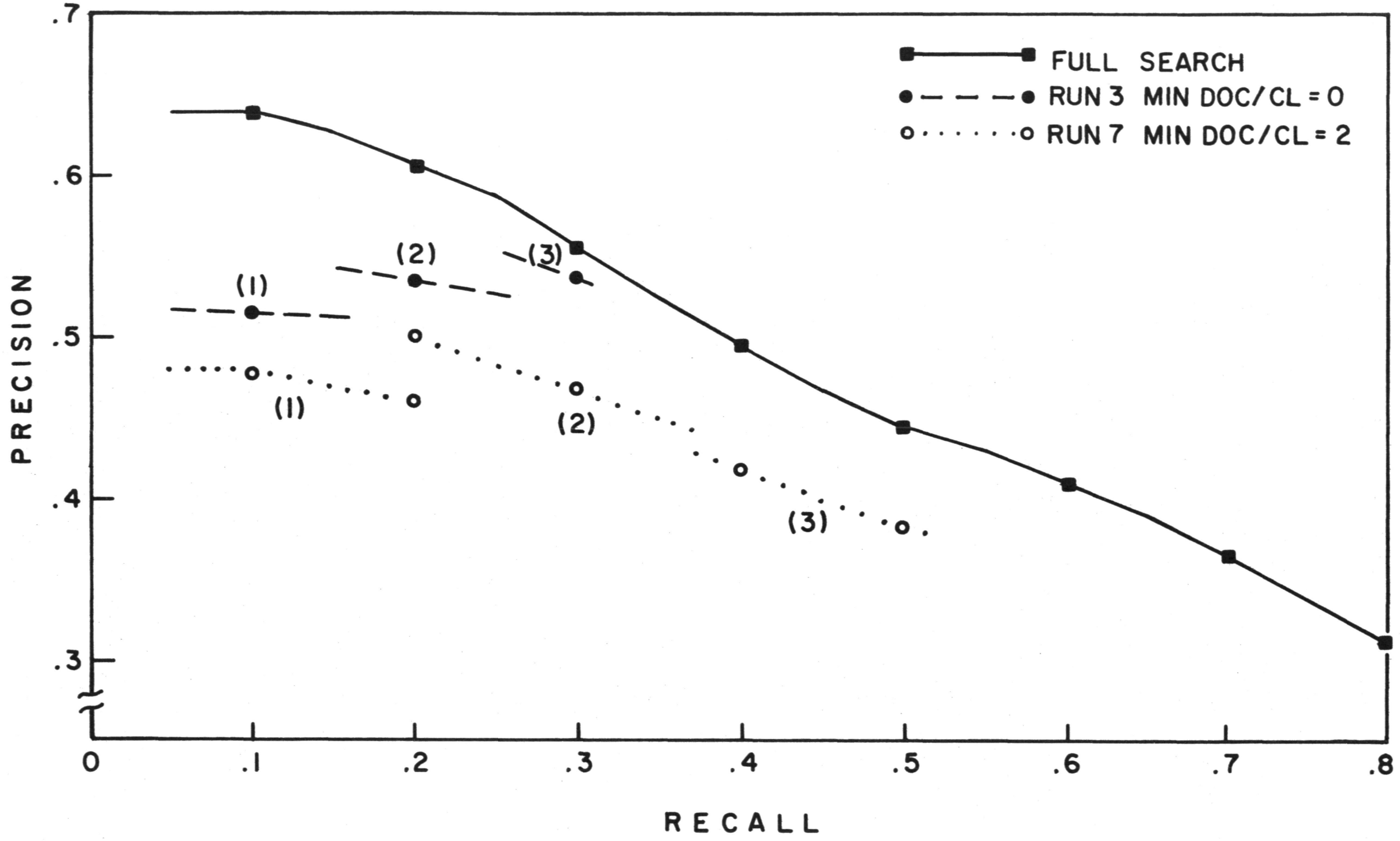
Run No.	Min. doc/clus	Recall Ceiling		
		1	2	3
3	0	.1573	.2568	.3192
7	2	.2007	.3725	.5129
1	5	.3117	.4056	.6434
10	7	.3163	.4546	.6358

Any of the percentages in the set of user percentages enclosed in the dashed lines meets the required specification; similarly for the machine percentages also shown in dashed lines. The intersection of these two sets is the smaller set drawn in dotted lines from which one selects a maximum recall ceiling of .3725 obtained from scanning two clusters in run seven.

Similarly in meeting the second set of specifications (i.e. to minimize the percentage the user will scan and to yield a recall ceiling of at least .4000) one would scan the documents in the three most relevant clusters of run seven obtaining an average recall ceiling of .5129 with the user scanning 24.1%. For the third set of specifications (to minimize the percentage the machine will scan and still yield an average recall ceiling of at least .2500) one would search the documents in the first cluster of run ten. One obtains an average recall ceiling of .3163 and the machine scans 25.2% of the collection. Note the results of one cluster of run one are practically identical.

Conclusions:

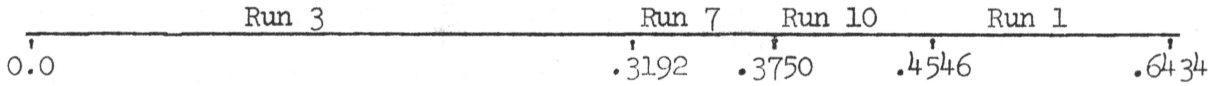
1. Based on the precision recall plot of Fig. 1, one may construct the following graph in which the run yielding the optimum precision recall plot is typed above the appropriate range.



Study 1 The Effect of Average Cluster Size on Recall and Precision

(n) number of clusters examined

Fig. 1



Run three is clearly best up to its recall ceiling. Run seven is best to a recall ceiling of .3750. From this point there is little difference between all runs valid over an appropriate range. It is not surprising that run three yields the best precision recall plot in its range. This run contains a very large number of clusters (34) and very few documents per cluster (2.4). One would guess intuitively that such an arrangement would yield comparatively high precision values since each retrieved cluster contains very few documents each of which correlate highly with the cluster centroid and thus correlate highly with the particular query. However, since so few documents are found in each retrieved cluster, one expects the recall ceilings to be very low.

2. For the user interested in a low user percentage, many clusters with few documents give better results. This is an obvious conclusion since fewer documents per cluster retrieved means fewer documents to be searched by the user.

3. For the user interested in a low machine percentage, many clusters with few documents per cluster also appears to be best. However, there should not be as many clusters as for a low user percentage, for in the latter case, the large number of clusters considerably increases the machine percentage.

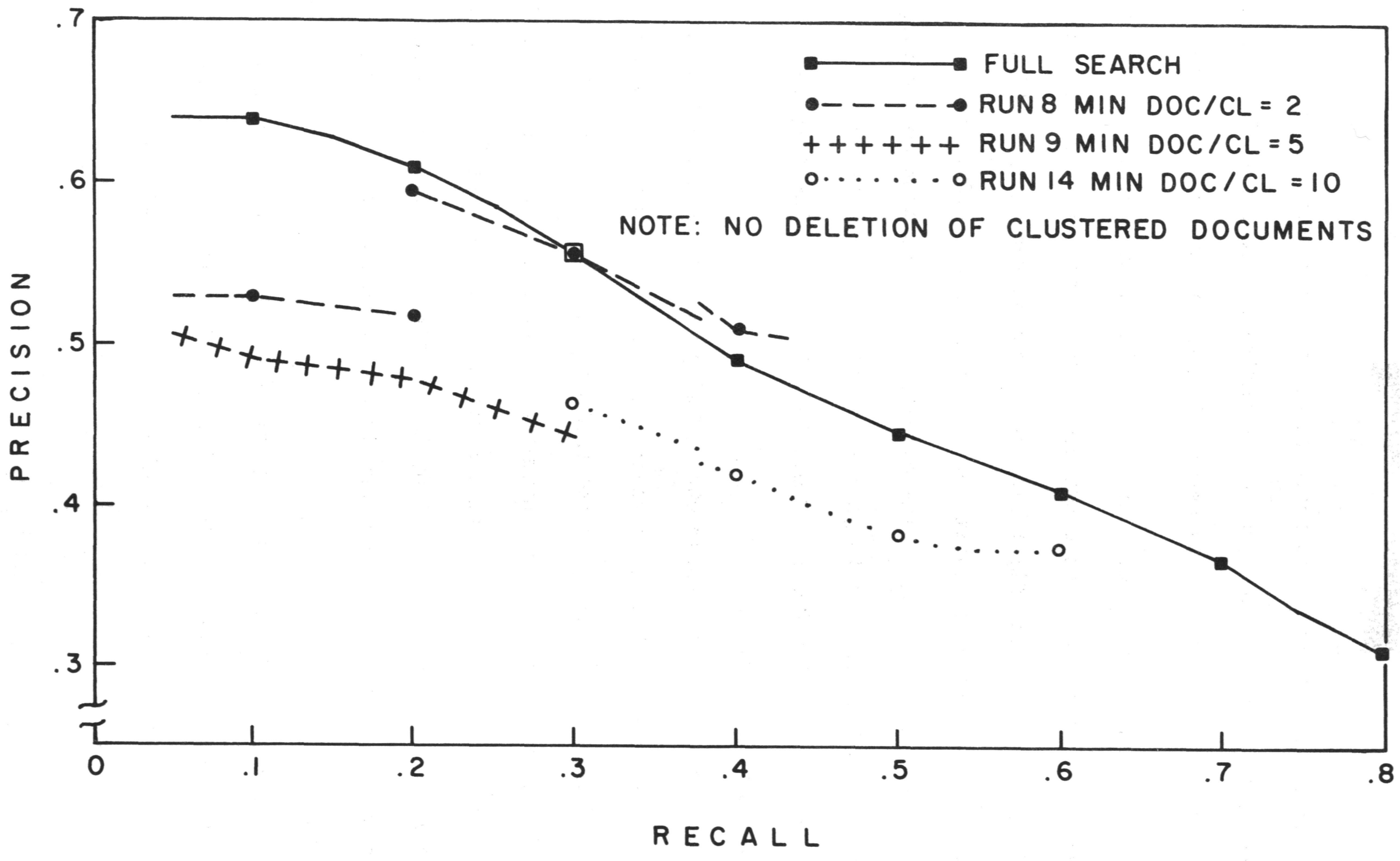
4. The user interested only in high recall requires many documents per cluster. Both his percentages are high since more documents are being searched. Runs one and ten appear equivalent.

5. Apparently precision recall plots improve with an increase in the number of clusters. Unfortunately generating more clusters increases the amount of computer time required. This problem is treated in study 3.

Run No.	No. of Clusters	Time Required (mins/secs)
3	34	14.37
7	14	10.58
1	7	6.52
10	7	6.38

Study 2: To determine whether it is better to use the entire collection for all density tests than to delete those documents already clustered in all subsequent density tests. The study consists of comparisons between groups of runs as follows: runs (1,9), runs (7,8), and runs (10,13). The only difference in each pair is that the second run of each pair considers all eighty-two documents for the density test whereas the first run deletes clustered documents. The minimum number of documents required per cluster is different in each of the three pairs.

Runs 8, 9, 13, and 14 collectively may serve as a supplement to the question of minimum number of documents required per cluster considered in the previous study. In the latter group of runs the minimum number of documents required per cluster is respectively 2, 5, 7, and 10 but each run requires that all eighty-two documents be used in all density tests. The output is shown in Fig. 2.



Study 2 The Effect of Average Cluster Size on Recall and Precision

Fig. 2

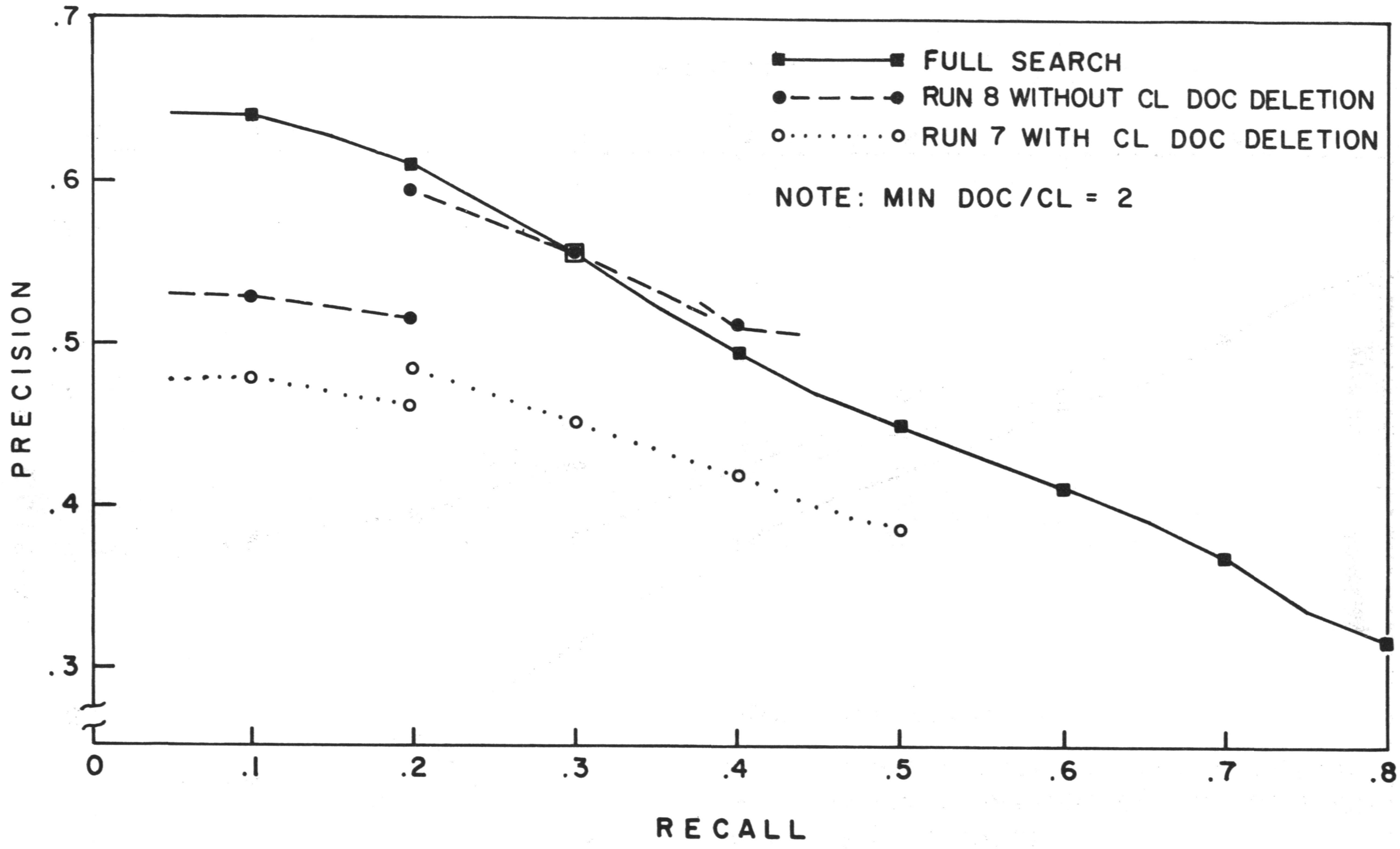
Conclusions:

1. Based on the three precision recall plots of Figs. 3, 4, and 5 (one for each pair) one sees that using the entire collection yields results at least as good, and in most cases better than those runs which delete documents. (In each of the three plots, the dashed line indicates the run in which the entire collection is considered for all density tests.) The only exceptions occur for the first cluster in graphs for (10,13) and (1,9) Figs. 4 and 5. These results deviate by a maximum of only .01. Thus the writers feel justified in reaching the overall conclusion that considering all documents in all density tests yields decidedly better precision-recall plots. Run 8 is especially worthy of mention since it is the only run which surpasses a full search.

These results are not unexpected. Using the entire collection in all density tests will obviously increase the number of clusters and correspondingly decrease the number of documents per cluster. As was explained in the previous study, this situation is conducive to high precision.

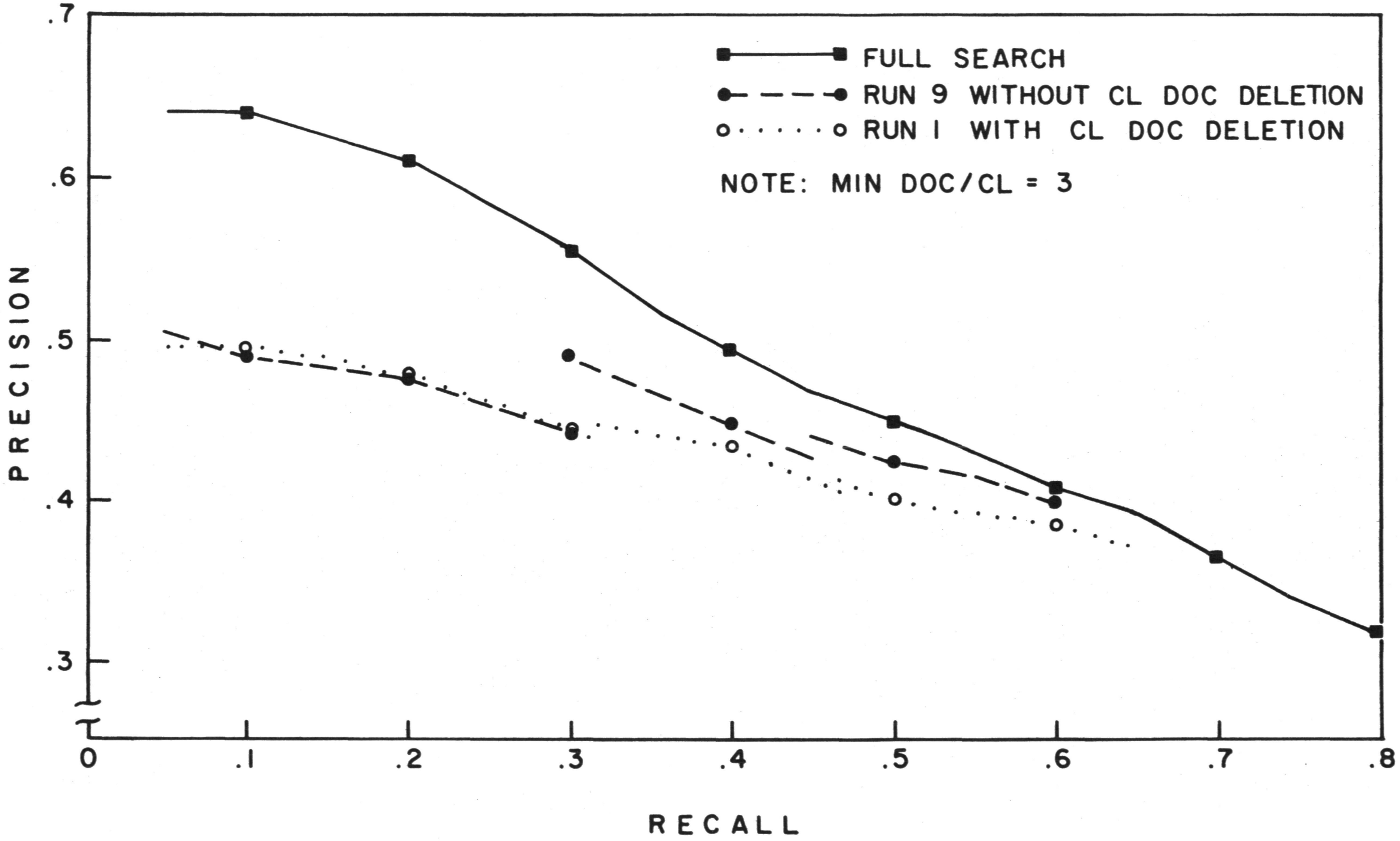
2. Based on the single precision recall plot for runs 8, 9, 13, and 14, the conclusion of the previous study is substantiated; i.e. the runs with the greatest number of clusters and fewest documents per cluster yield the best plots.

3. Conclusions may be drawn by comparing the average recall ceiling with the user percentage scanned for each pair. (For each pair the run which considers the entire collection is designated with an asterisk.)



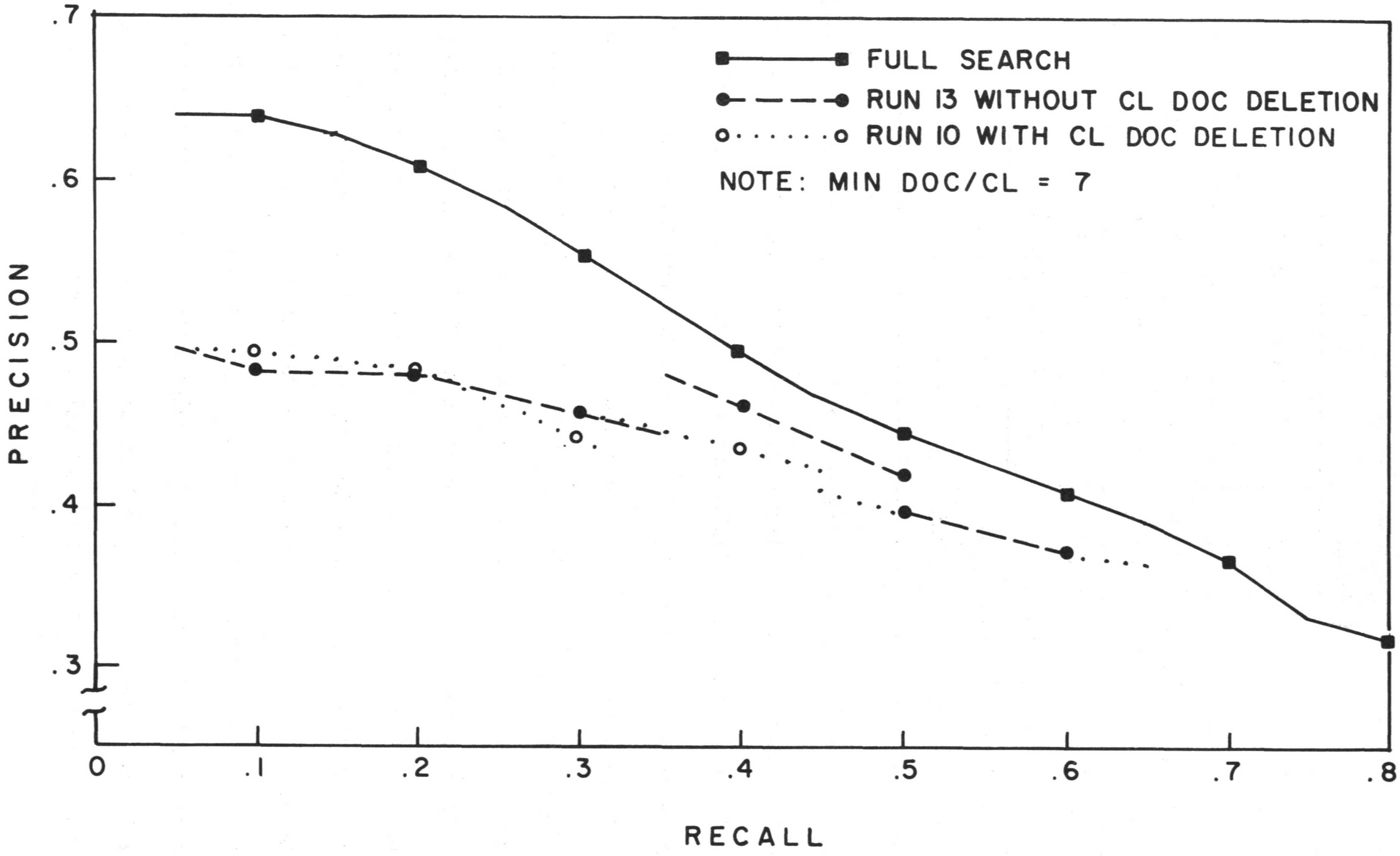
Study 2 The Effect of Deleting Clustered Documents from Subsequent Density Tests

Fig. 3



Study 2 The Effect of Deleting Clustered Documents from Subsequent Density Tests

Fig. 4



Study 2 The Effect of Deleting Clustered Documents from Subsequent Density Tests

Fig. 5

Run No.	Recall Ceiling			User Scanning %			Average docs/clus
	1	2	3	1	2	3	
7	.2007	.3725	.5129	8.2%	16.5%	24.1%	6.1
*8	.2097	.3817	.4305	5.9%	11.8%	17.5%	3.9
1	.3117	.4656	.6434	17.5%	33.4%	48.5%	12.4
*9	.3079	.4546	.5920	11.6%	23.0%	33.4%	8.2
10	.3163	.4546	.6358	16.7%	32.1%	46.8%	12.5
*13	.3565	.5033	.5825	12.5%	24.7%	36.3%	10.5

Note that for one and two clusters in all pairs, the recall ceilings are approximately equal, yet the user percentages are substantially reduced when all eighty-two documents are considered. In the third cluster the recall ceiling is reduced as well but the user percentages are down proportionately more. (A decrease of .0824 in recall ceiling corresponding to a decrease of 6.6% in user percentage; a decrease of .0514 in recall ceiling corresponding to a decrease of 15.1% in user percentage; a decrease of .0533 in recall ceiling corresponding to a decrease of 10.5% in user percentage).

Based on the above table, the user interested in a low user percentage would want the entire collection to be used in all density tests. This is to be expected since consideration of the entire collection in all density tests increases the number of clusters, decreases the documents per cluster, and therefore decreases user percentage since fewer documents are listed in each retrieved cluster. Unfortunately examinination of machine percentage versus recall ceiling does not establish such a well-defined pattern.

4. Determination of the optimum runs for the different user specifications yield the following: User 1 may obtain a recall of .3565 with the machine having to scan 25.9% and the user searching 12.5% of the collection. This result is slightly better than the one user 1 obtained in the previous study; in that study is obtained a recall ceiling of .3725 with the machine scanning 33.6% and the user scanning 16.5%. Users 2 and 3 obtain almost identical results in both studies. Unfortunately, no strong conclusions may be drawn from these parameters.

5. As in the previous study, the runs with the largest number of clusters yielded the optimum precision recall plots. These runs also had the largest cluster generation times. (Run eight with twenty-eight clusters required 14 minutes, 56 seconds whereas run thirteen required 8 minutes, 23 seconds and had seven clusters.) The following study attempts to keep the high precision-recall plots of the runs having many clusters, and simultaneously attempts to decrease the computer time required for generation.

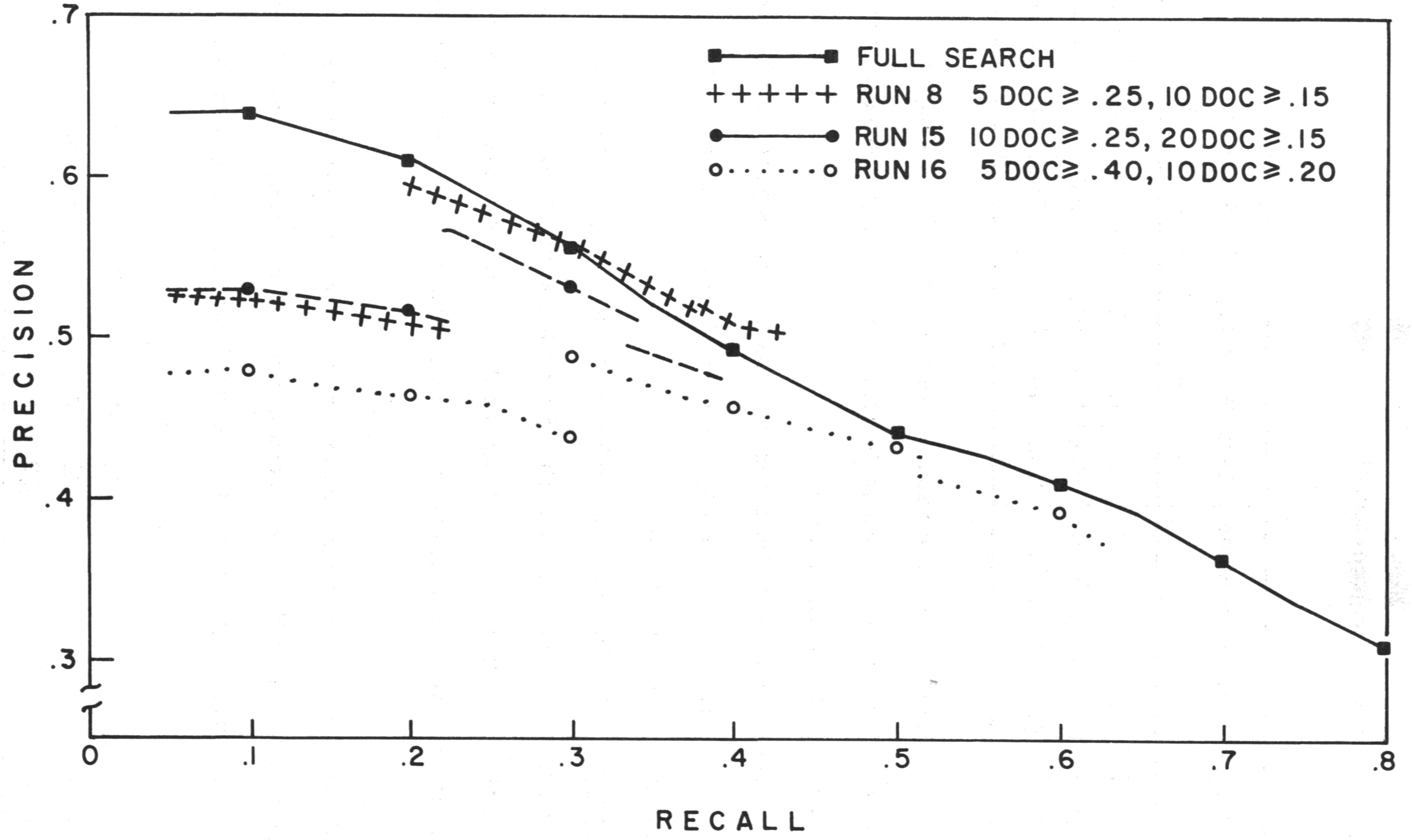
Study 3: To determine the effects of varying the density test parameters. From studies one and two it is concluded that the best precision-recall plots are obtained for those runs in which there are a large number of clusters and few documents per cluster. Unfortunately these runs are also characterized by large cluster generation times. Run eight of study two is especially important in that its precision-recall plot surpasses that of a full search. In an effort to keep the excellent plot and simultaneously reduce the cluster generation time, the density test parameters for run eight are made more rigid. It is hoped that the number of documents per cluster will remain low (consequently maintaining the high precision recall) and that simultaneously the number of clusters will decrease and therefore the time required for cluster generation will be reduced. Apparently these two conditions are contradictory. Note, however, there is a large amount of overlap (twenty-eight duplications) in run eight. If this overlap can be reduced, then both conditions may be met simultaneously. The following data are useful:

Run No.	Density Test		No. of Clusters	No. of Duplications	Docs/clus	Time
8	5 ≧ .25,	10 ≧ .15	28	28	3.9	14.56
15	10 ≧ .25,	20 ≧ .15	19	18	5.3	15.17
16	5 ≧ .40,	10 ≧ .20	8	4	10.7	11.38

Conclusions:

1. Run fifteen increased rather than decreased the cluster generation time. Evidently, the time involved in imposing a more rigid density test is greater than the time saved by generating fewer clusters. Run sixteen does manage to reduce the generation time but its density test parameters are too rigid as only eight clusters are formed. A compromise between the two runs is suggested but time did not permit other runs to be made.

2. The precision-recall plot of run fifteen (Fig. 6) is as good as that of run eight using one cluster. When two and three clusters are used, run eight is decidedly better. However, the plot of run fifteen is a good plot and is better than that obtained in any of fourteen other runs, excluding part of run three. The plot of run sixteen is clearly inferior to the other two.



Study 3 The Effect of Density Test Parameters on Recall and Precision

Fig. 6

3. The data below are useful.

Run	Machine Percentage			User Percentage			Recall Ceiling		
No.	1	2	3	1	2	3	1	2	3
8	40.0	46.0	51.7	5.9	11.8	17.5	.2097	.3817	.4305
15	31.2	38.9	47.0	8.1	15.8	23.8	.2237	.3384	.3833
16	25.7	42.0	54.3	17.2	32.3	44.6	.2941	.5194	.6223

Run sixteen clearly reduces the machine percentage. This is expected because of the few clusters present in this run. User percentage increases with increases in the cluster size.

4. Very definite effects can be obtained by varying density parameters. It is felt that this area is one of the most promising and considerable study should be devoted to it.

Study 4: To determine the effects of deleting low weighted concepts from the centroid vector. It was thought that low weighted concepts detract from the true picture of the centroid vector. Moreover by deleting such concepts it might be possible to save considerable computer time and/or storage.

First a word of explanation on how a centroid vector is generated. The weights of concepts in constituent documents are summed over all documents with the resulting vector of sums equal to the centroid vector. For example:

Document 1	(0	12	0	24	0	0)
Document 2	(36	0	48	0	0	48)
Document 3	(12	12	0	12	12	12)
Sum is Centroid Vector	(48	24	48	36	12	60)

The weighting scheme is arbitrary. In the runs of this report, it was performed as follows: If a concept appeared once in a document, it was assigned a weight of 12, twice a weight of 24, etc.

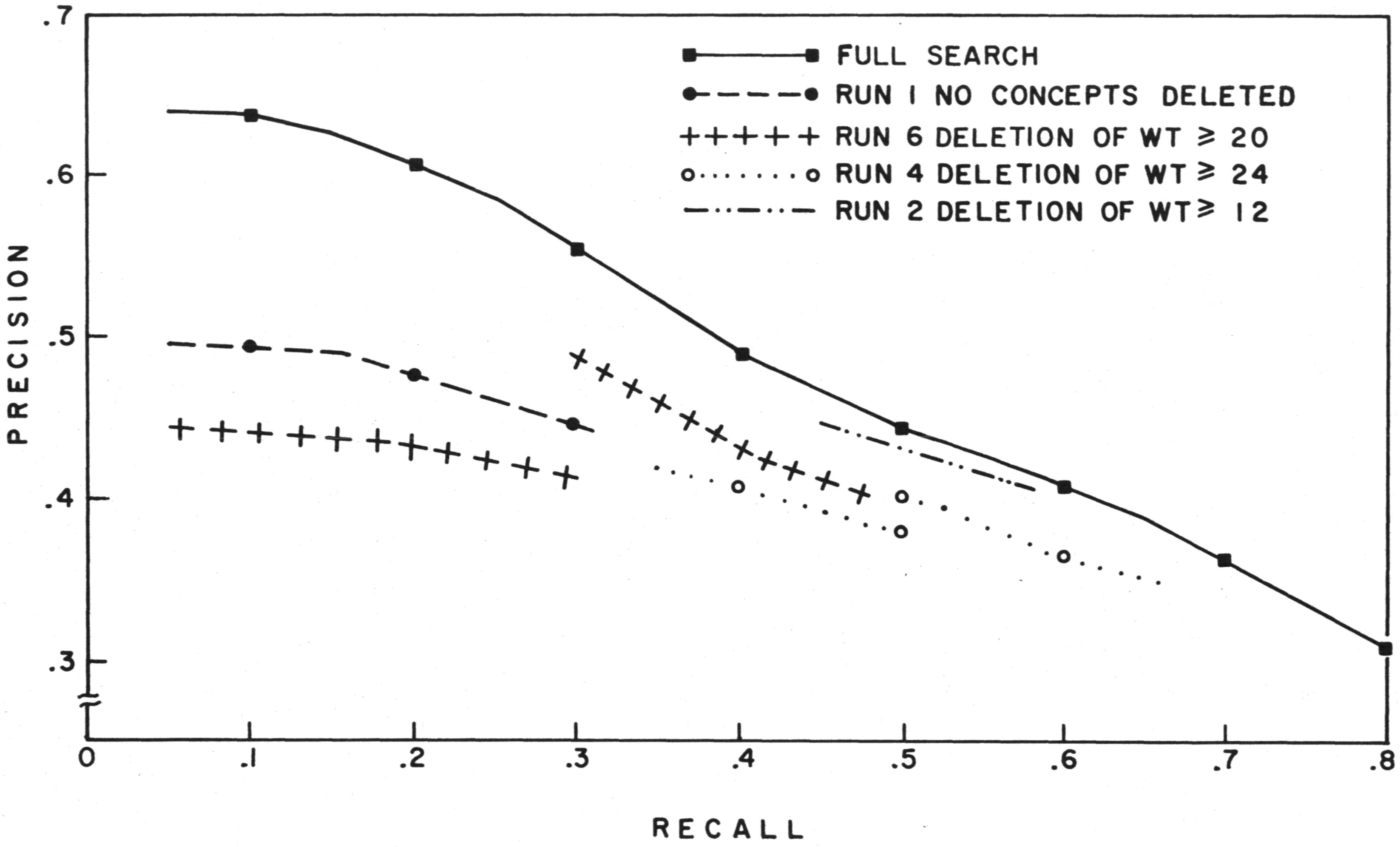
Conclusions:

1. By studying the precision recall plot of Fig. 7, one sees that there is no single run which is consistently best. Thus, the deletion of low weighted concepts appears to have little or no effect on a precision recall plot.

2. Based on the figures below one finds that it is possible to delete low weighted concepts with apparently no effect on recall ceiling, user percentage, or machine percentage. (Note that in the runs where the recall ceiling is low as in run two, the scanning percentages are proportionately lower as well.)

Run No.	Delete Wts.	Recall Ceiling		
		1	2	3
1	0	.3117	.4656	.6434
2	≦ 12	.2991	.3955	.5783
6	≦ 1% ≅ 21	.2990	.4727	.6259
4	≦ 24	.3431	.5035	.6692
11	≦ 36	.3085	.4912	.6317

Run No.	Machine Scanning %			User Scanning %		
	1	2	3	1	2	3
1	26.1	42.0	57.1	17.5	33.4	48.5
2	24.4	38.2	51.8	14.4	28.4	42.1
6	25.8	41.4	56.2	16.2	31.8	46.3
4	26.5	43.2	60.5	17.9	34.7	52.1
11	26.3	42.7	59.6	17.8	34.3	51.0



Study 4 The Effect of Deleting Low Weighted Concepts from Centroid Vector

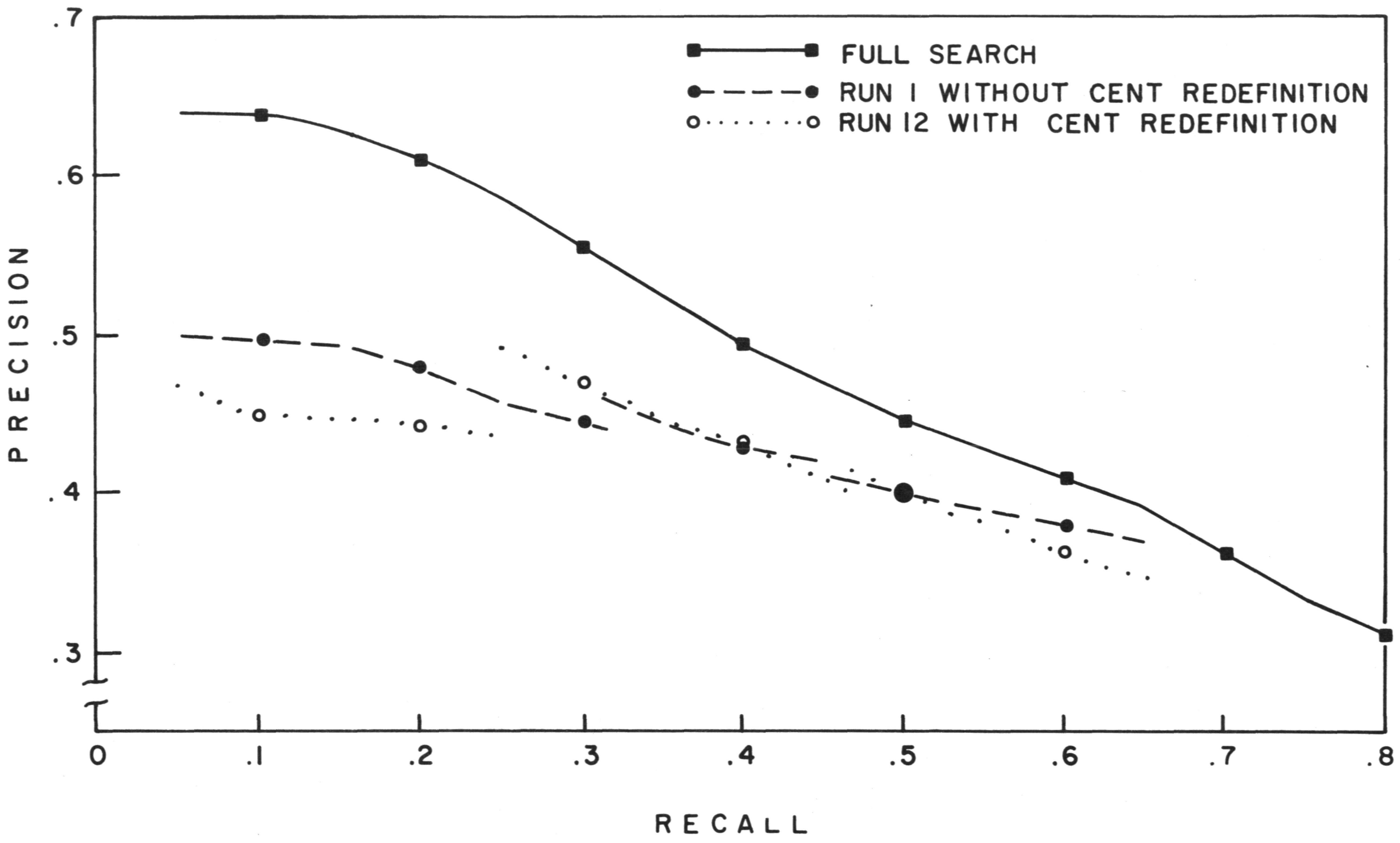
Fig. 7

3. Examining the times required for cluster generation, one fails to discover a definite relationship between deletion of weights and times required for generation. In conclusion, the authors feel that it is possible to delete low weighted concepts without either a beneficial or detrimental effect in recall ceiling, machine percentage, user percentage, the precision recall plot, or cluster generation times. The question of whether core storage is actually saved by the deletion remains open. Excluding the latter possibility, it appears that nothing is to be gained or lost by deletion of low weighted concepts from the centroid vector.

Study 5: To determine the effects of redefining the centroid vector. Remember that Rocchio's algorithm specifies that after the centroid vector has been generated it is matched against all documents in the collection, and that a set of documents is chosen to go into a cluster. It is possible that the set of documents chosen to go into the cluster differs slightly from the set originally used to constitute the centroid. This study investigates the effects of redefining the centroid vector to consist only of the set of documents actually contained in the cluster. This redefinition has a great deal of intuitive appeal. In the subsequent discussion run twelve is the run in which the centroid vector has been redefined. Run one has identical input parameters but its centroid vector remains constant. The output is shown in Fig. 8.

Conclusions:

1. Run one (the run in which the centroid vector is not redefined) gives a better precision recall plot a greater percentage of the time. The writers are unable to explain this result.



Study 5 The Effect of Centroid Vector Redefinition on Recall and Precision

Fig. 8

2. Both the machine and user percentages are extremely close using one, two, and three clusters in both runs. The average recall ceiling is decidedly better using one cluster in run one and very slightly better for two and three clusters in run twelve.

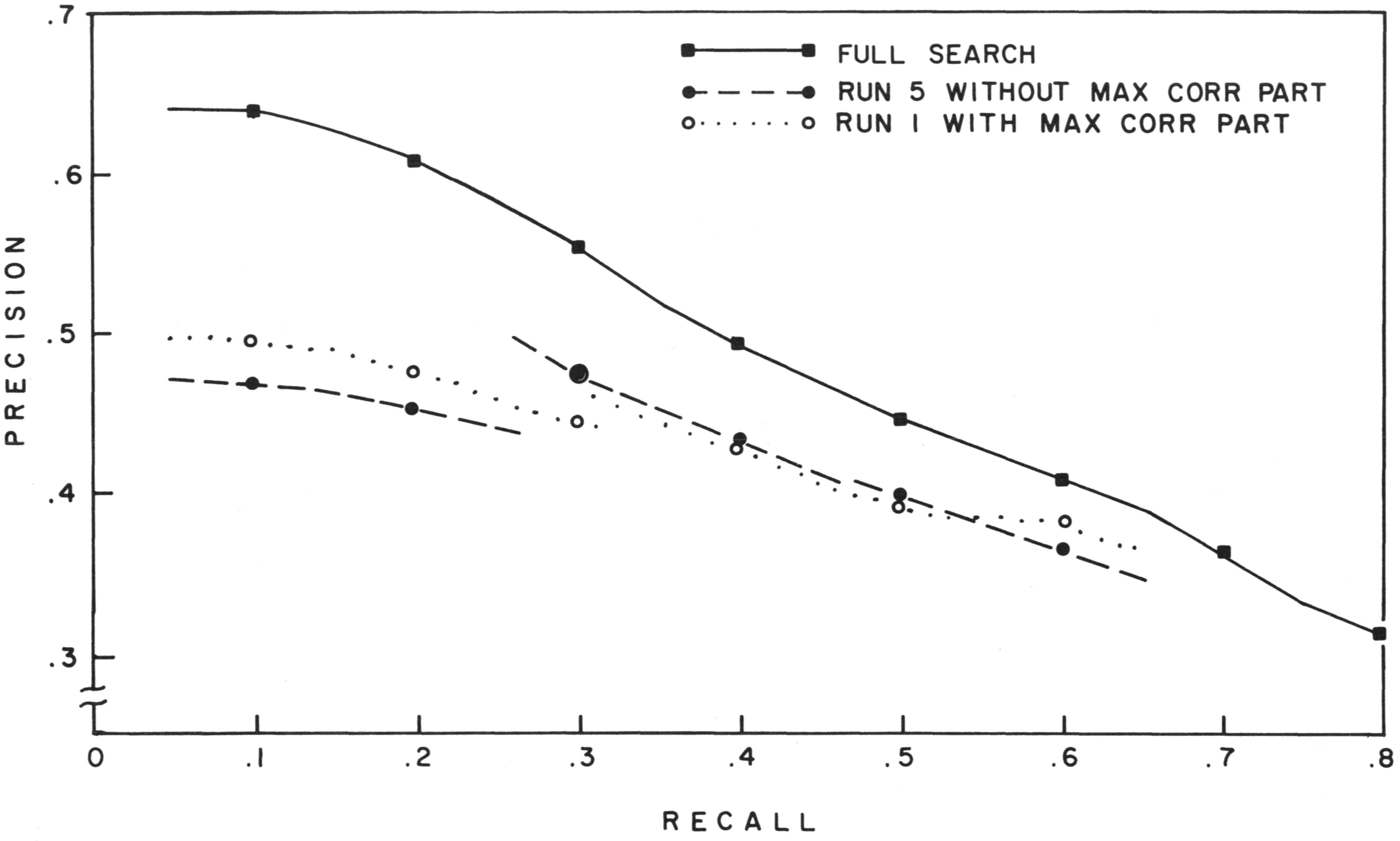
3. Obviously when the centroid vector is redefined one expects the cluster generation time to increase. This expectation is realized experimentally. Run one requires six minutes, fifty-two seconds. Run twelve requires seven minutes, fifty-five seconds.

4. Based on these observations, it would appear that the benefits to be derived from redefinition are not sufficient to merit the increase in computer time required for cluster generation. However, redefinition has such a strong intuitive appeal that the writers feel this study should be pursued further. A possible investigation would be a redefinition of the centroid vector immediately after the blended documents have been put into appropriate clusters.

Study 6: To determine the effects of eliminating the maximum correlation partition routine. (The function of this routine is described earlier). Run five is identical to run one except that the former omits the partition routine. Results are shown in Fig. 9.

Conclusions:

1. The precision recall plot appears to be better for run one a greater percentage of the time. One may guess this intuitively. The maximum correlation partition routine puts a document into the centroid with which it correlates highest. Hence, retrieved clusters which were generated under this routine are apt to have documents with higher correlations than those clusters generated without the routine. Thus, one expects improved precision.



Study 6 The Effect of the Maximum Correlation Partition on Recall and Precision

Fig. 9

2. The user and machine percentages are very close using one, two, and three clusters for both runs. The recall ceiling is decidedly better using one cluster for run one (as expected from the precision recall plot); little difference is found for three clusters.

3. As expected, run one has a larger cluster generation time (six minutes, fifty-two seconds) than does run five (six minutes, twenty-eight seconds).

4. The writers are unable to decide at this time whether the maximum correlation partition routine is worth the extra one-half minute of computer time. Further investigation of this topic is suggested.

C) Conclusions and Remaining Questions

It is felt that studies one, two, and four of this report are complete and resolve the questions:

1) Many clusters with few documents versus few clusters with many documents.

2) Consideration of the entire collection in all density tests.

3) Deletion of low-weighted concepts from the centroid vector.

The investigations of studies three, five, and six are only partially complete and should be pursued further. Among those problems untouched by this report, the following three appear to be most interesting.

1) Determination of search parameters; i.e. how many clusters are to be searched and how many documents in each cluster?

2) Effects of not blending all unclustered documents, or of blending only a percentage of unclustered documents.

3) Establishment of a cut-off criterion; i.e. should cut-off occur after the greatest difference in correlation coefficients, after the first difference which exceeds p, or after the first n documents?

References

[1] J. J. Rocchio, "Document Retrieval Systems — Optimization and Evaluation", National Science Foundation, Harvard Computation Laboratory, Cambridge, Mass., Report ISR-10 (doctoral dissertation), March 1966