X. Evaluation of Feedback Retrieval using Modified Freezing, Residual Collection and Test and Control Groups

C. Cirillo, Y. K. Chang, and J. Razon

Abstract

Three methods of feedback evaluation are experimented; Modified Freezing, Residual Collection and Test and Control Groups. Feedback runs are performed using each method on the SMART system, and results are compared against the previously used evaluation methods of total performance and full freezing, with respect to improvements of "feedback effect" evaluation.

1. Introduction

The two most common methods of evaluation of feedback retrieval systems both have weaknesses. Both total performance and feedback effect (full freezing) evaluation limit the attainable performance in later iterations.

In total performance the evaluation after feedback is very biased, since relevant documents already seen by the user are moved to the top of the ranking, thereby distorting the feedback evaluation, making it seem really good, while most of the improvement is gained simply by a reranking of documents already seen. This is known as "ranking effect", and the goal in feedback evaluation is to eliminate this ranking effect and develop a method which measures accurately only the "feedback effect", i.e. how much the new query is improved over the old query as far as the number and rank of <u>new</u> relevant documents retrieved.

Three methods of feedback evaluation, suggested by Ide in report ISR-15, are evaluated in this study. Part A evaluates the modified freezing technique. This is a method similar to the full freezing used by SMART except that certain nonrelevant documents are not frozen, in an attempt to give a more accurate picture of the "feedback effect". In Part B the method of residual collection evaluation is being used. Here both the ith and (i+1)st iteration queries are used to search the (i+1)st iteration residual collection in an attempt to isolate the "feedback effect" on the residual collection, and hence to measure it precisely. Part C considers the test and control groups method. A document collection is split into two halves — feedback runs are done on the test group, and the resulting modified queries are run on the control group, thus eliminating the "ranking effects" on the control group, resulting in an accurate evaluation of only "feedback effect".

Each method tries to isolate the "feedback effect", and an attempt is made to evaluate how accurately each method actually does measure the "feedback effect".

Part A

Evaluation of Feedback Retrieval Using Modified Freezing

1. Introduction

The method of full freezing is used in the SMART system, in an effort to eliminate the "ranking effect" and evaluate only the "feedback effect" in feedback evaluation. This method freezes the ranks of all documents presented to the user on earlier feedback iterations, and assigns the first document retrieved on the ith iteration a rank $iN+1$, where N documents are presented to the user (used for feedback) on each iteration. This measure of "feedback effect" is fairly accurate up to the first iteration. However, after that, any documents retrieved cannot be ranked higher than $2N+1$, and hence will have very little effect on the precision-recall curve.

A method suggested by Ide in report ISR-15 to evaluate these later iteration feedback improvements somewhat more effectively is the use of a modified freezing technique. She hints that evaluation by modified freezing might show later feedback iterations to be nearly as valuable as the first in moving the modified query towards the optimum query.

2. Modified Freezing

Modified freezing differs from full freezing as follows: in modified freezing all relevant documents retrieved on the ith iteration and used for feedback on the $(i+1)$st iteration have their ranks frozen, and all nonrelevant documents ranked above the last ranked relevant document used for feedback are also frozen. Nonrelevant documents ranked below the last relevant are not frozen. Hence the number of documents frozen on each iteration may vary, while in full freezing a specified number, N, are frozen on each iteration. In both methods, however, N new documents are retrieved (used for feedback) on each iteration.

The modified freezing algorithm is not presently implemented in the SMART system. In order to evaluate feedback by using the modified freezing method, one must simulate the method by re-ranking, by hand, the output from a previous search which uses the full freezing evaluation technique. A specific example should indicate exactly how this re-ranking is done. Consider query 25 of the ADIABTH collection, with 3 relevant documents. The initial query ranks the documents as follows:

Rank	1	2	3	4	5	...	15
Document	13R	53R	60	37	40	...	24R

Using feedback and full freezing, the results on the first iteration are:

Rank	Document	Correlation
1	13R	.8772
2	53R	.8103
3	60	.2902
4	37	.2770
5	40	.2834
6	24R	.5092
7	26	.3707
8	56	.3601
9	74	.3156
10	5	.2989

The first 5 (N) documents are frozen in this case, regardless of whether they are relevant or not. Document 24 moves up from rank 15 to rank 6 (1N+1). This is the best possible improvement and should be reflected in a sizeable increase in the precision-recall curve for the first iteration over that of the initial query. Now reranking to simulate modified freezing, using the correlations listed and freezing only up to rank 2 (the

last ranked relevant document retrieved in this case), the rankings are as follows:

Rank	Document	Correlation
1	13R	.8772
2	53R	.8103
3	24R	.5092
4	26	.3707
5	56	.3601
6	74	.3156
7	5	.2989
8	60	.2902
9	40	.2834
10	52	.2829

Ranks 3-5 are not frozen since they were nonrelevant documents ranked below the last retrieved relevant document. Document 24 moves up from rank 15 to rank 3. This once again is the best possible improvement in the feedback iteration. This example indicates one way in which modified freezing might be superior to full freezing as a method of evaluating "feedback effect". However, consider the case where a relevant document is ranked fifth. Then, if by feedback a relevant document is moved up from rank 15 to rank 6, it will still be ranked 6 by the modified freezing technique, and the precision-recall curves will be identical. However, it will seem as if the previous case had better feedback than the latter in the modified freezing evaluation, but in reality the feedback improvement in both cases is identical, each one giving the maximum improvement. This is a minor drawback to the modified freezing method.

If no relevant documents are retrieved on the initial query, on the first iteration one looks at the first 10 documents (the first 5 will be

identical to those retrieved on the first iteration, providing positive feedback is used), following the identical rules spelled out earlier.

3. Evaluation Results

In this section, copies of the output for searches of two document collections using feedback and full freezing are examined (the ADIABTH and the CRN 200 collections also used in Part B). Simulation of the modified freezing technique was accomplished by reranking the documents by hand, using the correlations listed in the output, as explained earlier.

For the ADIABTH collection, all 35 queries are used, and full freezing is compared to modified freezing on the first and second iterations. The resulting precision-recall curves turn out to be almost identical, with the modified freezing curves slightly higher than the full freezing curves. This is expected, since the relevant documents can only be ranked higher using modified freezing, not lower. Two reasons can be offered to explain why the average curves are so close together. First, the feedback in the ADIABTH collection is not as good as that in the CRN 200. Second, and more important, is the fact that all queries are used. In about half of the queries, the statistics using modified freezing and those using full freezing are identical. Three reasons may be given:

a) the feedback result is not good enough to enable relevant documents to have higher correlations than the unfrozen, previously retrieved nonrelevant documents;

b) all relevant documents are retrieved by the initial query;

c) no relevant documents are retrieved by the initial query.

There is also the case where a relevant document is the last retrieved docu-

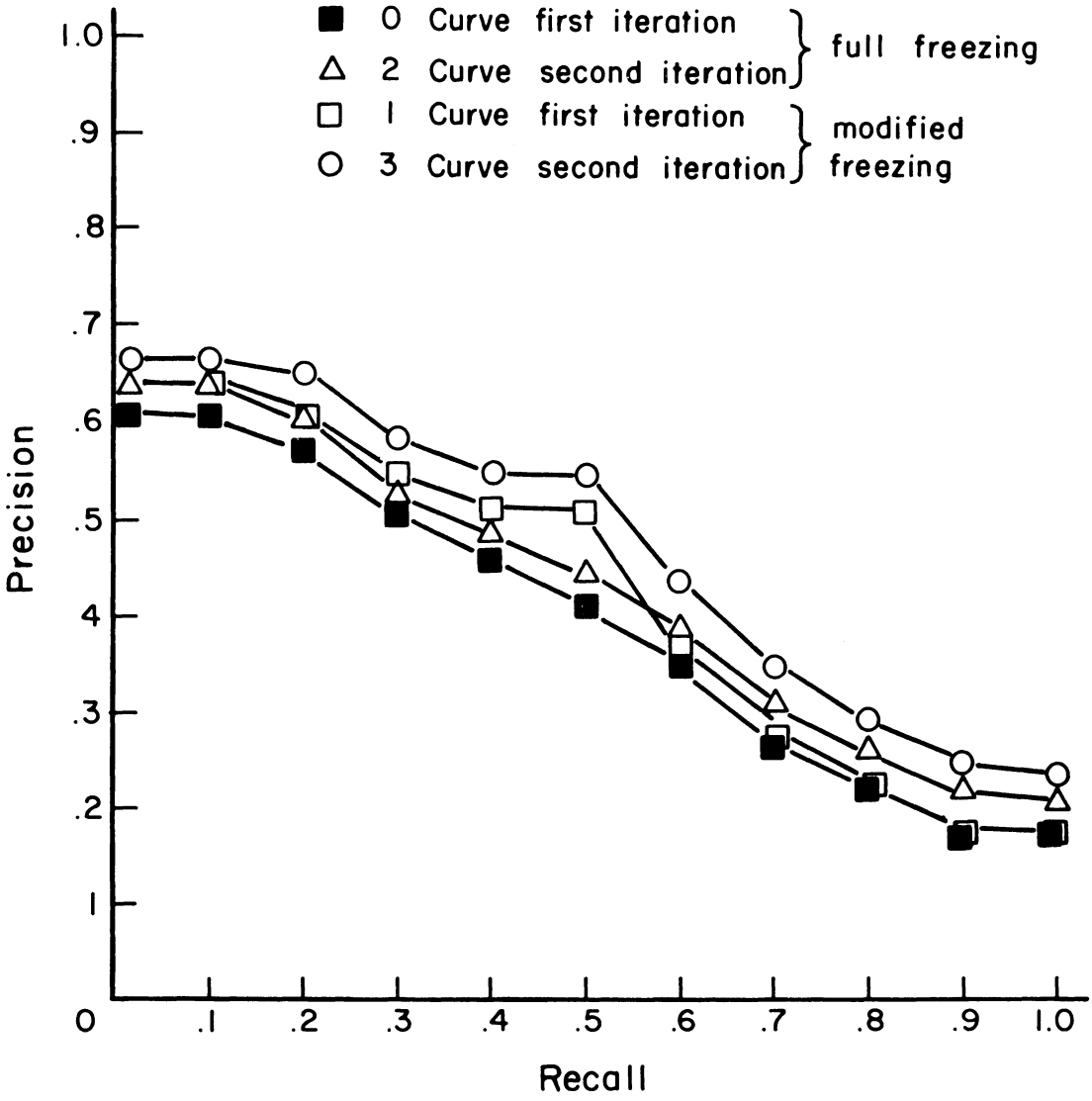
ment, as explained earlier. This is not to say that modified freezing is not inherently more useful than full freezing. In most of these cases one does not expect a high precision-recall curve, so that the resulting lower curve will correlate correctly with the higher curves gotten when the feedback is better.

For the CRN 200 collection, only the queries with different statistics for the modified and full freezing rankings are considered. This is done in order to isolate the advantages of modified freezing over full freezing. Only 24 out of the 42 queries yield different results for the two methods. However, this presents no problem, as explained above. The resulting precision-recall curve appears at the end of part A of this report. The key is as follows:

0 : 1st iteration full freezing P-R curve
1 : 1st iteration modified freezing P-R curve
2 : 2nd iteration full freezing P-R curve
3 : 2nd iteration modified freezing P-R curve.

4. Discussion

In the recall-precision graph of Fig. 1, the area between curves 0 and 2 is the feedback gain between the 1st and 2nd iterations, and the area between curves 1 and 3 is the same, the former using full freezing to evaluate it, the latter using modified freezing. The latter area is considerably greater than the former, and since both isolate the "feedback effect", the higher curves give a more reasonable picture of the improvement gained, as the curves are not damped by the freezing of so many nonrelevant documents. In fact, modified freezing evaluation seems superior to that of full freezing even on the 1st iteration. The modified freezing curves have a wider



Modified Freezing Evaluation

Fig. 1

range than the full freezing curves and hence can show more distinctly the difference between good and average feedback, due mostly to the decrease in the damping effect mentioned earlier. Hence the conclusion that modified freezing is superior to full freezing on a query-by-query basis.

This form of modified freezing can only be used with a positive feedback algorithm, since in negative feedback the nonrelevant documents are used to modify the query. If these are not frozen, they may be used again to modify the query, thus biasing the results. A small change in the algorithm can remedy this.

All in all, modified freezing does seem to be an improvement over full freezing as a method to evaluate the "feedback effect", especially on an individual query basis. However, the improvement tends to be swamped (as shown by the results on the ADI collection) by queries in which no difference appears between the two methods. It would seem worthwhile to include a modified freezing algorithm in the SMART system to be used as an option for individual query comparisons.

Part B

Evaluation of Feedback Retrieval Using Residual Collection Evaluation

1. Statement of the Problem

The measure of effectiveness of a relevance feedback system should be a measure of how many new relevant documents are retrieved as a result of feedback, as stated by Hall and Weiderman. In other words, the question should be "How close is the modified query to the optimum query for the documents not yet presented to the user?". Although this question is important for the evaluation of feedback strategies, neither the ordinary freezing method nor the modified freezing evaluation directly answers it. Therefore the residual collection evaluation method is used in an attempt to solve the problem.

The present problem and the method of solution are suggested by Ide in Section VII-B, report ISR-15.

2. Summary of Methods

Generally speaking, this method treats the remainder of the document collection, excluding those documents used for feedback, as a complete collection and the remainder of the relevant documents as a complete set of relevant documents, and then performs a total performance evaluation of the modified query in this new environment.

First one obtains the output of a search of a document collection using three iterations of full freezing, including the ranks of all the relevant documents for each query. To calculate the performance of the i^{th} iteration query in the $(i+1)^{st}$ iteration residual collection, all relevant documents not used for feedback retrieval on the $(i+1)^{st}$ iteration are to be reranked in the following way: the relevant documents in the i^{th} iteration are reranked by subtracting the number of documents used for feedback retrieval on the $(i+1)^{st}$ iteration from the original rank of these documents. If no rele-

vant documents remain the query is not used in the evaluation. Using these new ranks for the relevant documents, and the size of the $(i+1)^{st}$ iteration residual collection as the size of the document collection, the SMART routines RESCOL and AVERAG are called to calculate all measures and to plot recall-precision graphs.

Take as a specific example the evaluation of the second iteration query with respect to the third iteration residual collection in the ADIABT collection (82 documents, 35 quests):

1) Obtain a copy of ADIABT relevance feedback search output (5 new documents presented to the user and frozen on each iteration);

2) Since 5 documents are presented on each iteration, the size of the third iteration residual collection is 82 - 3x5 = 67 and all relevant documents with ranks larger than 15 as seen from the second iteration output are decreased by 15. For example, Q7 has originally 4 relevant documents: 7, 9, 19 and 40. On the second iteration, the output is as follows:

Rank	Doc	New Rank
1	19R	—
⋮	⋮	
13	40R	—
⋮	⋮	
15	69	—
16	7R	1 (=16-15)
17	9R	2 (=17-15)
⋮	⋮	

The number of relevant documents is 2, since there are only two relevant documents, 7 and 9, with ranks larger than 15. They are reranked to 16-15=1 and 17-15=2 respectively.

3) If no relevant document remains for a query, the query is dropped from the query sample. For instance, Q6 has only two relevant documents, 71 and 12, with ranks 3 and 11 respectively. Both ranks are less than 15, hence no relevant document remains in the third iteration residual collection and so the query Q6 is thrown away.

4) After reranking the original search output for each of the 35 queries, the RESCOL and AVERAG routines are called and thus the recall-precision curve, labelled 0, is obtained, This iteration is named RES23.

5) In order to compare the performances between the second and the third iteration queries, both with respect to the third iteration residual collection, those relevant documents in the residual collection of the third iteration are reranked similarly and curve 1 is obtained. This iteration is named RES33.

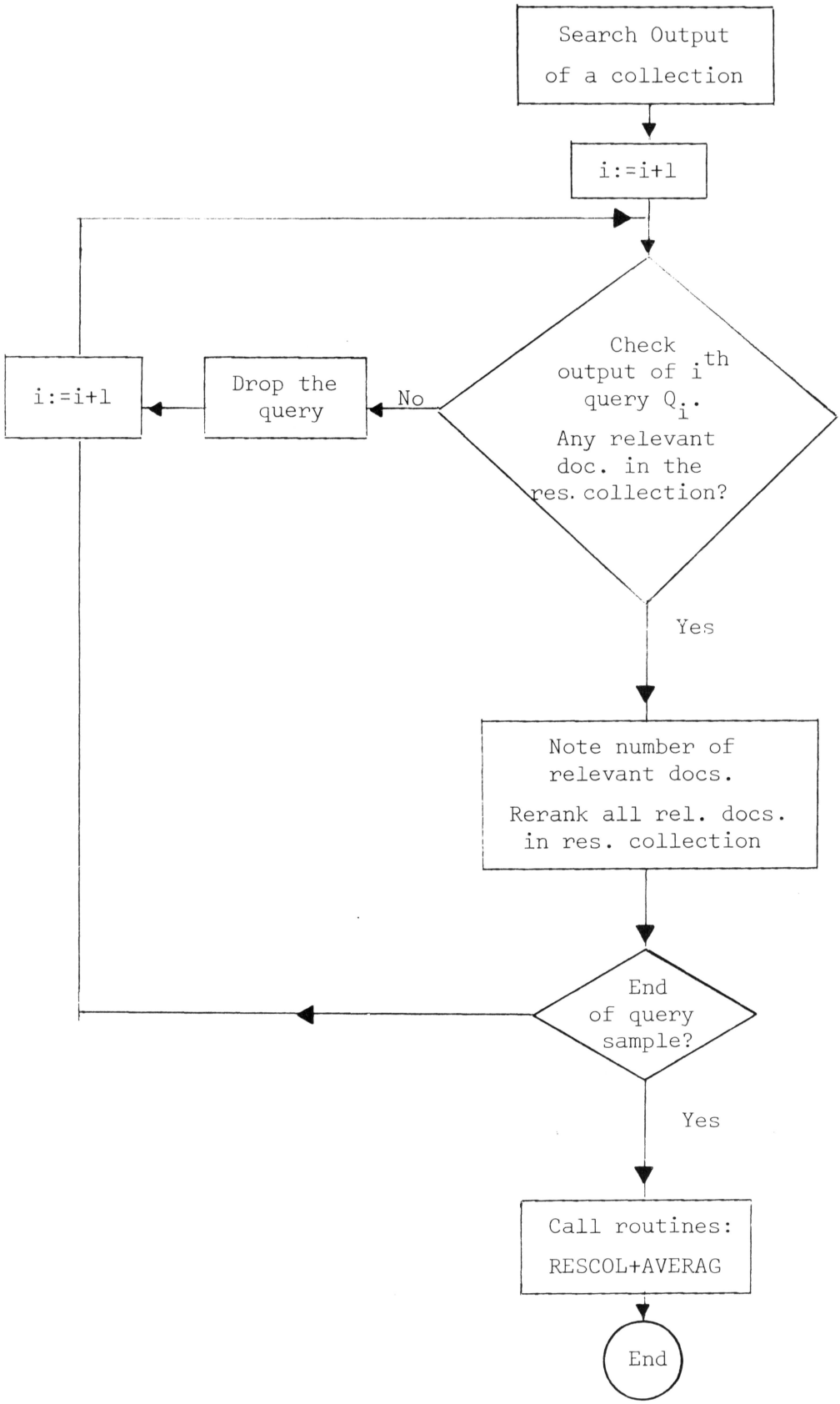
6) The original performance curves for the second and third iterations with respect to the 82 document collection are included in the same plot, labelled 2 and 3 respectively. These iterations are named RES-2 and RES-3.

3. Results and Conclusions

Three problems posed by Ide in ISR-15 are solved in the following way:

a) When all relevant documents are retrieved before all requested iterations are completed the query is dropped from the query sample.

b) Difficulties arise in averaging the performance of different queries because each query may have a different sized residual collection. In this project the number of documents used for feedback is the same for all queries on a given iteration. Therefore the size of the residual collection is fixed and no trouble arises. Otherwise recall



Simplified Residual Collection Evaluation System

Fig. 2

and precision could be averaged after a specific number of documents or after a certain percentage of the document collection had been retrieved.

c) A further difficulty may arise in comparing two methods of feedback which, for a given query, result in different generality numbers for the residual collections. As subsequent searches are made, the queries will be searching collections that include a different number of relevant items, and hence direct comparison (or averaging) of the results may not be valid.

d) The problem of reranking is handled in Section 2. Four computer runs have been performed; the output is given in Figs. 3-6.

Fig. 3: CRN2TH, 200 documents, 42 quests.
Initial and first iteration queries with respect to the first iteration residual collection.

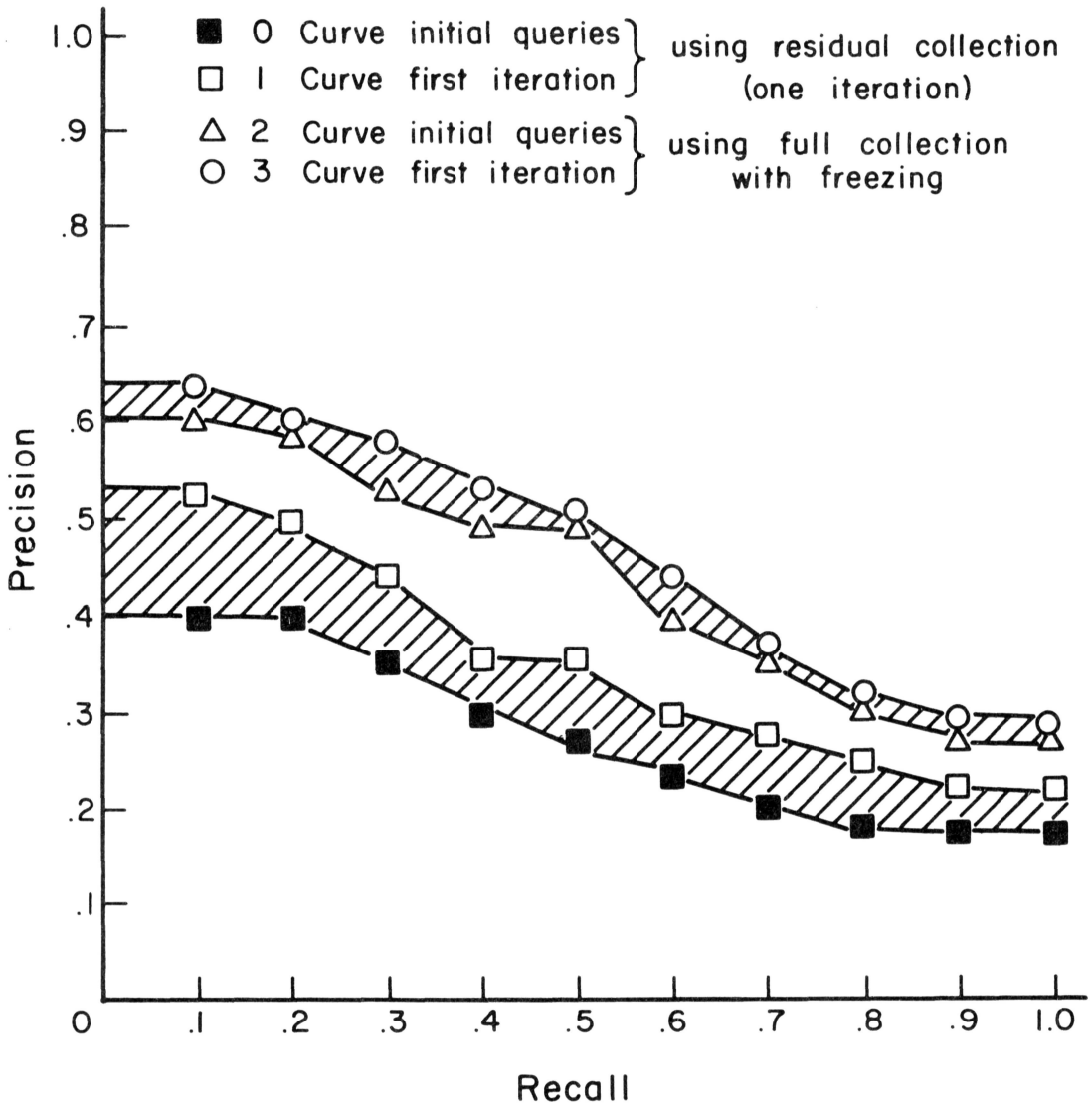
Fig. 4: CRN2TH, 200 documents, 42 quests.
Second and third iteration queries with respect to the third iteration residual collection.

Fig. 5: ADIABT, 82 documents, 35 quests.
Initial and first iteration queries with respect to the first iteration residual collection.

Fig. 6: ADIABT, 82 documents, 35 quests.
Second and third iteration queries with respect to the third iteration residual collection.

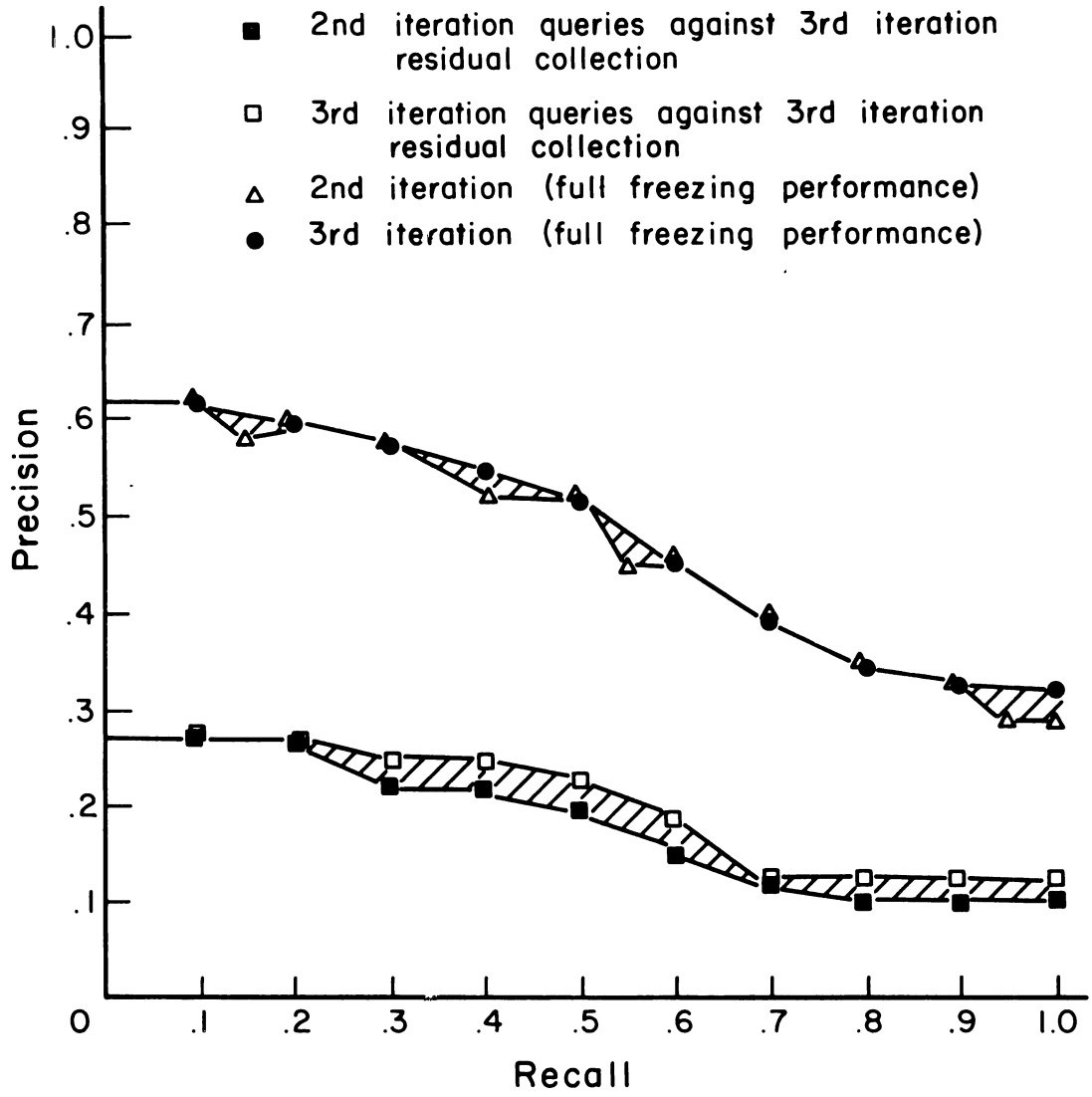
From the results shown, several conclusions can be drawn:

a) From the R-P curves for the intial and first iteration queries curve 1 is found to be quite a bit higher than curve 0 in both Fig. 3 and Fig. 5. This is as expected, because the modified query significantly improves the results in



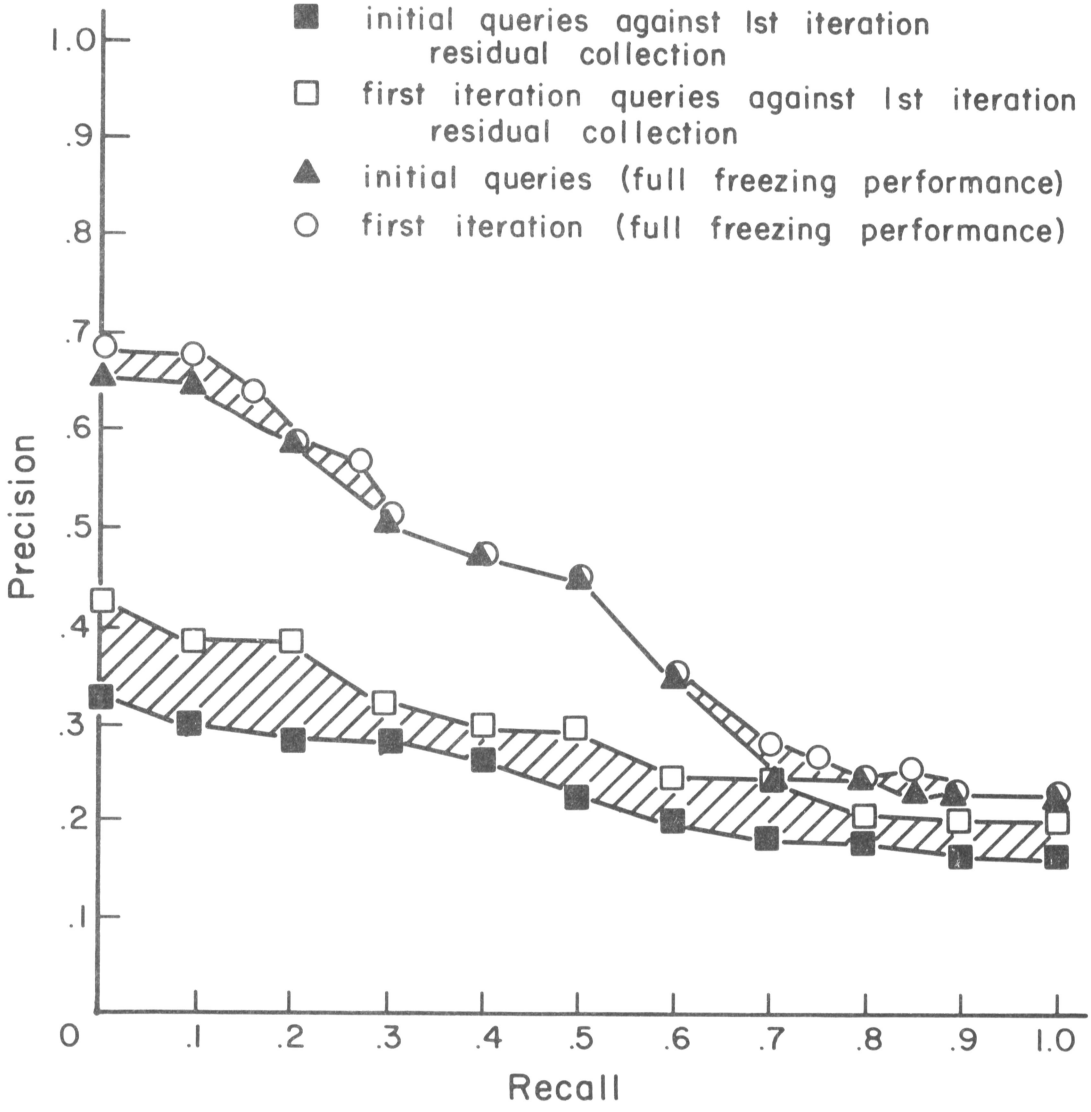
Cranfield Collection — Residual Collection Evaluation

Fig. 3



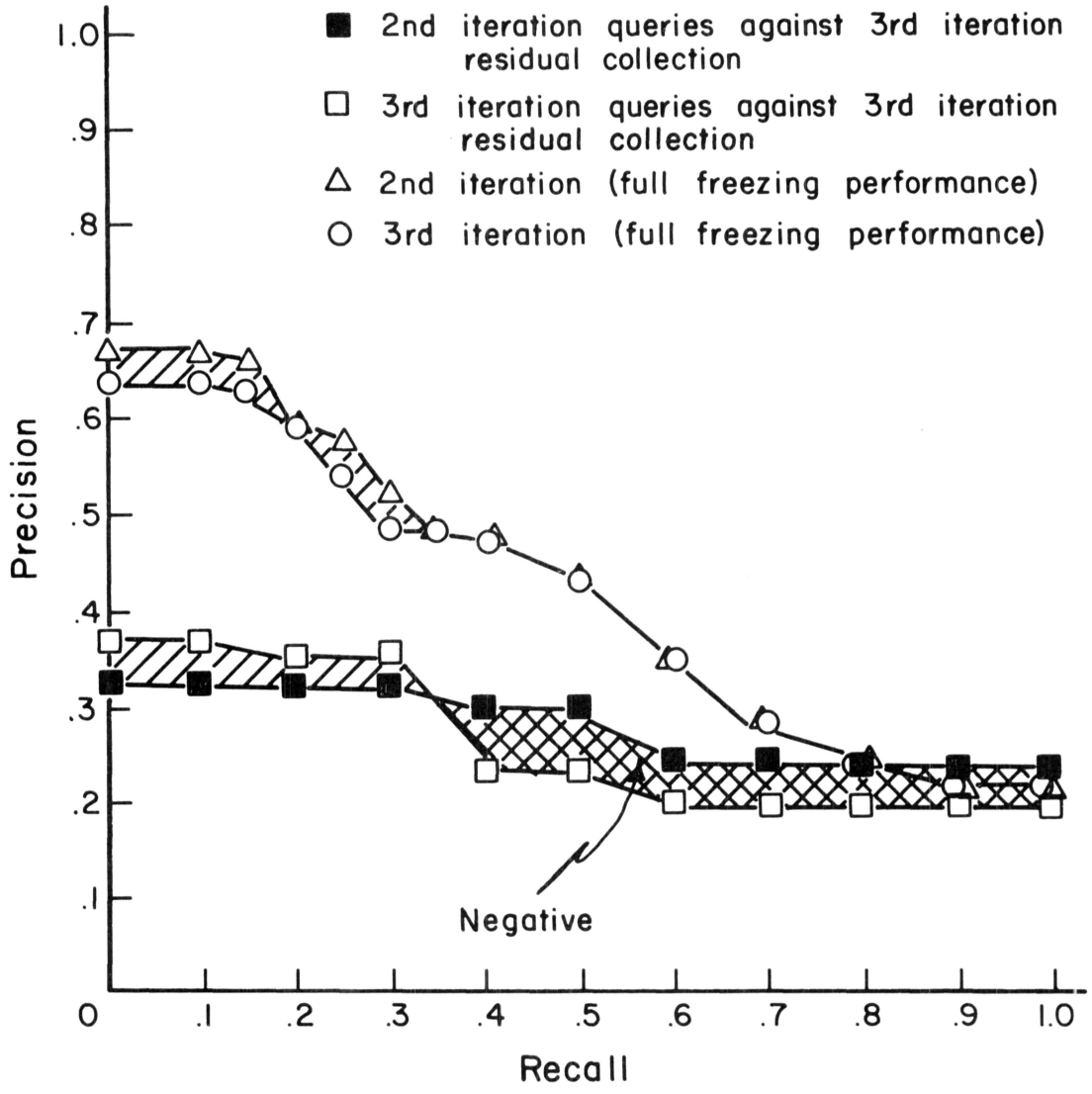
Cranfield Collection—Residual Collection Evaluation

Fig. 4



ADI Collection—Residual Collection Evaluation

Fig. 5



ADI Collection – Residual Collection Evaluation

Fig. 6

the first iteration. The residual collection evaluation method does show this marked difference in results, and this implies that a further iteration is worthwhile to get more new relevant documents.

b) By comparison of R-P curves for the second and third iteration queries (Fig. 4 and Fig. 6), one can find that in Fig. 4 curve 1 is almost the same as curve 0, and in Fig. 6 curve 1 is lower than curve 0. This can be explained by noting that after two iterations, the relevant documents are mostly already retrieved; the query could then be modified by weighting in the wrong direction, especially for the not-well-formed ADIABT collection. Thus for higher recall, curve 1 produces worse performance than curve 0, as shown in Fig. 6. These results imply that no further iteration is recommended, i.e., the user should look at more retrieved items on the second iteration, instead of performing a third iteration feedback.

c) Since the difference between the curve 1 and curve 0 is much larger than that between curve 3 and curve 2 (original freezing performance curves), and no ranking and freezing effects are involved in this evaluation method, it can be claimed that the residual collection evaluation method is better than either the freezing and modified freezing methods. However the reranking job must be done for each iteration, and the problems discussed in section 3 must be considered.

d) The relevance feedback searching algorithm appears to operate well since within two or three iterations, almost all relevant documents are normally retrieved.

e) Since the CRN2TH 200 documents and 42 quests are better formed and selected than the ADIABT collection, the performance curves are smoother than those of the ADIABT collection.

Part C

Evaluation of Feedback Retrieval using Test and Control Groups

1. Introduction

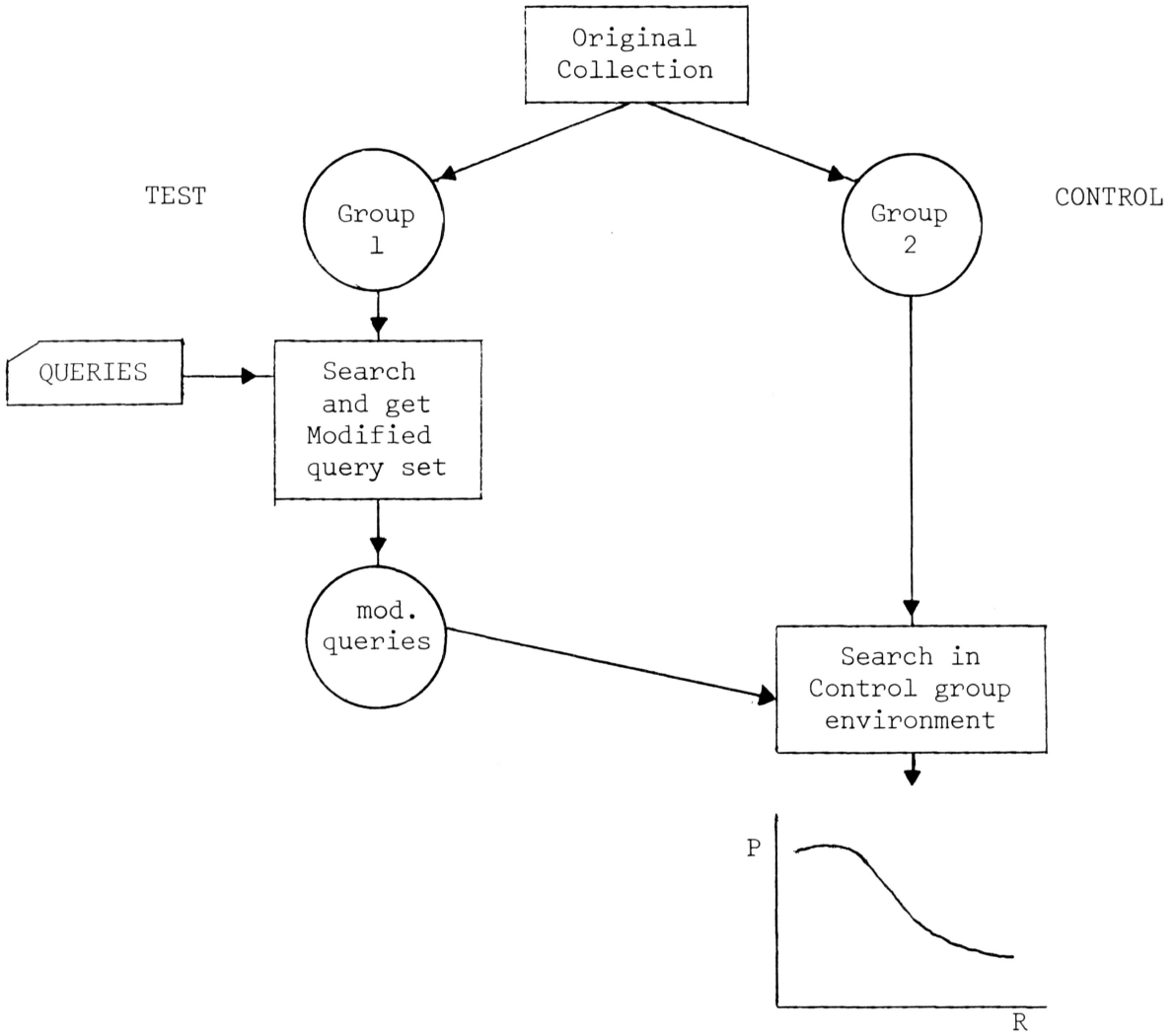
A third method of feedback retrieval evaluation which avoids the ranking effect problems is being used experimentally. The general scheme is as follows: A given collection is randomly split into two halves. One group is used to run an intial search and to modify the queries based on user relevance judgments, and the other group which has not been utilized to modify the queries is used to evaluate the performance of the feedback retrieval. Fig. 7 schematically represents this process.

2. Process Description

The collection CRN4S which includes 424 documents and 155 queries is being used. Two collections have been created based on odd and even document numbers. The Odd collection is used as the test group, while the Even collection is used as the control group.

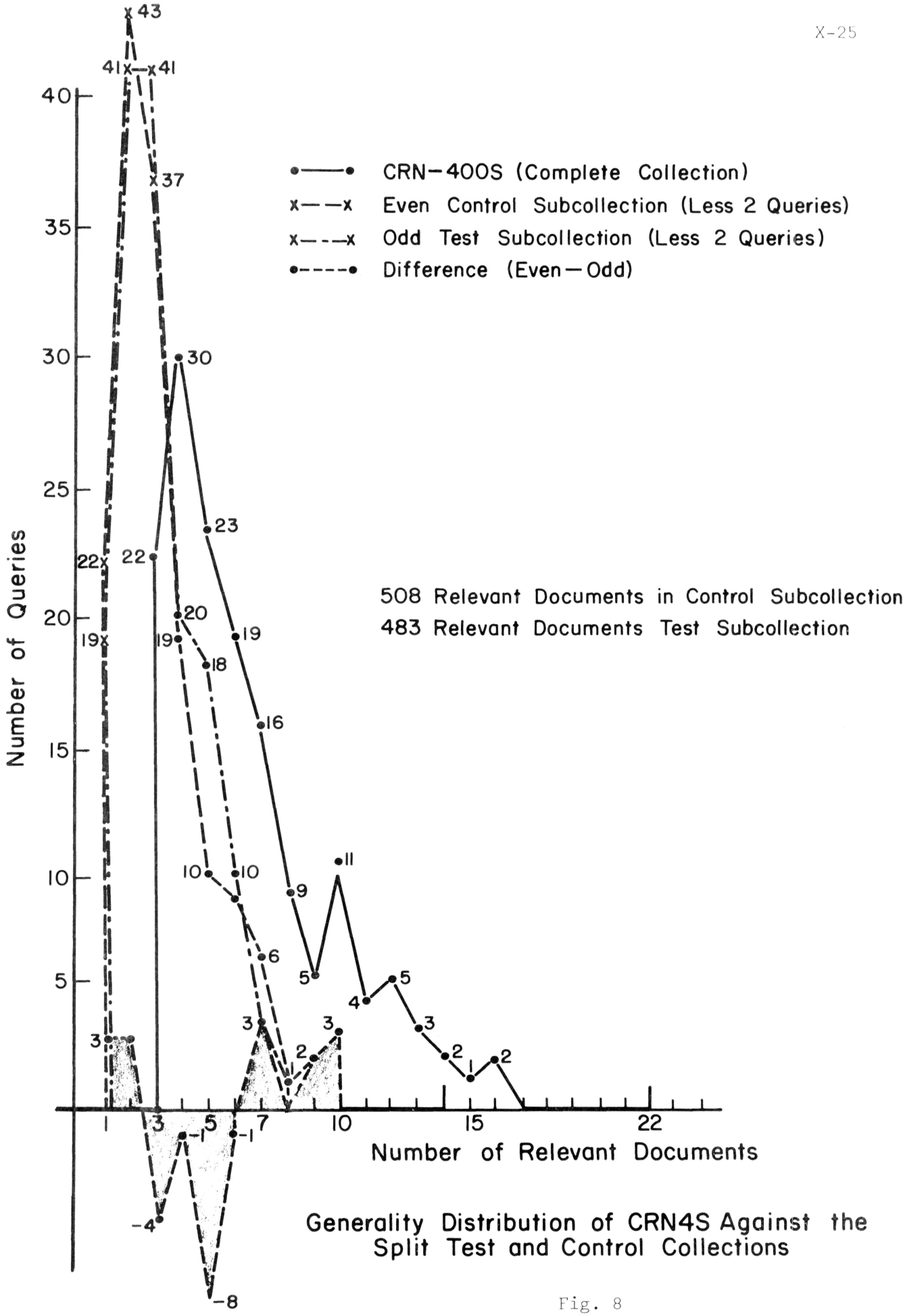
The reason for splitting the collection by using odd and even document numbers is simplicity. It is assumed that this process is sufficiently random to generate evenly distributed collections. From the original set of queries, two queries were deleted because they have no relevant documents in the test collection. Two query collections were then created, each one including the same number of queries (153) but with relevance decisions adjusted to interact with the Test and Control groups (see details in Section 3).

Fig. 8 shows the generality distribution of the original CRN-400 collection along with the two subcollections. The collections are quite balanced from the point of view of relevant documents (508 relevant documents in the Even collection and 483 in the Odd). The discrepancy between



General Flow Chart of the Process

Fig. 7



Generality Distribution of CRN4S Against the Split Test and Control Collections

Fig. 8

number of relevant documents per query in the Odd and Even collection is also small (see Fig. 9). Attention should be paid to the fact that the difference between the generality of the Even collection and the Odd collection, as represented by the dashed curve in Fig. 8, is due to 14 queries from the Odd collection centered at the 3-6 relevant document range against 15 queries from the Even collection spread over the rest of the whole range. This uneven distribution might cause discrepancies in the performance of the two collections.

The following steps are now carried out:

a) An initial full search (0 and 1 iteration) of the query sets against the test and control groups is performed, and averages are computed. The results will be preserved by SMART.

b) The results of the zero iteration between the two groups are compared and the similarity of the two subcollections is evaluated.

c) The relevance decisions of the queries which have been modified by the Odd-Test collection are changed by inserting the numbers of the relevant documents of the Even-Control collection as the relevant documents of the modified Odd query collection.

d) The feedback evaluation search is performed using the above query collection and the Control group

e) Results are evaluated.

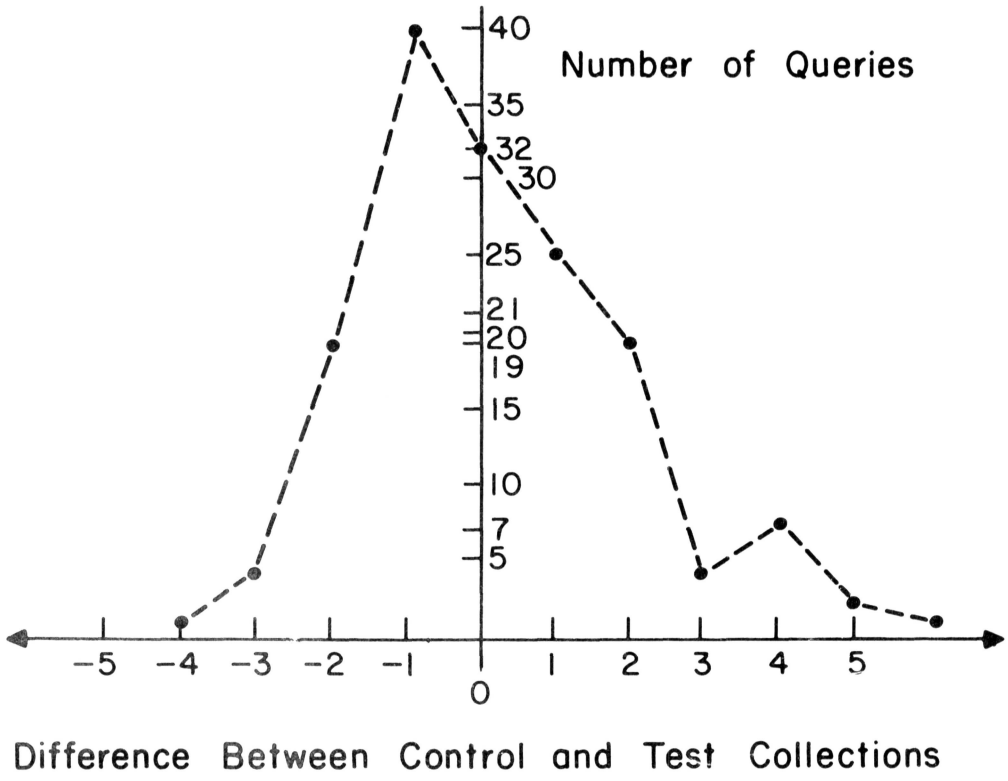
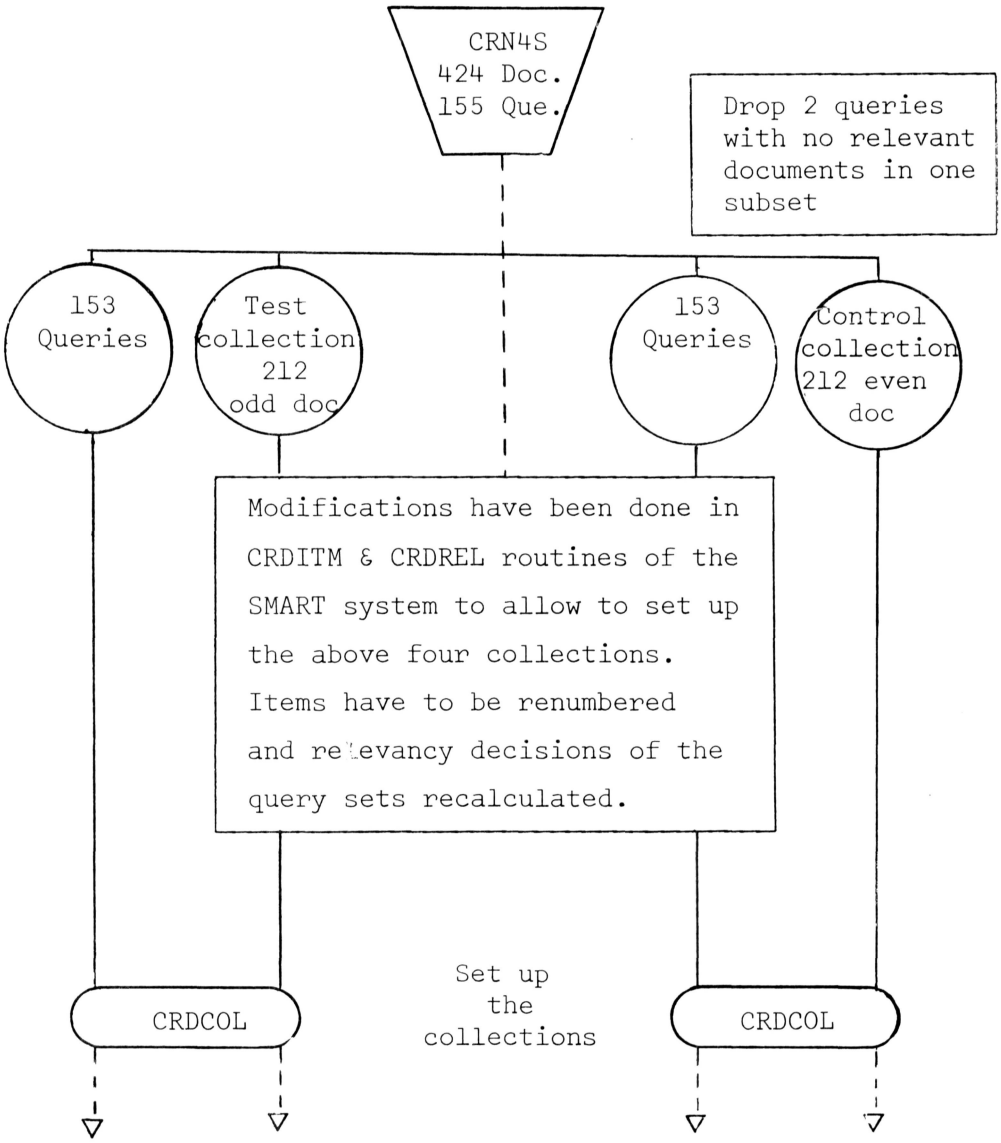


Fig. 9

3. Experimental Results and Evaluation

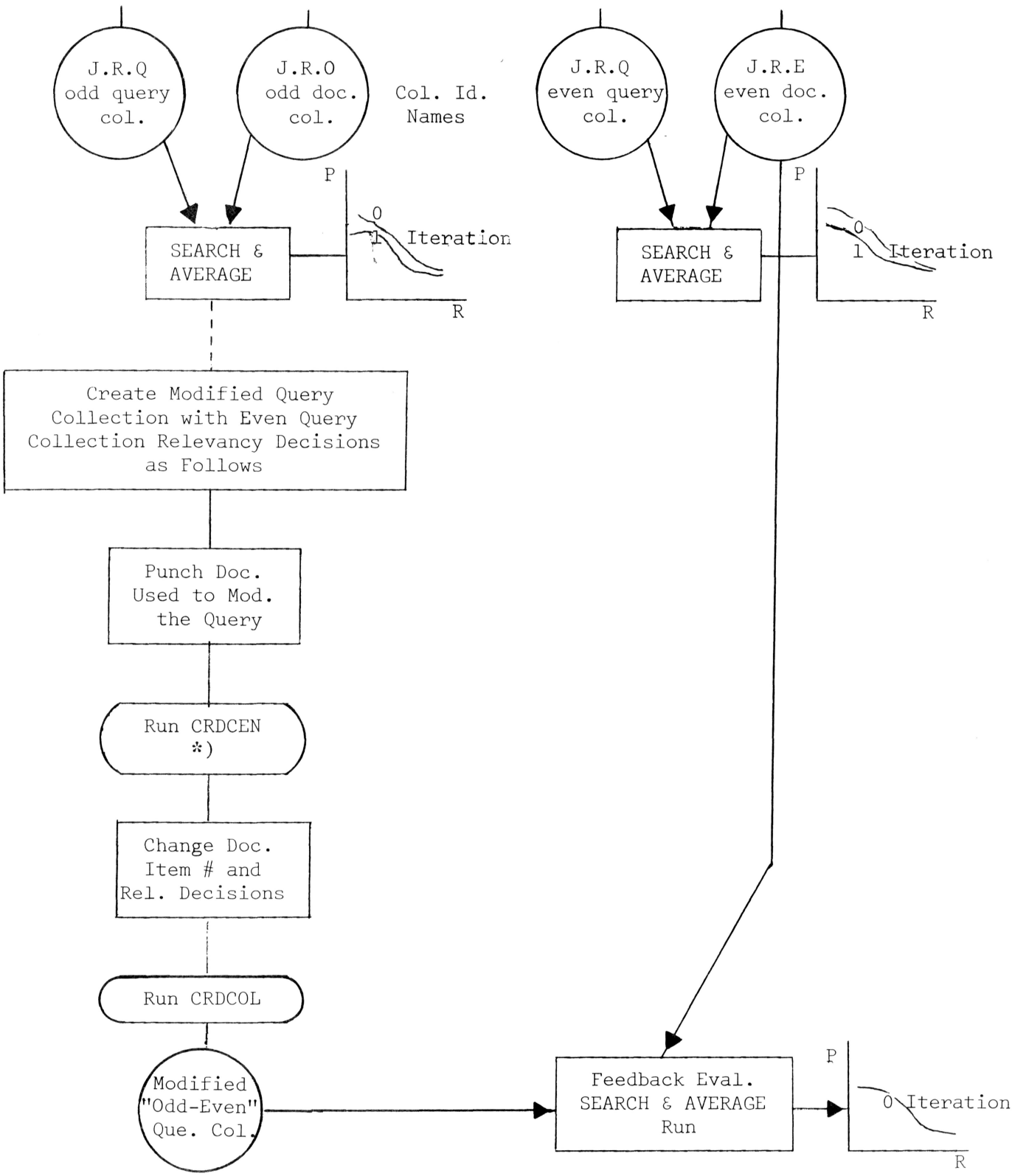
Fig. 11 represents the Recall-Precisicn graph obtained after the initial search (0-1 iterations). The exact curves (of the 0 iteration) printed by SMART appear in Fig. 12 (curves 0,1) and the difference between them is accurately displayed. Observing the results, it is seen that the two subcollections do not seem to be on the average entirely equivalent and for lower recall the Control collection performs better than the Test collection. This means that splitting the collection by odd and even document numbers is not good enough, at least in this case; care must be taken that the differences in generality between the two subcollections are small and evenly distributed. This could be done by shifting some documents back and

The following flowchart describes the above process:



Test and Control Group Evaluation

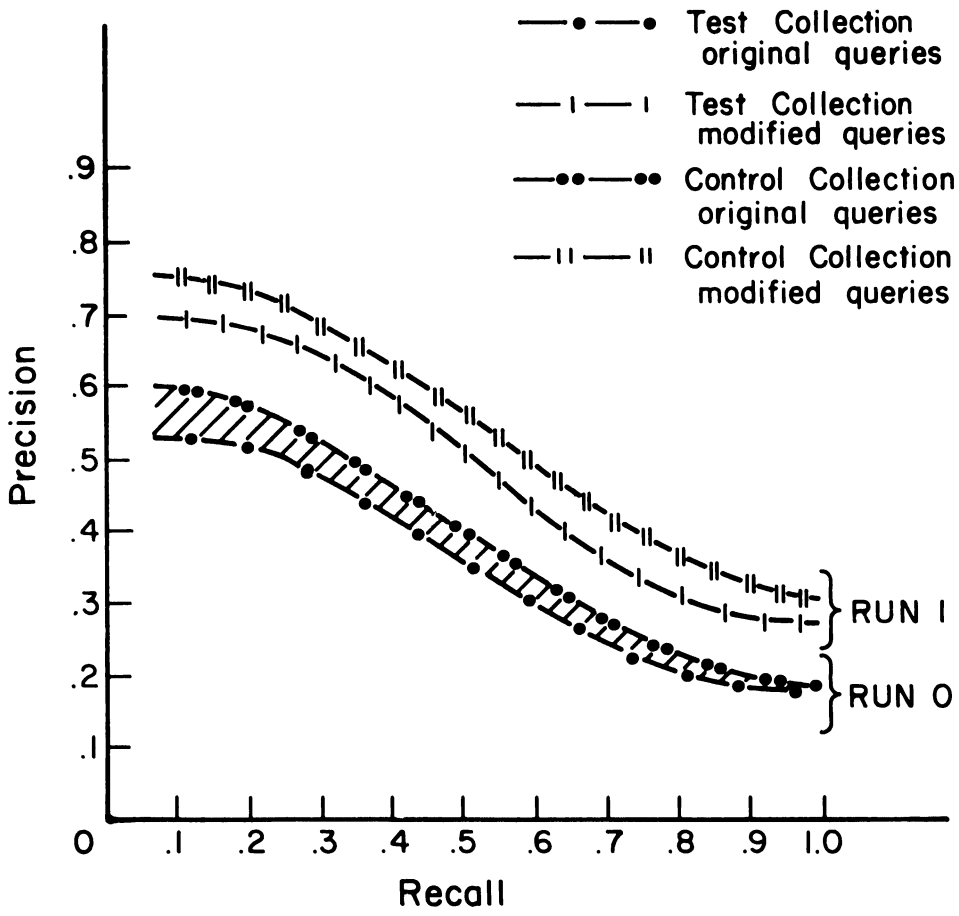
Fig. 10



*) No centroids are created

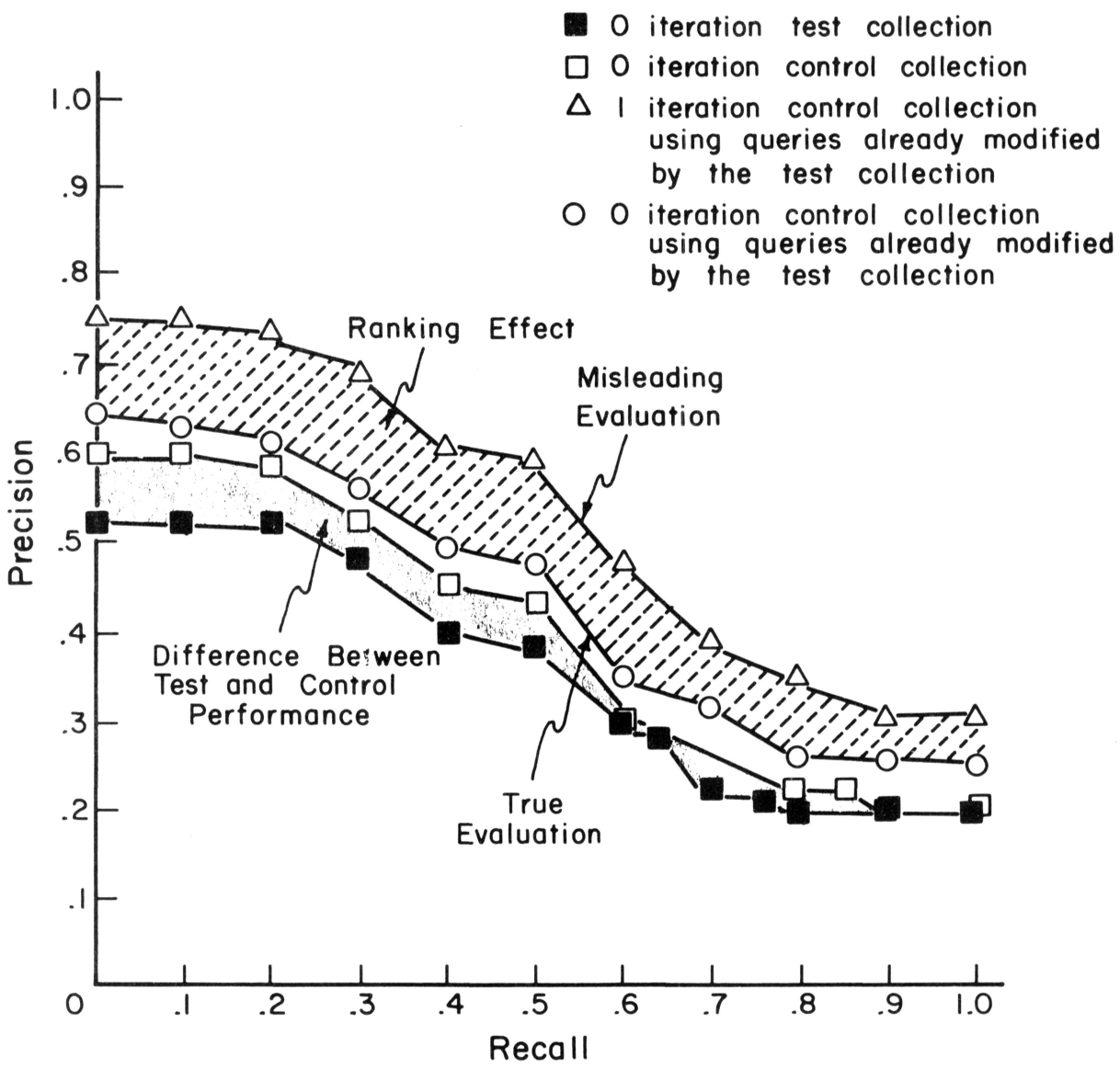
Test and Control Group Evaluation

Fig. 10 (contd.)



Recall–Level Averages 0,1 Iteration

Fig. 11



True Test and Control Evaluation

Fig. 12

forth until the proper distribtuion is found. A better way would be to identify queries that perform much worse in one group compared to the second group and to drop them. No attempt was made to correct the collection groups.

After executing steps c) and d) the last Recall-Precision curve is produced; it appears in Fig. 12 — curve 3 and it represents the true evaluation of the feedback retrieval.

This curve is obtained as a result of a zero iteration full search of the queries which have been modified by the Test collection and thus it is free from any ranking effects and at the same time ranks are assigned beginning from rank #1.

To bring all main results under one set of recall-precision curves an AVERAG run is performed on the results of the different searches. The graph of Fig. 12 contains the last result. Curves 0 and 1 describe the difference between the two collections (the black area). Curve 2 is the Recall-Precision curve obtained after the 1st iteration in the Control group. This is the curve that reflects the total performance of the feedback retrieval and which includes feedback effect as well as ranking effect. Curve 3 is the zero iteration result obtained by applying the modified queries obtained from the Test collection to the Control collection. It is free from ranking effects and reflects the "true" evaluation of the feedback retrieval.

Because of the differences in performance for the two subcollections it may be assumed that the difference between 1 and 3 may be greater for balanced collections. It is interesting to note that the pattern of the curves (2-3) is almost identical and the difference is constant (it is bigger in the 0.0-0.6 recall range and then drops down). This can be ex-

plained by the fact that ranking effects are on the average constant and the differences between 2 and 3 are due to this effect.

Another interesting phenomenon is the fact that curve 2 of Fig. 12 (1st iteration Control collection using queries already modified by the test collection) is almost identical to the 1st iteration search result using the original queries with the control collection. The fact that both of them are raised to the same level means that the performance of the queries modified by the different collections is on the average almost the same.

4. Conclusions

The experiment described above does show that test and control groups can be used for evaluating feedback retrieval. The fact that different collections are used for the evaluation is the main advantage of this method, since this permits the use of total performance as a measure of the feedback retrieval. More care should be taken in splitting the original collection in order to ensure more accurate results.

This method is effective principally as a tool for comparing the performance of different algorithms used to modify queries.

References

[1] Eleanor Rose Cook Ide, "Relevance Feedback in an Automatic Document Retrieval System", Master Thesis, Report ISR-15 to the National Science Foundation, Cornell University, Department of Computer Science, January 1969.

[2] Harold A. Hall and Nelson H. Weiderman, "The Evaluation Problem in Relevance Feedback Systems", Report ISR-12 to the National Science Foundation, Cornell University, Department of Computer Science, Section XII, June 1967.

[3] G. Salton, Automatic Information Organization and Retrieval , McGraw Hill Inc., 1968, Chapter 8.