VIII. New Experiments in Relevance Feedback

E. Ide

Abstract

New results are given for interactive user controlled retrieval strategies, using relevance feedback. Search strategies are included, based on the identification by the user of a fixed or variable number of relevant or nonrelevant documents. The evaluation results are based on experiments using 200 documents and 42 queries in the field of aerodynamics.

1. The Relevance Feedback Procedure

Automated information retrieval systems, like most mechanical processes, suffer from unavoidable inflexibility. The needs of users of a large information collection, especially a document collection, are too varied to be satisfied with any one full automatic retrieval algorithm. Users whose needs best match the assumptions built into the system are satisfied; others are not.

One suggested way to overcome this limitation is to employ feedback information from the user during the retrieval process. In a document retrieval situation, this could be accomplished as follows:

a) The user poses a request to the retrieval system.

b) The retrieval system returns some information (perhaps abstracts) about a specified number of documents judged relevant to the user's request.

c) The user selects from this set of initially retrieved items those documents which he deems relevant to the request, and feeds this information to the retrieval system.

d) Another retrieval search is performed incorporating these user judgments.

Steps c) and d) are iterated as often as desired.

Such an interactive process was proposed by Rocchio, who called it "relevance feedback" [1,2,3]. He showed that in a document retrieval system based on classification and using the cosine correlation function, the theoretically optimum query for retrieving a set of documents $R = \{r_i\}$ is given by the formula:

$$Q_{opt} = \frac{1}{n_r} \sum_{1}^{n_r} \frac{r_i}{|r_i|} - \frac{1}{n_s} \sum_{1}^{n_s} \frac{s_i}{|s_i|}$$

where: $R = \{r_i\}$ = the descriptor vectors of all documents in the collection which are relevant, according to the user, to the request;

$S = \{s_i\}$ = the descriptor vectors of all other documents in the collection, i.e. of all nonrelevant documents;

n_r = number of relevant documents in the collection;

n_s = number of nonrelevant documents;

$|r_i|$, $|s_i|$ = length of the document descriptor vectors r_i, s_i .

Of course, the sets R and S are not known to the system. On each iteration, however, the user feedback supplies information about two subsets, $R' \varepsilon R$ and $S' \varepsilon S$, where $R' = \{r'_i\}$ is the subset of relevant documents

retrieved and $S' = \{s'_i\}$ is the subset of nonrelevant documents retrieved. Therefore, the following formula is used by Rocchio to construct a new query from the query of the previous iteration:

$$Q_{i+1} = Q_i + \frac{1}{n'_r} \sum_{1}^{n'_r} \frac{r'_i}{|r'_i|} - \frac{1}{n'_s} \sum_{1}^{n'_s} \frac{s'_i}{|s'_i|} \qquad \text{(A)}$$

where n'_r (n'_s) is the number of relevant (nonrelevant) documents retrieved for feedback in the previous iteration.

Rocchio investigated relevance feedback using the above formula and 17 queries in the field of computer science and found that his algorithm does improve retrieval results [1,2,3].

Another investigation of a relevance feedback system was based on the "ADI collection", a collection of 82 documents presented at a conference on documentation. Thirty-five queries were constructed for this collection, and the documents considered relevant to those requests were specified by the two originators of the queries. The investigation of relevance feedback in the ADI collection was conducted by Riddle, Horwitz, and Dietz [4]. They used 22 of the 35 queries and studied a slightly different algorithm for modifying the search query using the following method of query modification:

$$Q_{i+1} = Q_i + \alpha \sum_{1}^{n_r} r'_i$$

2. The Experimental Environment

The document collection used in this study (the "Cranfield" collection) contains 200 documents from the field of aerodynamics, chosen from a library of 1400 documents. For this collection, there exist at present 42 queries, constructed by some of the authors of the 1400 documents; these requestors are also responsible for the relevance judgments.

The concept vectors describing document and queries are quite sparse for this collection. The maximum number of concepts used to describe one document is 85, out of a possible 552 concepts. The largest weight given to any concept in any document descriptor is 288. The query description vectors are sparser by one order of magnitude and shorter than the document descriptors. The maximum number of concepts used in a single query vector is 13; the largest weight in any query vector is 24. The largest number of documents relevant to a single query is 12, or six percent of the collection. The comparative brevity of the query vectors in this collection is typical for technical document retrieval, and provides a strong argument for the use of relevance feedback techniques. With relevance feedback, the user, in effect, provides a much more detailed query merely by citing a document; the document descriptor itself is used as the query.

The relevance feedback system being studied uses the following query-update formula:

$$Q_{i+1} = \pi Q_i + \omega Q_o + \alpha \sum_{1}^{\min(n_a, n'_r)} r_i + \mu \sum_{1}^{\min(n_b, n'_s)} s'_i \qquad (B)$$

where $n'_r + n'_s$ (see equation A, section 1) equals N, the number of documents retrieved for feedback.

The experimental variables are α, ω, π, μ, n_a, n_b, and N. The parameter α is positive, and weights all incoming relevant documents relative to the other contributors to the query (previous query, initial query, non-relevant documents). The parameter π permits the previous query to be increased in weight relative to the incoming documents. Q_o is the initial query, as opposed to the query of the previous iteration; ω permits the initial query to be used as part of the new query (see section 3B). The parameter μ should theoretically be negative, as it permits some significance to be attached to the nonrelevant documents retrieved. The parameter n_a (n_b) permits some specific number of relevant (non-relevant) documents to be used in the query even if n'_r (n'_s) is larger. It is assumed that the r'_i and s'_i are indexed in order of decreasing relevance (as determined by the system) to the query; that is, the n_a relevant documents (or n_b nonrelevant documents) used in the new query will be those closest in the descriptor space to the previous query. The flexibility of this formula permits the investigation of several feedback strategies.

The system also provides the following formula to simulate Rocchio's algorithm:

$$Q_{i+1} = \pi\, n'_r\, n'_s\, Q_i + \omega\, Q_o + n'_s \sum_{1}^{\min(n'_r,\, n_a)} r_i - n'_r \sum_{1}^{\min(n'_s,\, n_b)} s_i \qquad \text{(C)}$$

Formula (C) does not, however, normalize the vector lengths as does Rocchio's algorithm.

The document and query description vectors for both collections were constructed using a SMART thesaurus dictionary on the document abstracts and the queries [3]. The cosine correlation function is used to determine the order of retrieval.

3. Earlier Results in the Same Environment

An earlier study[5] uses the Cranfield 200 document collection and the relevance feedback system described in section 2. Three major variations in relevance feedback strategy are investigated:

1) The parameters π, ω, and α in formula B of section 2 are varied, holding μ equal to 0. This strategy is similar to the type investigated by Riddle, Horwitz, and Dietz [4]. The variation in results obtained when π, ω, and α are changed is slight, in fact, less than the variation found by Riddle, Horwitz, and Dietz, who used a different document collection.

2) The number of documents retrieved for feedback (N) is varied. N is set to 5, 10, and 15 documents. The improvement obtained by feeding back 10 documents instead of 5 is impressive; the further improvement obtained with 15 documents is less so.

   In addition, a "variable feedback" strategy is investigated, in which documents are retrieved until one relevant document is found, or until 15 documents have been retrieved. This strategy provides greatly improved precision at low recall but only slightly better precision at high recall than does the N = 5 strategy.

3) Two strategies using the information in retrieved nonrelevant documents are studied. In formula B (section 2), μ is set to minus 1 and n_b is set to 1 and 2. That is, the first nonrelevant document (the first two nonrelevant in the second strategy) is subtracted from the query. The strategy in which $n_b = 1$ (called "Dec Hi"), with N equal to 5, produces a performance comparable to that of a strategy using only relevant documents with N equal to 10.

The following conclusions may be reached from the results of the earlier study: The investigation supports relevance feedback as an information retrieval strategy. It also shows that varying the parameters in a query-update formula which uses relevant documents only is not a promising way to produce significant improvement in performance. The most promising strategies investigated, variable feedback and nonrelevant document feedback, require further study before they can be firmly recommended. The variable feedback strategy and the suggested combination of fixed and variable feedback should be investigated in a suitable evaluation system. The nonrelevant feedback strategies should be studied in a system which permits normalizing as in Rocchio [1]. Eventually, some combination of fixed and variable feedback may prove optimal in similar information retrieval environments.

4. Evaluation of Retrieval Performance

A) The "Feedback Effect" in Evaluation

The investigation described in section 3 uses a retrieval and evaluation method that has been assessed by Hall and Weiderman [6]. After

each iteration, all documents in the collection are ranked and the top-ranked N documents are used for feedback. Hall and Weiderman point out that evaluation of this retrieval technique takes into account two effects, which they call "ranking effect" and "feedback effect".

Relevance feedback in effect uses information from one or more document descriptors to modify the query descriptor. The relevant documents used for this purpose will be ranked higher by the modified query than previously, and the nonrelevant documents used will be ranked lower. The effect of these rank changes in "retrieved" documents is termed the "ranking effect". If the ranking effect is included in an overall performance measure, the measured change in performance between feedback iterations is quite impressive, as is seen in the results included in the earlier report described in section 3 [5]. This large change in "total performance" (including both ranking and feedback effect) indicates the extent to which the initial query has been perturbed toward the centroid of the relevant documents, and strongly supports Rocchio's theory.

Hall and Weiderman state that in an environment where the user must actively supply relevance judgments for feedback, changes in the ranks of documents which the user has already seen are of no interest to him. The user in such an environment is concerned primarily with the "feedback effect"; that is, the effectiveness of the modified query in bringing new relevant documents to his attention. They conclude that, though total performance is a valid measure of the effectiveness of relevance feedback in approaching the "ideal query", the feedback effect should be isolated and examined as well.

The present study evaluates feedback performance in the manner suggested by Hall and Weiderman by discarding the ranking effect and preserving only the feedback effect. The ranks of the top N documents retrieved in each iteration (the documents used for feedback) are "frozen" in all subsequent iterations, and only the remainder of the collection is searched using the modified query. Thus, in the present investigation, the N documents retrieved on any iteration are guaranteed to be N new documents; that is, documents not used for feedback on any previous iteration. Moreover, the performance measures for the first (second, third) iteration are calculated from a ranked document list in which the top N (2N, 3N) documents are the same as those retrieved previously. Only the changes in the ranks of documents not yet seen by the user is measured.

The evaluation described gives overall results that are deceptively low. Because the top ranks are frozen, no newly retrieved document can achieve a rank higher than that of any previously retrieved document. With a constant feedback strategy, therefore, on the first (second, third) iteration, the highest possible rank for a new document is N+1 (2N+1, 3N+1). For this reason, the feedback effect evaluation is a misleading measure of the performance of the retrieval system, and should be used in conjunction with other evaluation methods. Isolation of the feedback effect is primarily useful to compare different feedback strategies from the viewpoint of a user in an interactive retrieval environment.

B) Performance Measures

Several average measures of the performance of the tested retrieval algorithms on the 42 Cranfield queries are used in this report. Each measure is based on the concepts of "recall" and "precision". In evaluating an information retrieval system, an arbitrary cut-off is often employed, and documents above this cut-off are termed "retrieval". With such a cut-off, "recall" is the percentage of documents relevant to the user that are retrieved, and "precision" is the percentage of retrieved documents that are relevant. The average measures used in this study do not employ a cut-off, but evaluate the retrieval performance over the entire document collection. A discussion of the generalization of the concepts of recall and precision to such an overall evaluation is found in reference 7.

The average measures used herein are: Rank recall, log precision, normalized recall, normalized precision, and a curve reflecting average precision at each 10% of recall. The first 4 of these measures are defined in reference 7. The recall-precision curve used here differs from any used in previous studies.

Both the Quasi-Cleverdon curve, used in the earlier study (section 3), and the new curve used here are average plots of precision at each 5% or 10% of recall. Each query is averaged into each point of the plot. To accomplish this averaging process, an interpolation procedure is needed, since, for example, a query with two relevant documents can only achieve uninterpolated recall levels of 50% and 100%. The Quasi-Cleverdon curve and new curve are distinguished by the method of interpolation used.

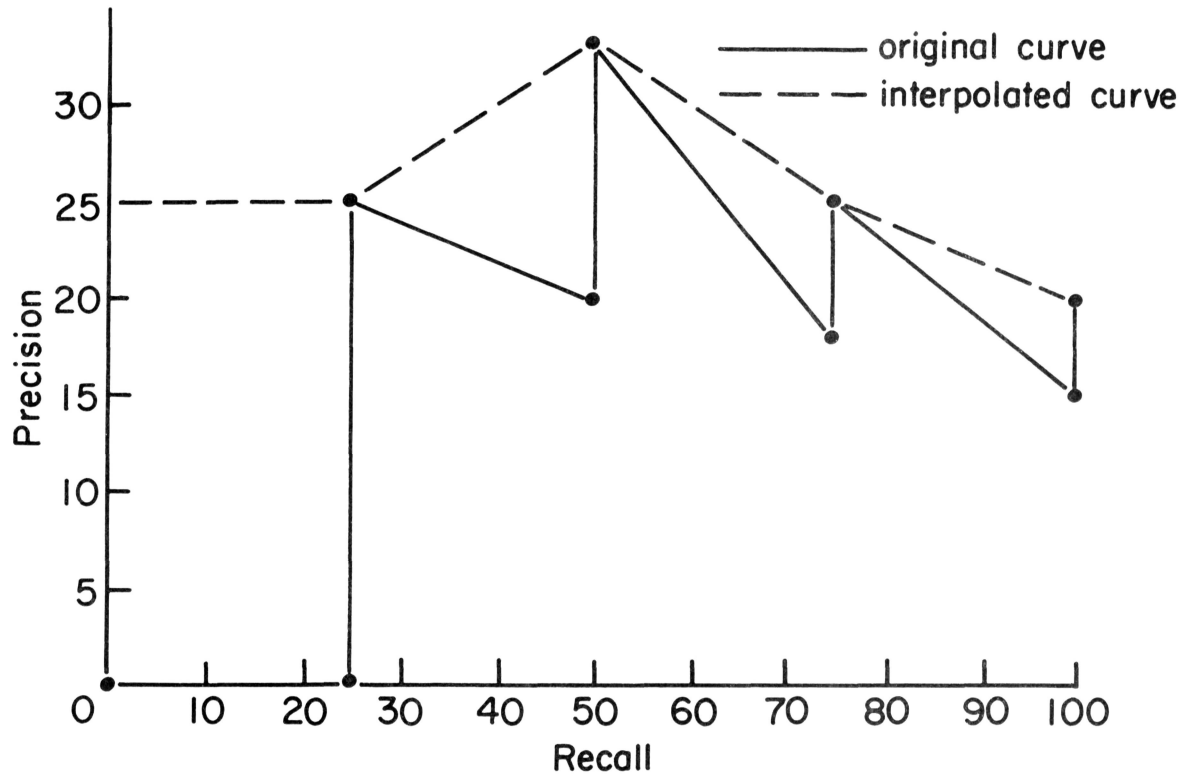
Figs. 1 and 2 show two graphs for a hypothetical query having 4 relevant documents. The relevant documents are assumed to be retrieved with ranks of 4, 6, 12 and 20. Thus, at 25% recall, the precision is 25%, at 50% recall, the precision is 33%, and so on. However, these values correspond actually to the highest possible precision points, since they are calculated just after a relevant document is retrieved. In this example, after 3 documents are retrieved, the precision is 0%, after 5 documents, the precision is 20%, and so on. This range of precision for each recall level is indicated by the top and bottom points in Figs. 1 and 2 at 25%, 50%, 75%, and 100% recall. The solid saw-tooth line connecting these points is not used for interpolation; it is rather intended to indicate the drop in precision between the actual recall levels for this query, as more nonrelevant documents are retrieved.

The Quasi-Cleverdon averages use a straight-line interpolation between peak points of precision, as indicated by the dashed line in Fig. 1. It has been argued that this interpolation is artificially high, since it lies at all points above the sawtooth curve, and thus, does not reflect,in any way, the precision drop as more nonrelevant documents are retrieved. The new averages of Fig. 2, use an interpolation that projects a horizontal line leftward from each peak point of precision, and stops when a higher point of precision is encountered. This new interpolation curve (the dashed line in Fig. 2) does <u>not</u> lie above the sawtooth curve at all points. When the precision drops from one recall level actually

achieved to the next, an immediate drop in precision after the first point to the level of the next point is indicated. For example, in Fig. 2, the precision value at 50% recall is 33%, but at 55% recall, the interpolated value used for the new averages is 25% precision. When the precision rises from one recall level to the next, however, the first precision point actually achieved is ignored for purposes of interpolation. The achieved precision of 25% at 25% recall in the example of Fig. 2 is ignored, and for all recall levels from 0 to 50%, an interpolated precision of 33% is used for the new averages. The proponents of the new interpolation argue that this method indicates in all uses a precision that the user could actually achieve, if he were to use hindsight by retrieving exactly the right number of documents. The new averages are now used both at Harvard and at Cornell in evaluating the SMART system.

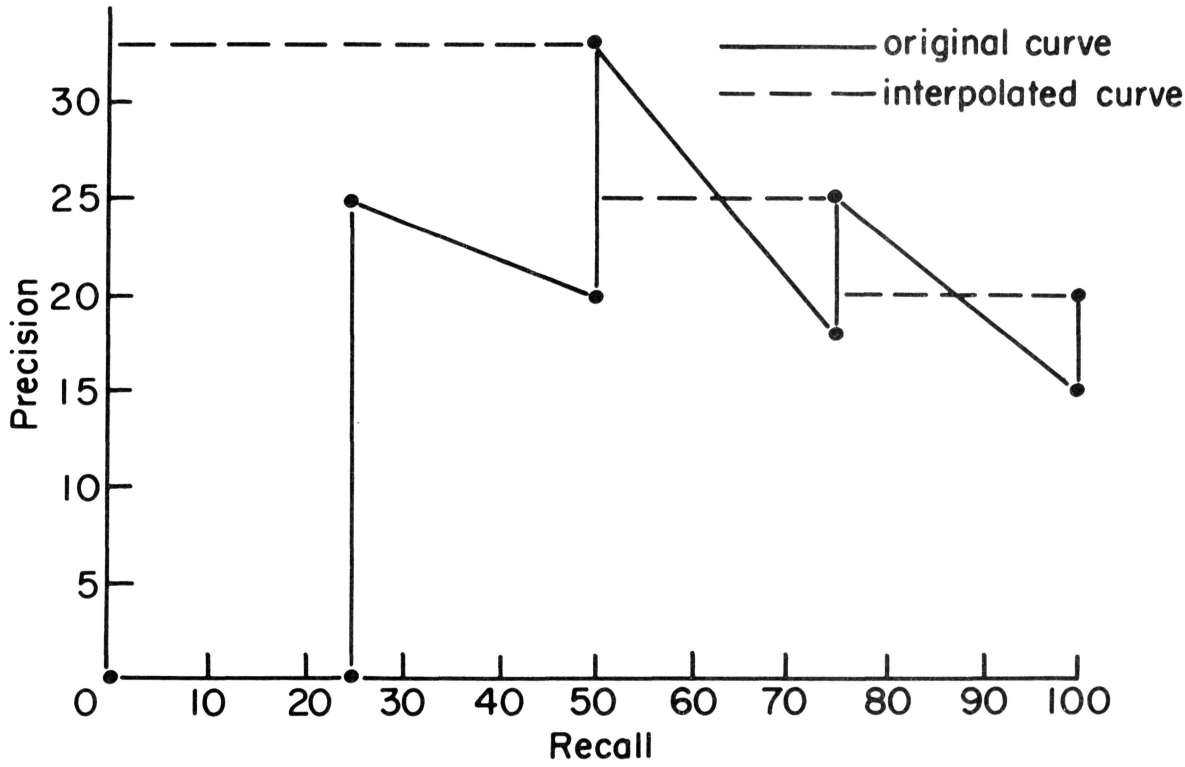
C) Statistical Tests

Several statistical tests are reported here using as input the rank recall, log precision, normalized recall, normalized precision, and 10 points from the new recall-precision curve. The statistical tests are intended to measure the "significance" of the average difference in values of these measures obtained for two iterations or two distinct search algorithms. The test results are expressed as the probability that the two sets of values obtained from two separate runs are actually drawn from samples which have the same characteristics. A small probability value thus indicates that the two curves are actually significantly different. If this probability for one measure is, for example, 5%, the difference in the two average values of that measure is said to be "significant at the 5% level".



An Illustration of the Interpolation Method Used for the Quasi-Cleverdon Recall-Precision Averages

Fig. 1



An Illustration of the Interpolation Method Used for the New Recall-Precision Averages

Fig. 2

Choice of a statistical method for calculating this probability is important. The present study uses two statistical tests, the familiar T-test and the Wilcoxon Signed-Rank Test (WSR) [8]. The T-test takes account of the magnitude of the differences, and assumes that the measures tested are normally distributed. The WSR test does not make this assumption. However, the WSR test takes account only of the ranks of the differences, ignoring their magnitude. Because this test does not assume normality of the input and because it ignores some information (magnitudes of differences), the WSR test is more conservative than the T-test. It is therefore less prone to the error of calling a result "significant" when it is not. Because information retrieval provides discrete rather than continuous data, and because only 42 data points (42 queries) are provided, the more conservative WSR test is preferable in the present evaluation.

5. Experimental Results

Results of three major areas of investigation are presented:

a) A comparison of two strategies that use only R', the set of relevant documents retrieved, to modify the query vector.

b) An investigation of the retrieval effect of the number of documents used for relevance judgment feedback on each iteration.

c) An investigation of strategies using the set S', that is, the nonrelevant documents retrieved, to modify the query. The statistical significance of the average results obtained is tested in each case.

A) Two Strategies Using Relevant Documents Only

In the earlier report, summarized in section 3 [5], several strategies using relevant documents only are compared. The differences in total performance found among these strategies were very slight. A feedback effect comparison of two "relevant only" strategies is made here. The strategies chosen are:

1) The straightforward strategy of setting α equal to 1, π equal to 1, and the other constants equal to 0, in formula B (section 2). The feedback formula in effect for this strategy is therefore:

$$Q_{i+1} = Q_i + \sum_{1}^{n'_r} r_i \quad .$$

This formula is not equivalent to any strategy used in the earlier study because the feedback effect evaluation provides new documents for feedback on each iteration (section 4A), while the total performance evaluation does not. The nearest comparable strategy from the previous study is the so-called "Q_o strategy" (also called "due only").

2) A strategy that gives added weight to the user's original query: $\pi = 1$, $\alpha = 1$, $\omega = 4$. This strategy is intended to compensate for the large difference in magnitude between document vector weights and query vector weights (section 2).

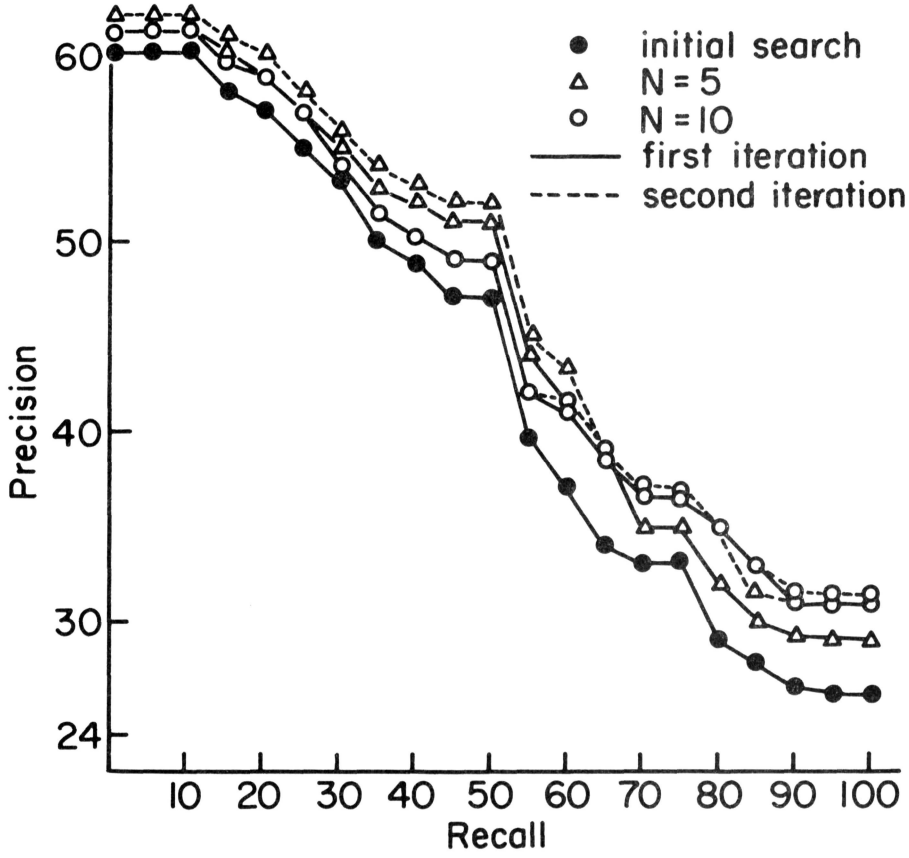
The difference in feedback effect between these two methods is trivial. For all average measures, the differences observed are less than one-and-a-quarter percent. The recall level averages for the second strategy, called "Q_o+", are presented in Fig. 6 to be described later. This results is hardly surprising, since for several "relevant only" strategies, the total performance results reported in the earlier study are nearly identical.

B) Varying the Amount of Feedback

In the earlier study (section 3), the improvement in total performance achieved by feeding back ten rather than five documents to the user is impressive. This difference, however, is primarily due to the ranking effect. The feedback effect results, shown in Fig. 3, are actually better at medium recall levels when only five documents are used for feedback. At high recall levels, the performance ahieved by feeding back five documents twice is roughly equal to that obtained by feeding back ten documents once. The average improvement in feedback effect gained by feeding back more documents on each iteration does not seem worth the cost of asking the user to make more relevance judgments.

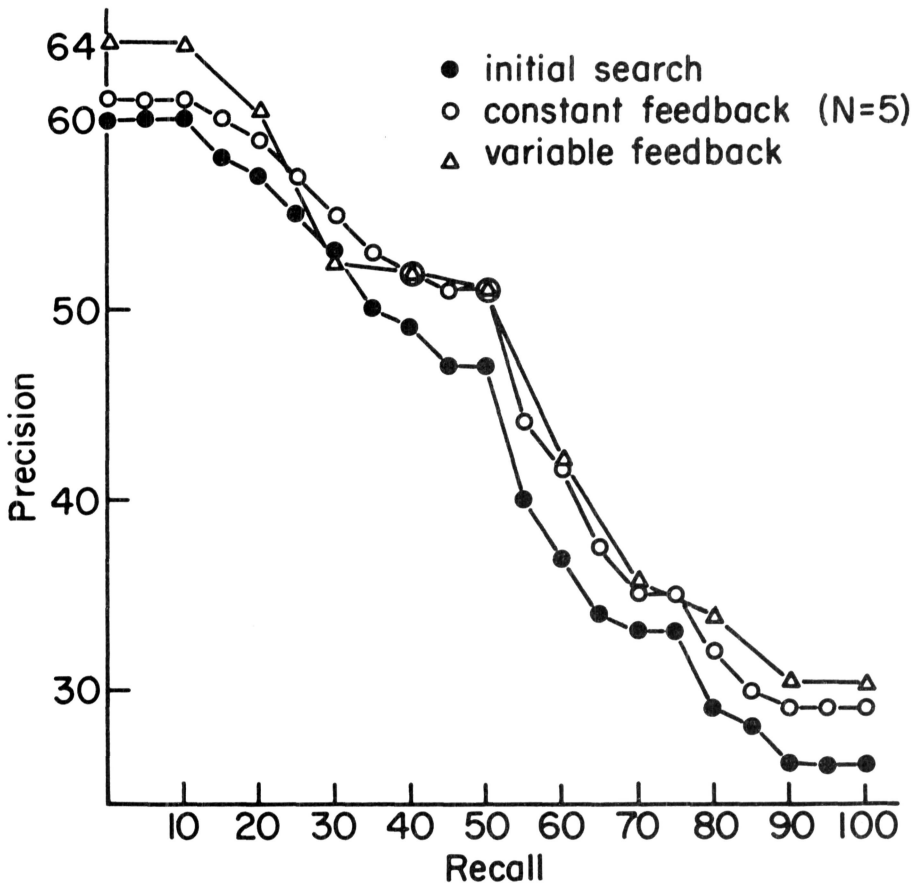
It must be noted, however, that the feedback effect evaluation gives an unfair advantage to runs using few documents for feedback. When five documents are used for feedback, ranks 1-5 are frozen on the first iteration and ranks 1-10 on the second. When ten documents are fed back, however, ranks 1-10 are frozen on the first iteration and ranks 1-20 on the second. The difference in results caused by increasing the number of documents fed back is therefore minimized by the evaluation process.

A variable feedback strategy is tested in the earlier study. The user is asked to search the retrieved list from the top until he finds one relevant document or has seen fifteen documents. The total performance reported indicates that the user who does not require high recall (50% recall or less) reaps considerable benefit from this strategy, but that high recall performance is very little better than that obtained by a constant feedback with N equal to 5. The feedback effect evaluation produces



Increment Only Feedback of Five or Ten Documents

Fig. 3



Constant Feedback Compared to Variable Feedback - First Iteration

Fig. 4

a different picture. In Fig. 4, one iteration of variable feedback is compared to one iteration feeding back 5 documents. The "average user" must look at 4 documents from the initial search to find one relevant document. The feedback effect performance at 10% recall is better for the variable feedback strategy. At 20% recall, the two strategies are the same, at 30% variable, feedback is worse. At medium recall levels, the two strategies are again approximately the same. At recall levels above 70%, however, the variable feedback strategy gives higher precision.

The variable feedback results are also affected by the feedback effect measure. Since for 75% of the queries five or fewer documents are used for feedback, variable feedback receives the "unfair advantage" noted earlier. This advantage should, however, be most evident at low recall, whereas the major improvement observed here is at high recall.

The apparent inconsistency of a large total performance improvement (from constant feedback to variable feedback) at low recall, with a feedback effect improvement at high recall is easily explained. As was mentioned in section 4A, the feedback effect evaluation prevents a document retrieved by feedback from achieving a higher rank than any document retrieved earlier. The total performance improvement at low recall indicates that relevant documents retrieved early (including the first relevant document retrieved, which is used for feedback) make larger jumps in rank with variable feedback than with constant feedback. These jumps are inhibited by the artificial rank ceiling imposed by the feedback method of evaluation. The feedback effect shows improvement at high recall, indicating that even at high recall levels, new relevant

documents appear sooner in the retrieved list with variable feedback than with constant feedback. The total performance results favor variable feedback primarily for the user who does not require high recall. The feedback effect results support this type of variable feedback as a strategy to be generally recommended in an interactive environment.

Two further investigations must, however, be considered. First, the performance of several iterations of the variable feedback strategy should be investigated using the feedback effect evaluation. Second, the results presented here are valid for a hypothetical "average" user. An examination of subgroups of queries would show whether or not certain types of users achieve better results with the constant feedback strategy.

C) Strategies Using Nonrelevant Documents

In the earlier report (section 3), a strategy using nonrelevant documents displays a total performance similar to that achieved using relevant documents only and using twice as many documents for feedback. This nonrelevant document strategy, called "Dec Hi" in both the earlier and the present reports, uses the retrieval formula:

$$Q_{i+1} = Q_i + \sum_{1}^{n'_r} r_i - s_1 \quad ,$$

where s_1 is the first nonrelevant document retrieved.

In the previous study, it was recommended that Rocchio's relevance feedback strategy, a strategy using all documents retrieved, be tested.

The present study tests Rocchio's strategy, but without normalizing the lengths of the documents used to modify the query. Since the cosine correlation is used to rank documents for retrieval, this normalization is a theoretical necessity [2].

The results of the significance tests on the three strategies, Q_o+ (see section 5), Rocchio (Rocchio's algorithm without normalization), and Dec Hi are given in Fig. 5. Fig. 5 shows the significance levels of the differences among the three strategies for two iterations, using the less conservative T-test. It is evident that the differences in averages among the three strategies described are not significant.

Two comparisons are of particular importance: The differences in normalized recall between the "relevant only" strategy, Q_o+, and the two nonrelevant document strategies, Rocchio and Dec Hi, on the first iteration. The five percent difference between Q_o+ and Dec Hi is significant at the 6% level, and the six percent difference between Q_o+ and Rocchio is significant at the 3% level, according to the T-test. However, the Wilcoxon Signed-Rank Test (WSR) indicates that the two algorithms do not give significantly different results. The significance level comparing Q_o+ and Dec Hi is 46%, and that comparing Q_o+ and Rocchio is 48%

These different significance levels must be considered in the light of the characteristics of the two significance tests. The T-test takes account of magnitude, the WSR test considers only rank. Evidently, differences favoring Q_o+ and differences favoring the nonrelevant document strategy are mixed in rank, producing "insignificant" results for the WSR test. Yet, some of the results favoring Q_o+ (not <u>all</u>, because the ranks are mixed) must be very large in magnitude, to give "significant"

	Rocchio versus Q_o+				Dec Hi versus Q_o+				Dec Hi versus Rocchio			
	Iteration 1		Iteration 2		Iteration 1		Iteration 1		Iteration 1		Iteration 2	
Rank Recall	-0.1	99.8	.1	96.1	.3	76.8	-.4	81.3	.4	61.3	-.5	55.9
Log Pre.	-0.4	61.7	-.2	81.0	.0	93.3	-.3	83.7	.4	32.9	.0	94.8
Normed. Rec.	-6.1	3.4*	-4.4	17.7*	5.4	5.7*	-7.0	9.4*	.6	75.8	-2.6	36.4
Normed. Pre.	-2.9	15.4	-2.0	43.0	-2.1	31.4	-3.0	29.6	.8	55.0	-1.0	53.8
Rec. Level 10%	.0	99.4	-.2	87.9	.2	84.2	-.2	89.0	.2	66.1	.0	96.6
20%	.3	83.4	.3	83.5	.6	57.4	.3	81.4	.4	48.9	.0	93.8
30%	.1	94.7	.2	89.1	.6	61.3	.3	83.8	.5	49.8	.1	85.0
40%	-.1	93.6	.4	78.9	.7	58.1	.9	56.1	.8	29.2	.5	40.8
50%	.1	92.3	.6	69.8	.9	48.5	.8	60.7	.8	24.7	.2	71.8
60%	-.5	95.4	-1.2	63.7	.8	55.5	.8	62.6	1.3	39.7	2.0	11.3
70%	.7	68.0	1.0	54.6	.4	61.4	.7	51.2	-.3	92.3	-.3	94.6
80%	1.1	42.5	.9	50.8	1.1	42.7	.5	70.0	-.1	98.3	-.5	47.6
90%	.7	55.8	.0	95.1	.6	67.9	-.1	50.9	-.2	77.4	-1.1	16.2
100%	.6	66.4	.0	92.3	.5	72.5	-.8	60.3	-.1	89.5	-.8	31.3

* Significance of Normalized Recall for WSR test:

Rocchio versus Q_o+	Iteration 1	24.1
	Iteration 2	48.9
Dec Hi versus Q_o+	Iteration 1	98.5
	Iteration 2	45.6

Q_o+: $Q_{i+1} = Q_i + \Sigma r_i + 4Q_o$

Rocchio: $Q_{i+1} = n'_r n'_s Q_i + n'_s \Sigma r_i - n'_r \Sigma s_i$

Dec Hi: $Q_{i+1} = Q_i + \Sigma r_i - s_1$

Comparing Three Feedback Strategies, First and Second Iterations
Differences and Significance Levels

Fig. 5

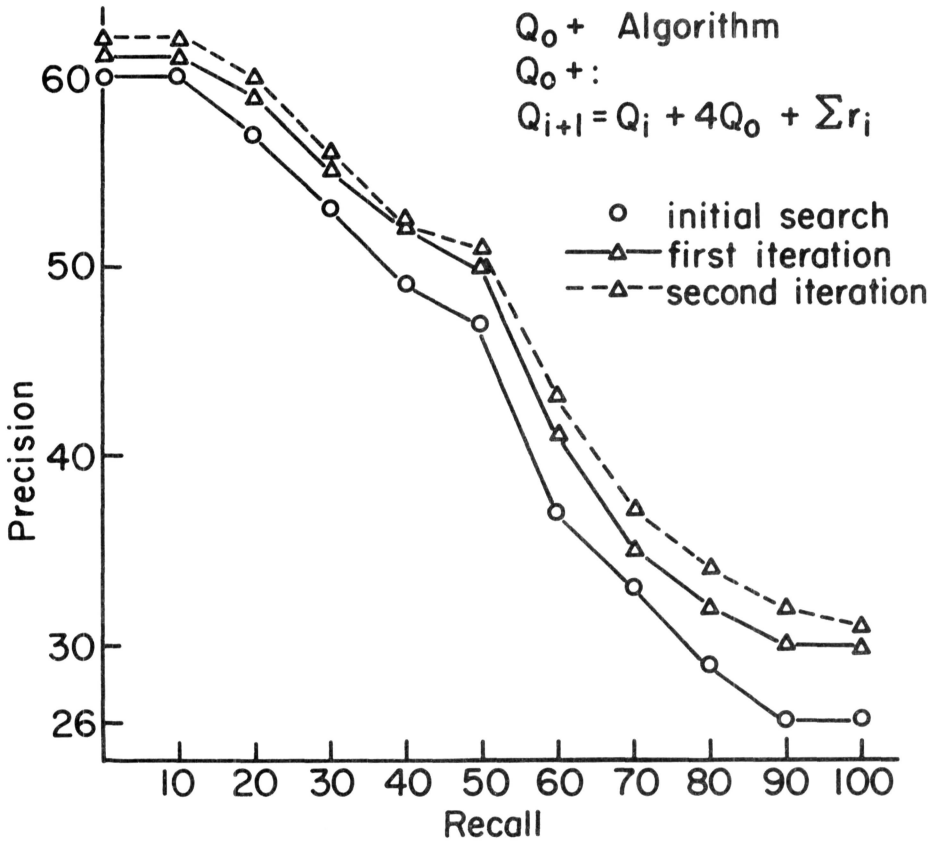
indications on the T-test. Thus, for a few queries, the Rocchio and Dec Hi algorithms must be much less effective than Q_o+ as measured by normalized recall.

The total performance normalized recall measured in the previous study (section 3) was also low for Dec Hi, compared with a "relevant only" strategy. The explanation given in the earlier report is equally valid for the feedback effect results reported here. In brief, the use of nonrelevant documents for feedback seems to raise the ranks of fairly high-ranking relevant documents, and at the same time, lower the ranks of some low-ranking relevant documents.

The significance levels obtained by comparing the first and second iteration results to the initial search result, within each of the three strategies, are very informative. Figs. 6, 7, and 8, show the performance of algorithms Q_o+, Rocchio, and Dec Hi respectively. The significance of the gap between the initial search and each iteration is tested, using the more conservative WSR test.

Looking at the three recall-precision graphs, the average performance of the three algorithms seems quite similar. Fig. 5 shows that, in fact, the differences in average performance are not significant. Yet, the significance levels displayed in Fig. 6 differ greatly from those displayed in Figs. 7 and 8.

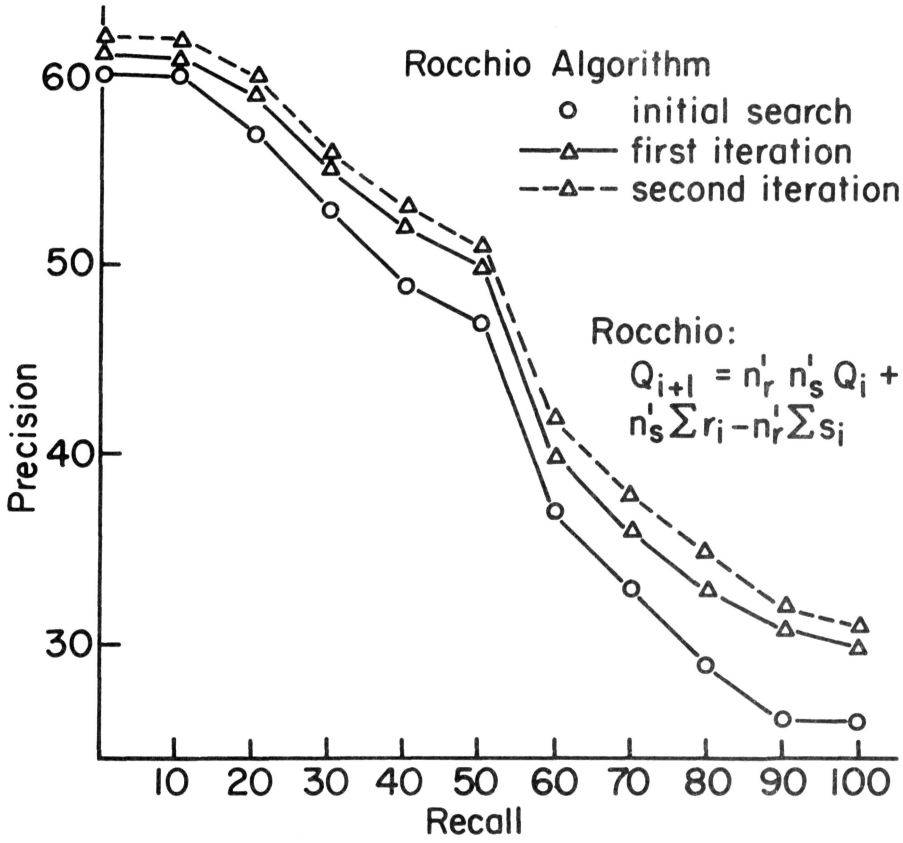
For the Q_o+ strategy, the differences between the initial search and each feedback iteration are significant. On the first iteration, the four overall measures and the precision differences from 20% through 50% recall are significant at the 5% level or less, and only at 70 and 80%



	First Iter. vs. Initial Search		Second Iter. vs. Initial Search	
	Percent Difference	Percent Significance	Percent Difference	Percent Significance
Rank Recall	4.1	3.8	5.1	1.9
Log Precision	2.3	4.1	2.8	1.2
Normalized Recall	3.1	0.6	3.7	0.1
Normalized Precision	2.7	1.8	3.6	0.5
Recall Level 10%	1.0	6.8	1.8	1.3
20%	1.8	4.3	2.5	0.8
30%	2.2	1.5	3.0	0.9
40%	3.0	0.5	3.7	0.4
50%	3.1	0.8	3.7	1.1
60%	3.5	5.7	5.7	2.8
70%	2.4	29.2	4.4	19.5
80%	3.2	18.3	4.9	4.5
90%	3.6	6.1	5.6	2.4
100%	3.6	6.1	5.2	2.6

Comparison of First and Second Iterations to Initial Search
Differences and Significance Levels - Q_o+ Algorithm

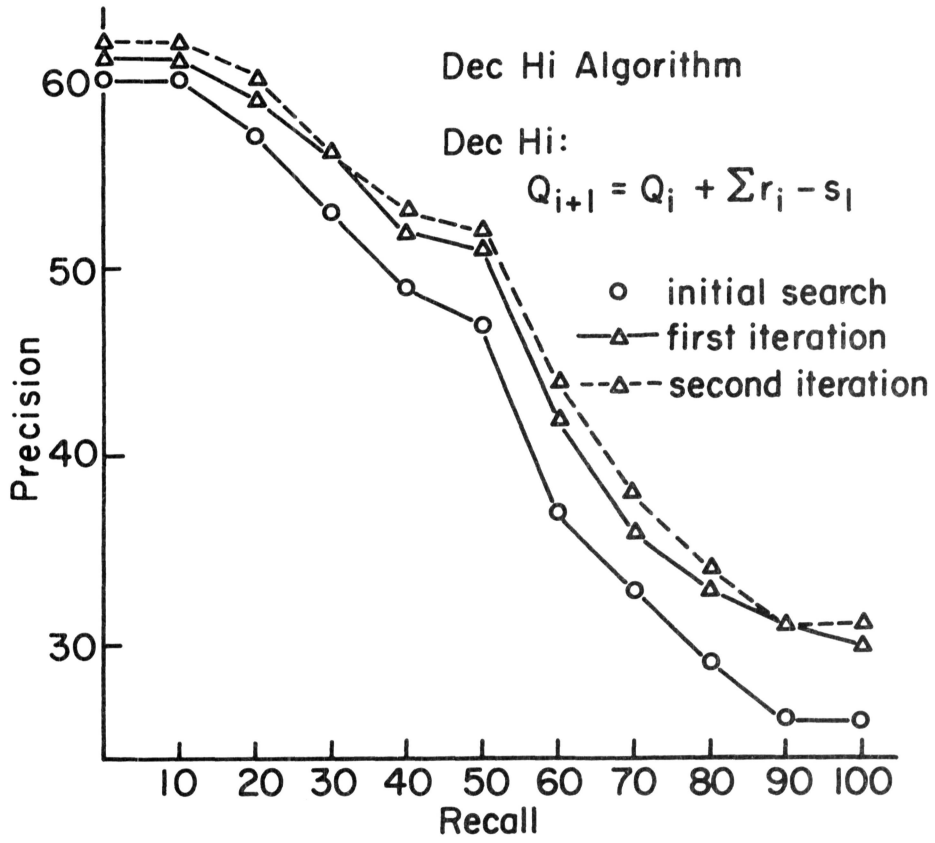
Fig. 6



	First Iter. vs. Initial Search		Second Iter. vs. Initial Search	
	Percent Difference	Percent Significance	Percent Difference	Percent Significance
Rank Recall	4.0	19.1	5.3	5.4
Log Precision	1.8	32.6	2.6	8.2
Normalized Recall	-3.0	60.0	-0.8	23.5
Normalized Precision	-0.1	55.5	1.6	20.8
Recall Level 10%	1.0	51.8	1.6	30.7
20%	2.1	25.8	2.8	12.4
30%	2.3	14.9	3.2	5.3
40%	2.9	10.4	4.1	2.2
50%	3.2	8.8	4.3	3.0
60%	3.1	47.7	4.5	28.7
70%	3.1	43.2	5.4	12.0
80%	4.3	22.1	5.8	8.5
90%	4.4	19.1	5.1	5.7
100%	4.2	21.1	5.3	7.4

Comparison of First and Second Iterations to Initial Search
Differences and Significance Levels - Rocchio Algorithm

Fig. 7



	First Iter. vs. Initial Search		Second Iter. vs. Initial Search	
	Percent Difference	Percent Significance	Percent Difference	Percent Significance
Rank Recall	4.3	13.6	4.8	9.5
Log Precision	2.3	14.5	2.6	9.2
Normalized Recall	-2.3	45.1	-3.3	19.6
Normalized Precision	1.5	25.9	0.6	14.1
Recall Level 10%	1.2	30.9	1.6	21.5
20%	2.4	13.3	2.9	8.4
30%	2.8	8.3	3.3	7.1
40%	3.7	5.5	4.7	1.8
50%	4.0	4.1	4.5	2.3
60%	4.4	14.5	6.5	5.2
70%	2.8	39.1	5.1	16.7
80%	4.3	16.7	5.4	8.5
90%	4.2	17.3	4.6	10.5
100%	4.1	17.9	4.5	10.9

Comparison of First and Second Iterations to Initial Search
Differences and Significance Levels - Dec Hi Algorithm

Fig. 8

recall are the precision differences not significant at the 10% level*. On the second iteration, the performance difference is significant at the 5% level for all points except 70% recall. For the Rocchio strategy, however, only one measure (precision at 50% recall) shows a significant difference between the first iteration and initial search at the 10% level or less. Even on the second iteration, only 8 of the 14 differences are significant at the 10% level or less, two at the 5% level or less. The significance of the corresponding differences for the Dec Hi strategy are similar.

This difference between strategies in the significance of the improvement obtained by feedback leads to a general conclusion: Performance on all measures is less consistent for the nonrelevant document strategies than for the Q_o+ strategy. However, since the average magnitude of this improvement is equal for the three algorithms (from the significance results presented in Fig. 5), it must be true that the Rocchio and Dec Hi strategies are better for some queries and worse for others than is the more consistent Q_o+ strategy.

The greater variance of nonrelevant document strategies is therefore demonstrated not only by the normalized recall, but by all performance measures. The results given here seem to indicate that for some types of queries, nonrelevant documents should be used for feedback, but

---

*For these comparisons, a one-tailed significance level is appropriate, since performance is expected to improve. To obtain one-tailed values, the reported two-tailed values must be divided by two. That is, the probability that the first iteration is no better than the initial search is 5% or less except at 70 and 80% recall.

for others, only relevant documents should be used. If the queries appropriate to each strategy could be distinguished easily before the retrieval operation, performance of the system could be improved by choosing the appropriate strategy for each query. Procedures for distinguishing such subgroups of queries must be investigated.

6. Summary and Recommendations

The isolation of the feedback effect adds to an understanding of relevance feedback in an automatic interactive retrieval system. The present investigation supports the earlier finding [5] that changing the constant formula parameters in the simplest algorithm — using only relevant documents — has little effect on retrieval. This study contradicts the earlier conclusion concerning the optimum amount of feedback. Looking only at the feedback effect, returning ten rather than five documents no longer seems worth the extra user effort. However, feedback effect evaluation tends to minimize differences caused by varying the number of documents used for feedback.

The combination of the results for total performance and feedback effect favors the general use of "variable feedback", in which the user searches the retrieval list only until one relevant document is found. In feedback effect, strategies using nonrelevant documents no longer display an average performance superior to strategies using relevant documents only. Significance tests indicate that "relevant only" strategies are superior for some queries and nonrelevant document strategies for others.

Several areas of investigation are recommended. At least one more iteration of variable feedback, and at least two iterations of the "combination strategy" recommended in the earlier total performance study [5] should be investigated using feedback effect evaluation. Significance tests on the variable feedback results should also be obtained. Rocchio's strategy with normalization should be investigated.

Results of comparing relevant only and nonrelevant document strategies strongly suggest that an investigation be made of subgroups of queries. These types of algorithms seem appropriate to different groups of queries. It would be useful to be able to choose the appropriate strategy for a query before retrieval, by examination of the query. For variable feedback, an investigation of query subgroups should be performed to determine whether or not some identifiable group of users is short-changed by this strategy.

The use of query subgroups, however, raises questions of sample adequacy. Even if the 200 documents and 42 queries of the Cranfield 200 collection are representative of a typical retrieval environment, the statistical dangers of dividing 42 queries into small subgroups must be considered. Investigation of relevance feedback should therefore be continued in a more adequate environment. The Cranfield 1400 document collection, available to the SMART system would provide significantly larger samples of documents and queries.

References

[1] J. J. Rocchio, Relevance Feedback in Information Retrieval, Scientific Report No. ISR-9 to the National Science Foundation, Section III, Harvard Computation Laboratory, August 1965.

[2] J. J. Rocchio, Document Retrieval System Optimization and Evaluation, Harvard University Doctoral Thesis, Scientific Report No. ISR-10 to the National Science Foundation, Harvard Computation Laboratory, March 1966.

[3] J. J. Rocchio, G. Salton, Search Optimization and Iterative Retrieval Techniques, Proceedings of the Fall Joint Computer Conference, Las Vegas, November 1965.

[4] W. Riddle, T. Horwitz, R. Dietz, Relevance Feedback in Information Retrieval Systems, Scientific Report No. ISR-11 to the National Science Foundation, Section VI, Department of Computer Science, Cornell University, June 1966.

[5] E. Ide, User Interaction with an Automated Information Retrieval System, Scientific Report No. ISR-12 to the National Science Foundation, Section VIII, Department of Computer Science, Cornell University, June 1967.

[6] H. Hall, N. Weiderman, The Evaluation Problem in Relevance Feedback Systems, Scientific Report No. ISR-12 to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, June 1967.

[7] G. Salton, The Evaluation of Automatic Retrieval Processes, Selected Test Results using the SMART System, American Documentation, Vol. 16, No. 3, July 1965.

[8] F. Wilcoxon, Some Rapid Approximate Statistical Procedures, American Cyanamid Company, Stamford, Connecticut, 1949.