## VI). Experimental Results

The results of this study are presented in five general sections:

a) A comparison of the improvement in retrieval performance observed for the Cranfield 200 document collection with that obtained for the ADI 82 document collection used by Riddle, Horwitz, and Dietz [11].

b) An investigation of strategies that use only R', the set of relevant documents retrieved, to update the query. The different algorithms are obtained by varying the parameters $\pi$, $\omega$, and $\alpha$ in the query-update formula. The "increasing alpha strategy" of Riddle, Horwitz, and Dietz is included among the methods tested.

c) An investigation of the effect of the number of documents given to the user for feedback on each iteration.

d) An investigation of strategies that use both relevant and non-relevant documents retrieved to update the query.

e) An investigation of the retrieval characteristics of selected subgroups of queries.

### A. Comparison of the Cranfield and ADI Collections

The initial search results, before feedback, for the two collections are essentially the same except at the ends of the recall-precision curves. Below 30% recall, the precision of the ADI initial search is from 2 to 7% better than that of the Cranfield initial search. Above 80% recall the precision in the Cranfield initial search is from 2 to 6% better.

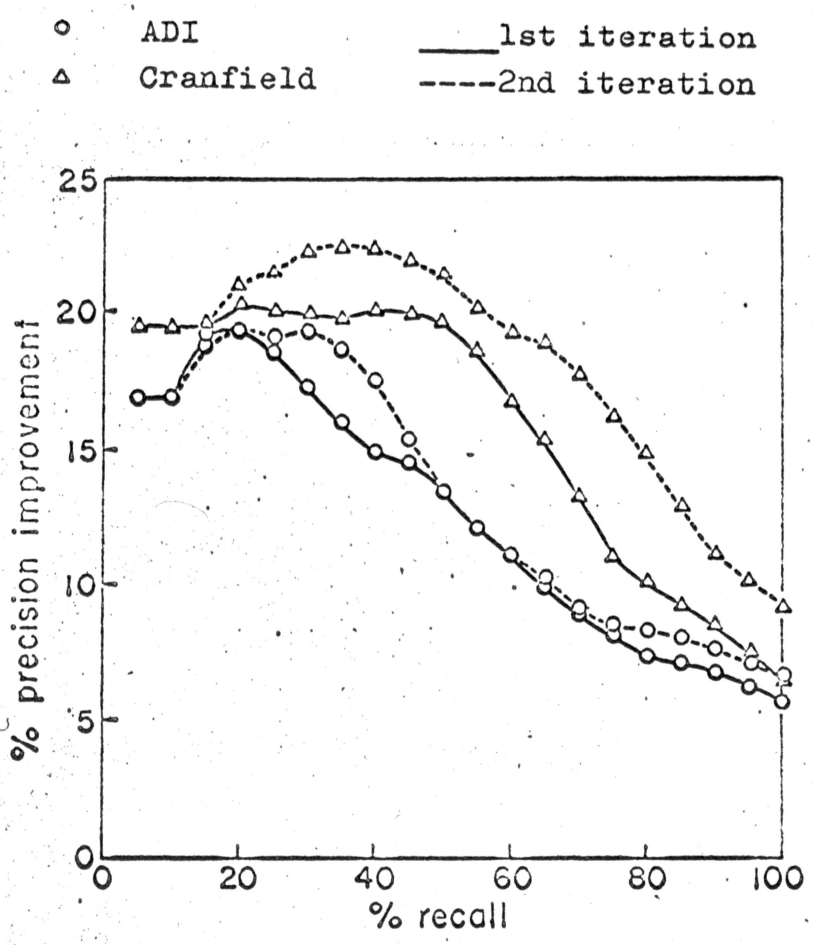
This result is interesting because there is reason to expect that performance in the Cranfield collection would be worse. Cleverdon and Keen point out that in a collection with a higher "generality number", that is, with a higher ratio of relevant documents to collection size, performance is better with respect to precision [18]. The average generality number of the ADI collection is over twice that of the Cranfield collection. The generality number in a collection of practical size would be even lower than that of the Cranfield collection.

Because the initial search results differ, the total performance improvement caused by feedback in the recall-precision curve is used for comparison of the two collections. All thirty-five queries are used to search the ADI collection. The "increasing alpha strategy" of Riddle, Horwitz, and Dietz is the update formula, and five documents are given the user on each iteration.

Figure 3 shows the differences in total performance precision for all recall levels between the initial search and the first and second feedback iterations for each collection. In the Cranfield collection, relevance feedback causes greater improvement than in the ADI collection. Also, the second iteration results in a greater improvement over the first in the 200 document collection. The difference in generality between the collection would be expected to cause less improvement in the larger collection [18].

N=5

"increasing alpha strategy"




Improvement Over Initial Search
ADI and Cranfield Collections

Total Performance Recall-Precision Curves

Figure 3

The greater effect of relevance feedback in the Cranfield collection could be due to any or all of the following factors:

a) The difference in subject and in the language of the subject. It is possible that the terminology of aerodynamics is 'harder', that is, more limited and precise, than the vocabulary of the newer field of computer science. Retrieval of documents from a harder subject area would be expected to be better.

b) The difference in collection scope. The ADI collection covers a wider subject area within computer science than does the Cranfield collection within aerodynamics. A narrower subject area should provide better retrieval.

c) The difference in variability within the collections. The 200 documents were chosen from 1400 documents concerned with aerodynamics. The 82 document collection consisted of short papers presented at a single conference. Since the Cranfield 200 documents vary more in such parameters as vector length and terminology, relevant documents might be easier to distinguish from non-relevant documents.

d) The difference in query construction and relevance judgments mentioned in Section III. It is encouraging to find that in the more realistic Cranfield environment, relevance feedback causes more rather than less improvement in performance.

B. Strategies Using Relevant Documents Only

Two of the experiments of Riddle, Horwitz, and Dietz [11] are repeated for the Cranfield collection. To simulate their

experiments with equation D of Section IV, the parameter $\alpha$ is varied; $\pi$ is kept equal to 1, and $\mu$ and $\omega$ equal to 0. Both the "increasing alpha" and "constant alpha" strategies are employed.

Figure 4 clarifies the effect of the "increasing alpha strategy" and the "constant alpha strategy" for the first, second, and third iterations of a feedback run, evaluating total performance. The $R^0$ column shows the factors which multiply a relevant document retrieved on the initial search, $R^1$ shows the multipliers affecting a relevant document which is not retrieved until the first iteration, and $R^2$ shows the multipliers affecting a document not retrieved until the second iteration. Figure 4 assumes that a document once retrieved is retrieved on all succeeding iterations; in the experimental system this assumption is generally correct.

It is clear that both the constant and increasing alpha strategies give a document retrieved on an earlier iteration more significance in later queries. On the third iteration, the constant alpha strategy assigns to a document retrieved on the initial search three times the significance it gives a second iteration document (the respective multipliers are 3 and 1). The increasing alpha strategy assigns to an initial search document twice the significance of a second iteration document (the respective multipliers are 6 and 3). This effect stems from the use of the previous query $Q_i$ as an element in the equation. To assign

| | iter | $Q_o$ | $R^0$ | $R^1$ | $R^2$ |
|---|---|---|---|---|---|
| TOTAL PERFORMANCE | | | | | |
| | 1 | 1 | 1 | 0 | 0 |
| "increasing alpha strategy" | 2 | 1 | 3 | 2 | 0 |
| | 3 | 1 | 6 | 5 | 3 |
| | 1 | 1 | 1 | 0 | 0 |
| "constant alpha strategy" | 2 | 1 | 2 | 1 | 0 |
| | 3 | 1 | 3 | 2 | 1 |
| | 1 | 1 | 1 | 0 | 0 |
| "$Q_o$ strategy" | 2 | 1 | 1 | 1 | 0 |
| | 3 | 1 | 1 | 1 | 1 |
| FEEDBACK EFFECT | 1 | 1 | 1 | 0 | 0 |
| "feedback increment" | 2 | 1 | 1 | 1 | 0 |
| | 3 | 1 | 1 | 1 | 1 |

Effects of 'Relevant Only' on the Multipliers
of Documents Retrieved on Three Successive Iterations

Figure 4

the same significance to relevant documents whenever they are retrieved, it is necessary to substitute $Q_o$ for $Q_1$ in the formula; that is, to let $\pi = 0$ and $\omega = 1$ in equation B. This is called the "$Q_o$ strategy" in Figure 2.

Riddle, Horwitz, and Dietz [11] report that for the 82 document collection, the "increasing alpha strategy" performs somewhat better than the constant alpha strategy. In the Cranfield collection, the three strategies shown in Figure 2 give essentially the same results when $N = 5$. Using the $Q_o$ strategy with different relative values of $\omega$ and $\alpha$ also does not change performance. Query update parameters (in equation D) for the six experiments performed are shown in Figure 5. Among all six experiments, the differences in normalized precision and recall are less than 0.75% for all iterations.

In total performance, six strategies using only relevant documents differ very little. Three additional 'relevant only' algorithms are compared using feedback effect evaluation. One of these strategies sets $\alpha$ and $\pi$ equal to 1 in formula D. This strategy, called Feedback Increment, is not equivalent to the constant alpha strategy because the feedback effect evaluation provides new documents for feedback on each iteration. Figure 5 shows that the Feedback Increment strategy gives the same weighting effects as does the $Q_o$ strategy. These two strategies are identical on the first iteration, but on subsequent iterations the feedback effect evaluation may retrieve different documents.

$N=5,\ n_a=N,\ n_b=0,\ \mu=0$

| TOTAL PERFORMANCE | $\pi$ | $\omega$ | $\propto$ |
|---|---|---|---|
| | 1 | 0 | 1 |
| increasing alpha | | | 2 |
| | | | 3 |
| constant alpha | 1 | 0 | 1 |
| $Q_o$ strategy | 0 | 1 | 1 |
| $Q_o$ weighting query double | 0 | 2 | 1 |
| $Q_o$ weighting query half | 0 | 1 | 2 |
| $Q_o$ weighting query six times | 0 | 6 | 1 |
| FEEDBACK EFFECT | | | |
| Feedback increment | 1 | 0 | 1 |
| Feedback $Q_o$+ | 1 | 4 | 1 |
| Feedback Rocchio + | $n_r n_s$ | 0 | $n_s$ |

Query Update Parameters for Relevant Only Strategies

Using Only Relevant Documents

Figure 5

Another strategy using feedback effect evaluation, called $Q_0+$, gives added weight to the original query on each iteration by setting $\omega$ equal to 4 ($\alpha = 1$, $\pi = 1$). A third strategy is Rocchio +, the Rocchio strategy without non-relevant documents. In effect, $\alpha$ equals $n_r n_s$ and $\pi$ equals $n_s$, so $\alpha$ and $\pi$ vary with each query.

Differences in feedback effect among these three methods are trivial. For the two overall measures and the recall-precision curves, the largest difference is 1.25 per cent. The document curves are more sensitive in general to performance differences, especially in recall. The largest difference is 3% in recall at a 40 document cut-off. Most differences in all measures favor the $Q_0+$ strategy.

The 200 document collection seems quite insensitive to variations in the parameters $\pi$, $\omega$, and $\alpha$. The considerations mentioned in section VI-A are probably relevant here also. This insensitivity indicates that perhaps the performance for the Cranfield collection is more stable in general than for the ADI collection. Evidence of comparative stability is also reported by Lesk and Salton [19]. The performance differences between automatic use of the word stem thesaurus and a regular subject-area thesaurus (see Section II) are less pronounced in the Cranfield 200 collection than in the ADI collection.

It is evident from the reported experiments that the weight assigned to the original query has little effect on retrieval. This finding tends to support the conclusion of

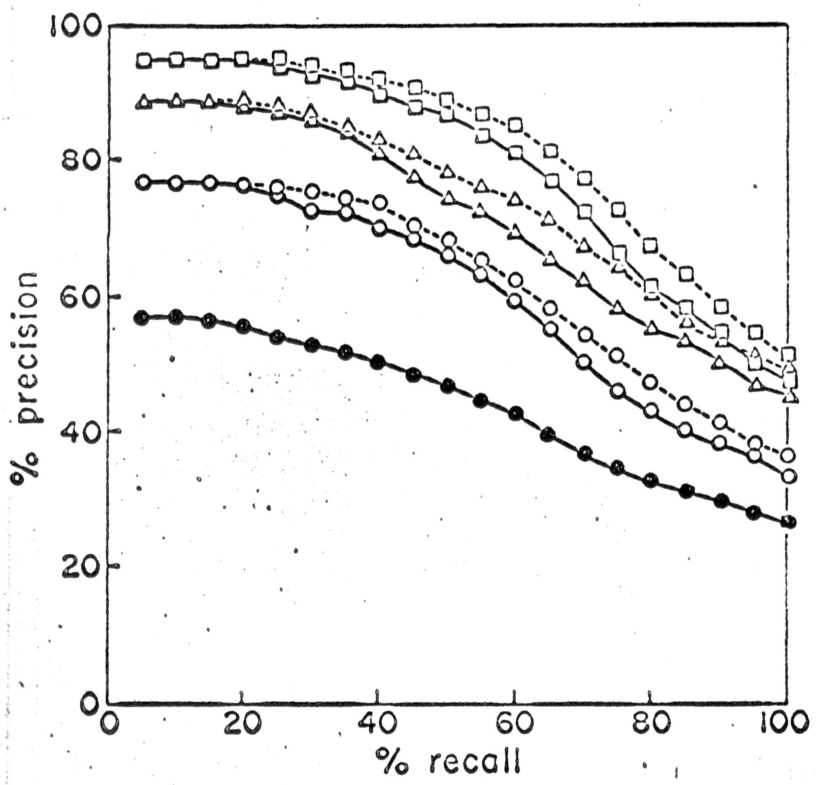
Crawford and Melzer [12] that the original query is not needed after the initial search (Section III). The advantage of their strategy over equation B is probably not caused by the omission of the initial query when relevant documents are found, but by the non-relevant document feedback used when no relevant documents are found (see Section VI-D).

C. Amount of Feedback Output

The number of documents fed to the user is a critical parameter in a relevance feedback system. Of course, performance improves when the user supplies more information. This improvement must be evaluated in terms of the extra effort required of the user.

Figure 6 shows the total performance of the "increasing alpha strategy" when 5, 10, and 15 documents are fed to the user for relevance judgments. The total performance improvement between the N = 5 and N = 10 curves might justify doubling the number of relevance judgments the user must make; that is, a hypothetical "average" user might be willing to double his effort to achieve such an improvement. Tripling the feedback to produce the N = 15 curve might not be justified by total performance, especially at the high recall end of the curve.

Caution is necessary in interpreting the feedback effect evaluation when N is varied, because 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




Number of queries (out of 42)
retrieving no relevant in the first N:

| N = 5 | N = 10 | N = 15 |
|---|---|---|
| 11 | 5 | 2 |

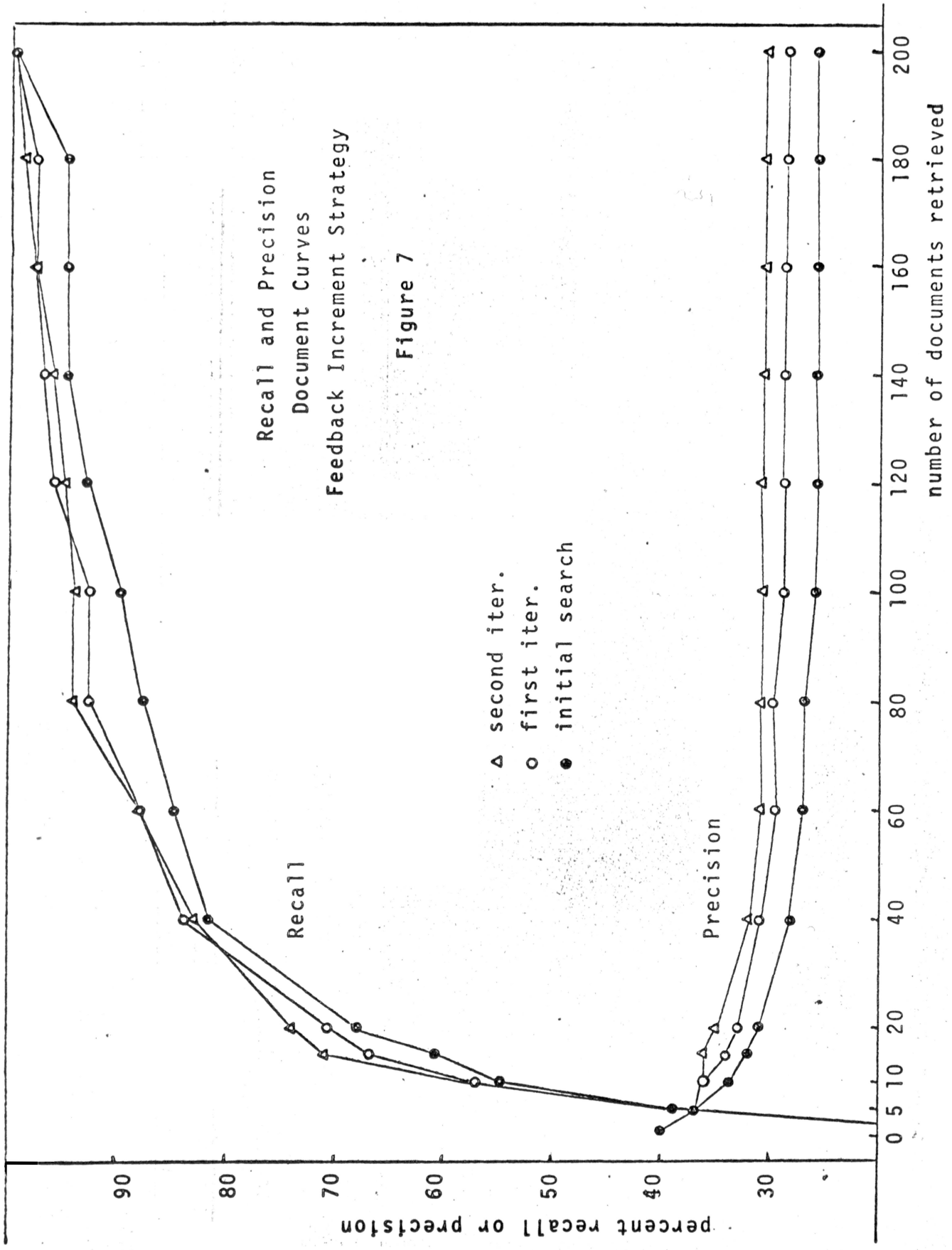
Varying the Number of Feedback Documents
Total Performance Recall-Precision Curves

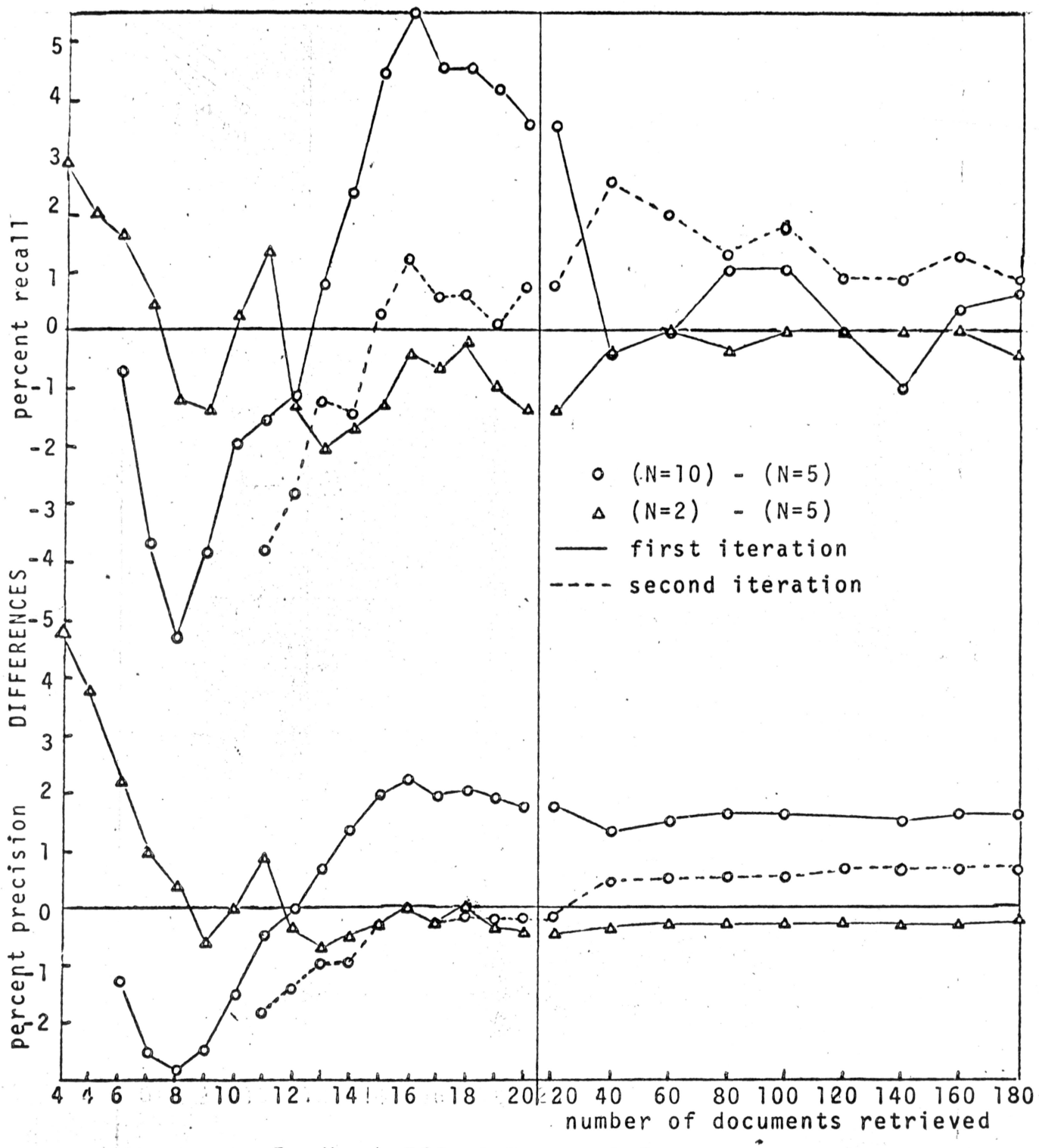
Figure 6

and ranks 1-10 on the second (see Section V-C). 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 feedback effect process.

Recall-precision curves for the feedback effect are not presented in this section, because the minimizing effect described above is averaged into different recall levels for different queries. For example, assume that query a has four relevant documents and query b has two. Each query retrieves the first relevant document with rank 8. When five documents are used for feedback, each retrieves the first relevant document with rank 6. When ten documents are used, of course rank 8 is 'frozen' by the feedback effect evaluation. Consider the effect of these queries on the recall-precision averages for the first iteration. When ten documents are used for feedback, neither query improves these averages. When five are used, query a improves the recall-precision averages from 5% to 25% recall, but query b improves the recall-precision averages from 5% to 50% recall.

Thus it is hard to judge the significance of any difference in results caused by differences in feedback output. Figure 7 shows the document curves for two iterations of feedback with N equal to 5, using the Feedback Increment algorithm. Figure 8 uses these curves as a norm for compar-


Recall and Precision
Document Curves
Feedback Increment Strategy
Figure 7
Recall
Precision
△ second iter.
○ first iter.
● initial search
percent recall or precision
number of documents retrieved
0 5 10 20 40 60 80 100 120 140 160 180 200
30 40 50 60 70 80 90




Feedback Effect Document Curves

Recall and Precision Differences

Varying the Number of Documents Retrieved
(N=5 normal, N=2 and N=10 compared)

Figure 8

ison with the results of less feedback and of more feedback. The document curves of Figure 7 are represented in Figure 8 by the straight line at zero difference. The differences between N=10 and N=5 for two iterations and between N=2 and N=5 for one iteration are graphed. The N=2 differences, indicated by '$\Delta$', are positive at first because ranks 3-5 are frozen for the N=5 curve. This initial advantage fades after 10 documents and the N=2 results are lower than the N=5 norm thereafter. The N=10 curves for both iterations are affected by the feedback effect evaluation. The first iteration gains higher performance than N=5 after 13 documents. The second iteration curves cross the N=5 norm after 15 documents, even with ranks 1-20 frozen. After 40 documents have been retrieved, the differences in both recall and precision for the first iteration are slight; the N=10 advantage on the second iteration is slight but consistent.

After 20% of the collection has been searched, the differences in feedback effect observed in Figure 8 are quite low. However, the marked improvement in early retrieval caused by additional feedback might justify the additional user effort and system output required, especially if further feedback iterations are desired. Moreover, it is important to note that certain users get no benefit from any feedback strategy using only relevant documents. These are the users who find no relevant in the first N documents retrieved. For N=5, 10, and 15,

the number of queries retrieving no relevant on the initial search is given in Figure 6, in the table below the graph. This table probably explains much of the performance difference among the three strategies. Eleven queries in the N=5 case produce the same low performance on the initial search, first iteration, and second iteration. These low results are averaged into all the N=5 curves. The N=10 curve is pulled down by only five such

queries, and the N=15 curve by only two. In the N=5 case, one quarter of the users are not assisted by the chosen feedback strategy, a large proportion for a practical retrieval system. For these unlucky users, feedback of more documents is worth the effort.

A variable feedback strategy is here proposed which might save effort to the average user and give better service for more effort to the user who does not find a relevant document early in the initial search. Each user is fed retrieved documents until he finds one relevant document that he hasn't seen on any previous iteration. The relevant document found is immediately used to produce a new query. The success of this strategy depends on the ability of a single relevant document to improve the retrieval performance.

Figure 9 shows the total performance results of two iterations where the "user" is instructed to search the retrieved documents until he finds one new relevant document or until he has seen 15 documents. The "$\triangle$" curve in Figure 9 shows what happends when the user is instructed to find <u>two</u> new relevant documents. Only one iteration of the latter scheme was run because several queries do not have four or more relevant documents.

The first iteration feeding back one relevant document begins near the N=15 curve of Figure 6 but by 50% recall has dropped near the first iteration N=5 curve, which has been superimposed on Figure 9. The table below the graph shows

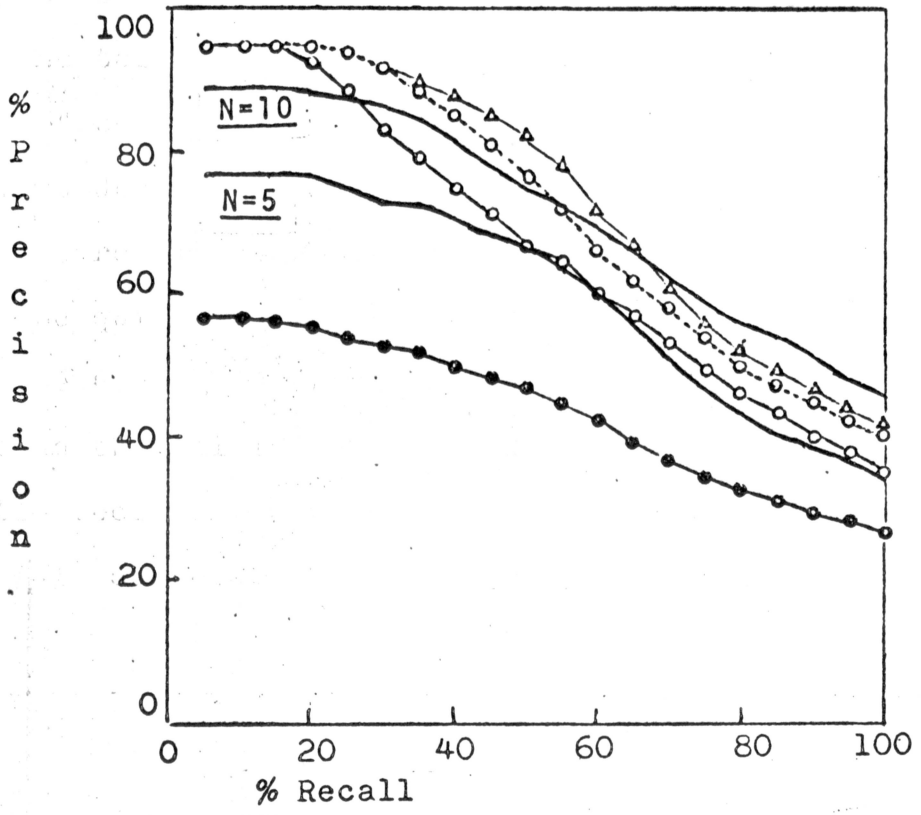
that the "average" user had to scan only four documents for feedback in order to achieve the performance displayed in the first iteration "o" curve. By contrast, each user looked at exactly 5 documents to produce the first iteration N=5 curve. The first iteration strategy of feeding back only one relevant document gives equal or better performance for less average effort.

The second iteration "o" curve requires the average user to search 5.9 more documents, or a total of 9.9 documents. This curve drops below the first iteration N=10 curve (10 documents scanned) at roughly 55% recall (the first iteration N=10 curve from Figure 6 is superimposed on Figure 9). The user desiring high precision and who may be less interested in high recall might be wise to feed back one relevant document for each of two iterations. However, the user needing higher recall should instead look at ten documents retrieved on the initial search. (These statements apply to the "average" user). It is also seen from the table below the graph that for the second iteration "o" curve one quarter of the users cannot find a new relevant document, and thus after the first iteration these users search 15 documents to no avail. Such performance would be quite annoying in practice.

The average user who searches for two relevant documents in the initial output looks at 7 documents. His recall-precision curve ("$\Delta$") drops below the first iteration N=10 curve at 65% recall. Although 9 out of 42 users

o—o initial search
—— 1st iteration
--- 2nd iteration
—— (no character) superimposed curves from figure 6, 1st iter

o feedback 1 new relevant
Δ feedback 2 new relevant




| | o 1 relevant | | Δ 2 relevant | combined strategy |
|---|---|---|---|---|
| | Initial output | 1st iter output | Initial output | Initial output |
| Avg. no. of documents searched | 4.0 | 5.9 | 7.0 | 6.4 |
| No. of users (of 42) not finding n new relevant in the first 15 documents retrieved | 2 | 11 | 9 | 2 |

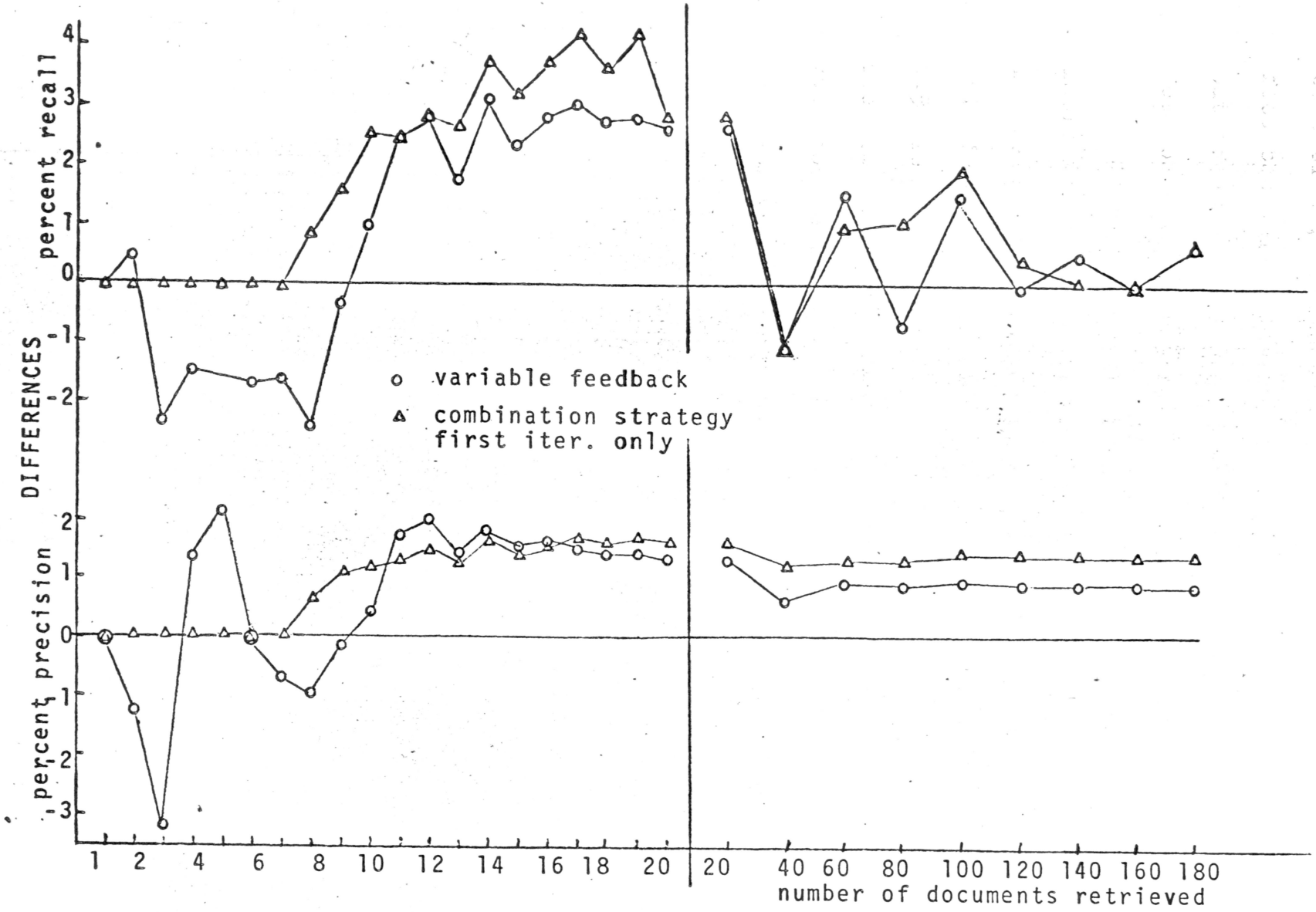
Variable Feedback
Feeding Back Only the First or First Two New Relevant Documents Found Within the First Fifteen Documents Retrieved, Total Performance, $Q_o$ Strategy.

Figure 9

do not find two relevant documents in the first 15, all but two of them find one relevant to feed back to the system.

The user who feeds back two relevant documents on one iteration ("Δ ") achieves better performance than does the user who feeds back one relevant document on each of two iterations (second iteration "o"). This result shows that the second relevant document retrieved on the initial search is more valuable for feedback than the first new relevant retrieved on the first iteration by the total performance retrieval method. Feeding back one relevant document on the first iteration evidently pushes down some relevant documents that are valuable for retrieval. This finding provides a strong argument against the proposed variable feedback strategy, at least where high recall is desired. Perhaps some sort of combination strategy might be optimal; for instance, the user could be instructed to feed back all relevant documents in the first five retrieved, but if none are found in the first five to keep looking and feed back the first relevant document found.

One iteration of feedback effect performance of variable feedback and of the combined strategy described above is presented in Figure 10. The differences between variable feedback (feeding back one relevant) and constant N=5 feedback are graphed ( o ). After nine documents have been retrieved the feedback advantage using the first relevant document for feedback is evident. The combination strategy (graphed ▲) that retrieves at least five documents




Variable Feedback and Combination Strategy Compared to N=5 Norm
Recall and Precision Differences, Feedback Effect Document Curves

Figure 10

shows a greater improvement over N+5 than does variable feedback. This result shows that this variable feedback algorithm is advantageous only for those queries that retrieve no relevant among the first 5 documents. For those that do retrieve relevant within the first five documents, the constant N=5 strategy gives better feedback effect results. Figure 9 shows that the combination strategy requires the average user to look at 6.4 documents, as compared to the 4.0 documents retrieved by variable feedback.

Total performance and feedback effect results support three conclusions about feedback strategies that use only relevant documents:

First, retrieving more documents does improve both types of performance (except where the rank-freezing of feedback effect evaluation prevents any improvement). Further, notable improvement can be obtained by searching further if no relevant documents are retrieved in the first N.

Second, the total performance for 5, 10, and 15 documents indicates that when N is constant for all queries, the average increment of improvement obtained tends to become smaller as more documents are used for feedback.

Third, however, a comparison of the variable feedback and combination strategies show that the first relevant document retrieved generally does not contain enough information for retrieval, and that documents retrieved soon after the first (N=5) can add useful information. In their study cited in Section III, Crawford and Melzer used only one 'very relevant' document for retrieval. The finding of this study would indicate that if several documents are

almost equally relevant, all of them should be used.

These three conclusions lead to a recommendation that N be set to some value that most users would consider reasonable, but that for some queries N should be raised until at least one relevant document is retrieved. This recommendation endorses the "combination strategy" for a retrieval system using only relevant documents.

D) Strategies Using Non-Relevant Documents

Rocchio's update formula (equation A) considers the information contained in the set of non-relevant documents retrieved (S) to be as important as that contained in the set of relevant documents retrieved (R). If this is the case, the strategies so far examined disregard half the available feedback information. Further, information from non-relevant documents retrieved on the initial search might help those users who retrieve no relevant documents on the initial search (see Section III and reference 6). Figure 6 shows that there are eleven such users out of 42 when N equals 5.

However, problems arise in using the non-relevant documents in the SMART experimental system. There is no provision for negative weights in the query vector. Also, queries and documents cannot be normalized to the same length for query updating. There is some danger, therefore, that the query will be reduced to nearly the zero vector when the documents of S are subtracted from it. Riddle, Horwitz, and Dietz [11] try to avoid this danger in their "negative heuristic strategy" by feeding back only the first two non-relevant documents retrieved.

The Rocchio strategy adjusts the multipliers for each query so as to weight the original query, the sum of the relevant, and the sum of the non-relevant equally, and uses all retrieved documents.

This study compares the Rocchio and 'negative heuristic' strategies using total performance and feedback effect measures. All comparisons are made with N equal to 5. Figure 11 compares the $Q_o$ strategy (see Section VI-B) with a strategy called 'Dec Hi', that decrements each query by subtracting from it the first retrieved non-relevant document. In the query update formula (equation D), the parameter values and effective update formulas for these strategies are:

$Q_o$: $\pi = 0,\ \omega = 1,\ \alpha = 1,\ \mu = 0,\ n_a = N,\ n_b = 0$

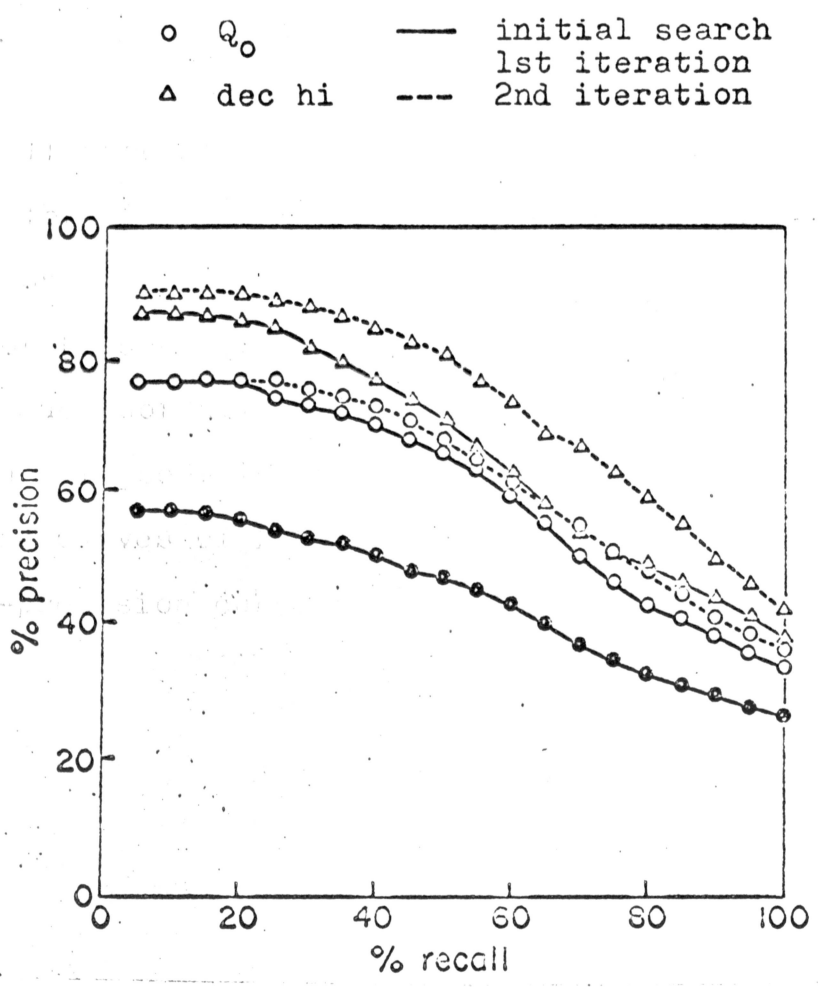
$$Q_{i+1} = Q_i + \sum_{1}^{N} r_i$$

Dec Hi: $\pi = 0,\ \omega = 1,\ \alpha = 1,\ \mu = -1,\ n_a = N,\ n_b = 1$

$$Q_{i+1} = Q_i + \sum_{1}^{N} r_i - s_1$$

Figure 11 shows that the average results are consistently better for the "dec hi" strategy, especially on the second iteration.

For this experiment, the implications of the total performance normalized precision and recall results given in Figure 12 seem inconsistent with those of the recall-precision curves of Figure 11. On the first iteration, the recall-precision curve for the dec hi strategy is above




Decrementing the Highest Non-relevant Document
Total Performance Recall-Precision Curves

Figure 11

| | | Init | 1 | 2 |
|---|---|---|---|---|
| Normalized Recall | $Q_o$ | 88% | 90 | 90 |
| | Dec High | 88 | 87 | 90 |
| | Dec 2 Hi | 88 | 85 | 89 |
| Normalized Precision | $Q_o$ | 68 | 75 | 76 |
| | Dec High | 68 | 76 | 81 |
| | Dec 2 Hi | 68 | 75 | 80 |

Normalized Results for Non-relevant Document Strategies
Total Performance

Figure 12

that for $Q_0$ at all recall levels. However, the normalized recall for the first iteration is lower for dec hi, although precision is one percent higher. (On the second iteration, the normalized recalls are the same, and the normalized precision for dec hi is five percent higher). This apparent paradox can be understood by considering the normalized recall measure.

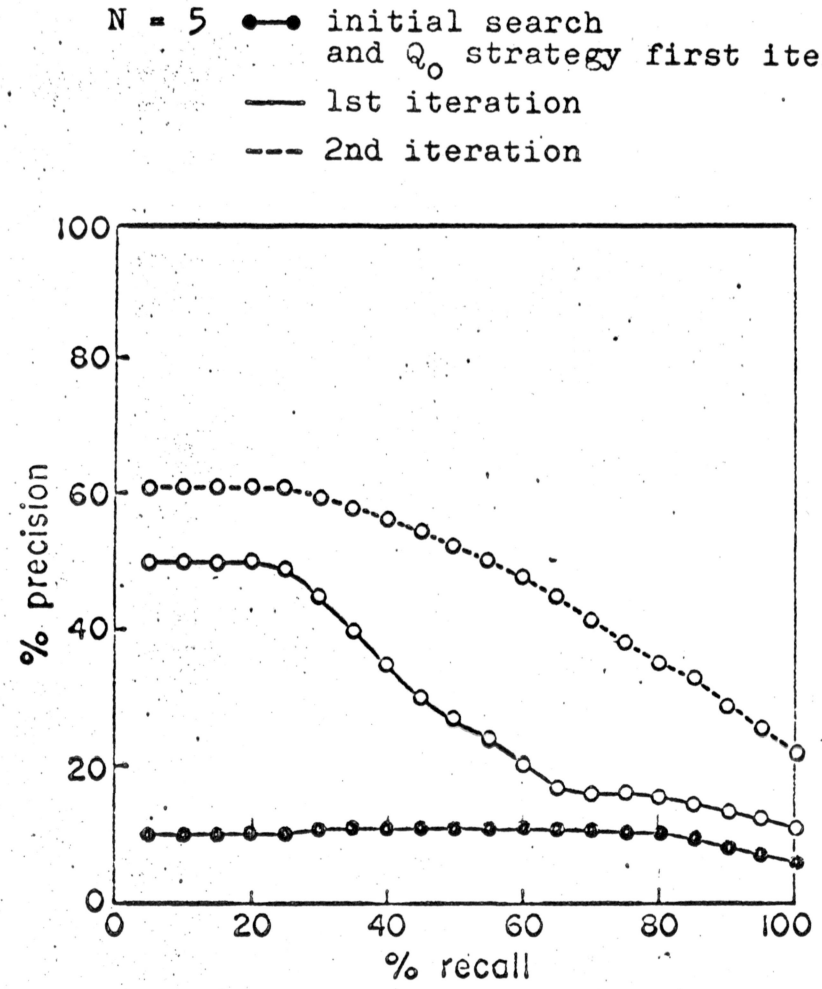
Each document retrieved is assigned a "rank" in order of retrieval (rank 1 is the document retrieved first). The normalized recall measure is based on the sum of the ranks of all relevant documents in the search. A change in rank affects this measure equally regardless of the magnitude of the rank. That is, a change from rank 195 to rank 191 is equivalent to a change from 5 to 1 in its effect on normalized recall. The same is not true for normalized precision. It seems evident that while the dec hi strategy increases the rank (1 is considered highest) of some of the relevant documents, it decreases the ranks of others that are, on the average, of lower rank already. This explains the phenomenon of higher precision at all levels of recall but lower overall normalized recall.

Figure 13 shows how much the dec hi strategy helps the 11 users who receive no relevant documents in the first 5 on the initial search. For the "inc only" strategy, the initial search and all subsequent iterations are the same for these 11 users; the precision being about 10 percent. Feeding back one non-relevant document fetches at least one

relevant document on the first iteration for 7 of these 11 users. For some of these queries, some low ranking relevant documents are pushed still lower at first. The relevant documents which are raised to the first 5, however, provide a second iteration query which often raises these same low documents again. The first iteration curve thus shows the most improvement at low recall, while the second iteration shows great improvement all along the recall-precision curve.

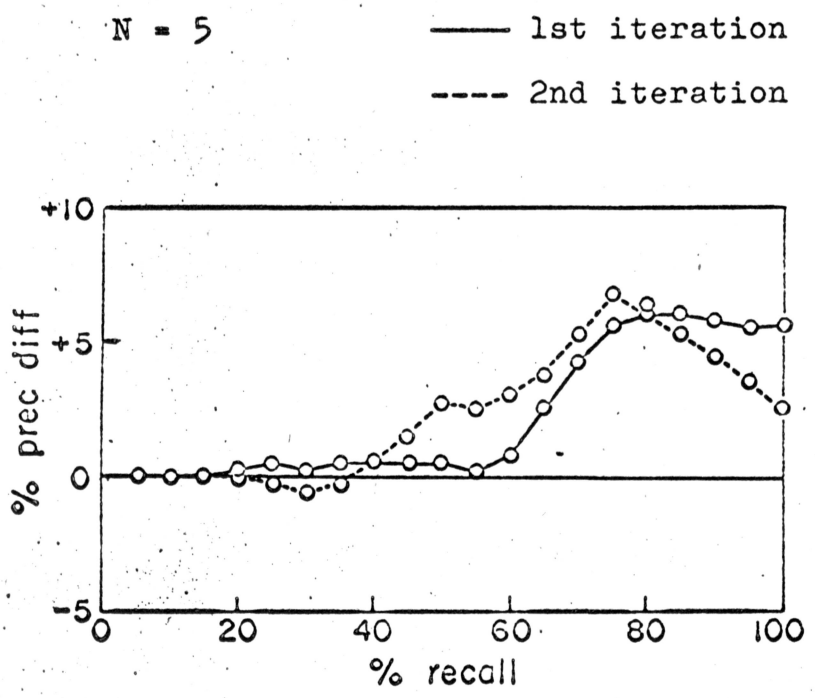
Since the improvement in total performance for the 11 "bad" queries is so striking, it is natural to wonder whether this strategy is helping or hurting the other 31 users. Figure 14 shows a different curve for the dec hi and $Q_0$ strategies run only on the 31 queries that retrieve at least one relevant in the first 5 documents. A point above the zero line indicates that dec hi is better than $Q_0$ at that recall. Both iterations are better for dec hi, especially at the high recall end of the curve, where they differ by as much as six percent. Since the dec hi strategy improves the results even for the "good" queries, a heuristic strategy that selects only some of the queries (as does the "negative heuristic strategy" of Riddle, Horwitz, and Dietz) for the dec hi algorithm appears unnecessary in this environment.

Figure 15 represents a total performance difference curve comparing the "dec hi" strategy with the alternative of decrementing the query by subtracting the <u>two</u> highest non-relevant documents retrieved on each iteration (called




Decrementing the Highest Non-Relevant Document
on Eleven Bad Queries
Total Performance Recall-Precision Curves

Figure 13



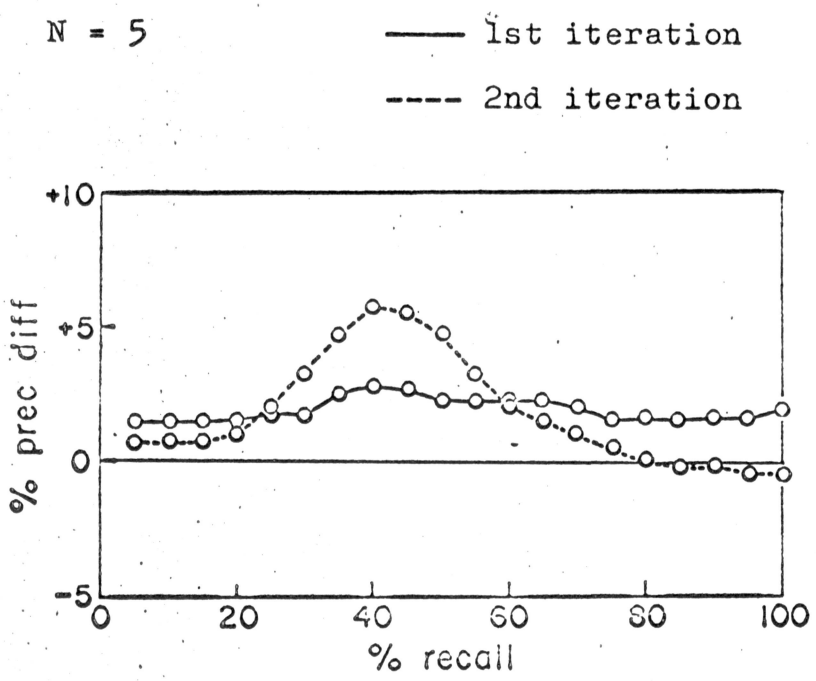

(Dec Hi) - ($Q_o$)

For the 31 Good Queries, a Comparison Between Decrementing the Highest Non-Relevant and Incrementing the Relevant Only Total Performance Difference

Figure 14

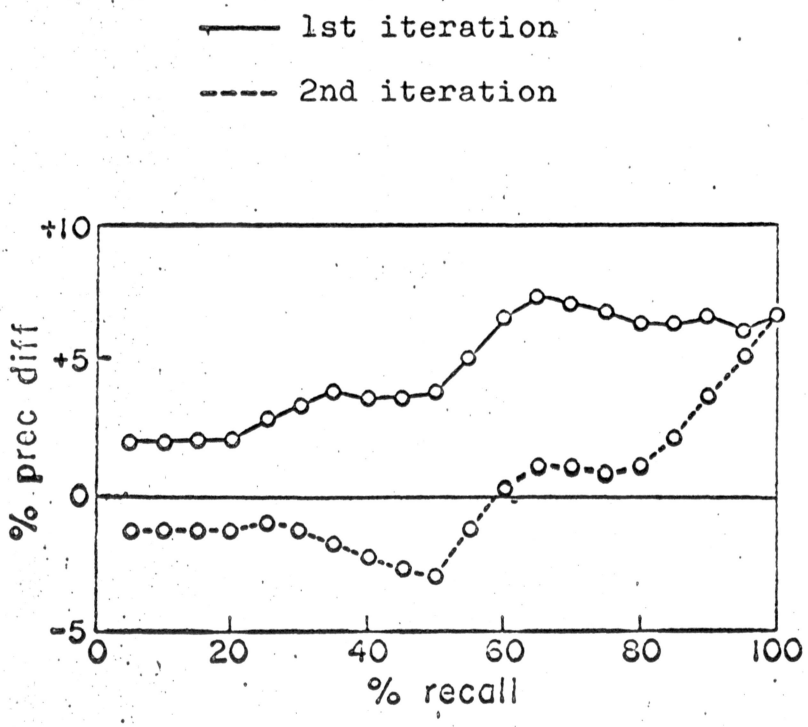
"dec 2 hi"). It shows that decrementing only one non-relevant gives generally better results; the largest difference being a five percent hump at 40% recall in the second iteration. In Figure 12 the normalized measures show the same relationship; the dec 2 hi strategy is one or two percent lower on each iteration than is dec hi. This result may be due to the danger mentioned earlier, that the non-relevant documents may be subtracting out most of the query. (Only one query completely disappears using this strategy, and it is erased also by the dec hi and Rocchio's algorithms.) It might be possible to overcome this "disappearing query" phenomenon by juggling the parameters $\pi$, $\omega$, $\alpha$, and $\mu$, without introducing the complications of Rocchio's normalizing method.

It was mentioned in Section VI-C that much of the improvement between the N=10 and the N=5 curves of Figure 6 might be caused by the improvement on the six queries that fetch a relevant document within the first 10 but not the first 5 on the initial search. Seven users of the unlucky 11 are helped by the dec hi strategy; that is, the dec hi strategy provides useful feedback for one more user than does the relevant document strategy with N=10. It is pertinent to ask if the dec hi algorithm has, in fact, attained the total performance of the N=10 curve. Figure 16 shows a difference curve between the N=10 curve of Figure 6, and the dec hi curve of Figure 11. The N=10 curve is higher for the first iteration, over five percent higher at the high recall end of the curve. This is understandable in view of the lowering of low-ranking relevant documents on the first




(Dec Hi) - (Dec 2 Hi)
A Comparison of Decrementing
the Highest or the Two Highest Non-Relevant Documents
on Each Iteration
Total Performance Difference

Figure 15




(Feedback 10) - (Dec Hi)

A Comparison Between Feeding Back 10 Documents
but Increasing Relevant Only,
and Feeding Back 5 Documents
but Decrementing the Highest Non-Relevant
Total Performance Difference

Figure 16

iteration, discussed earlier in this section. For the second iteration, the dec hi curve is slightly better at the low recall end and only slightly worse at recalls between sixty and ninety percent. Considering that the dec hi curve only requires half as much user effort (5 documents scanned instead of 10), the total performance results strongly favor this non-relevant document retrieval strategy.

The feedback effect results are not as encouraging. Using the two overall normalized measures and the recall-precision curves, three strategies are compared and their differences tested with the two significance tests described in Section V-A, the T-Test and the WSR Test. The three strategies that are compared in feedback effect performance are:

Feedback Dec Hi: The Dec Hi strategy with the feedback effect retrieval method, using the first retrieved non-relevant document.

Rocchio: Rocchio's recommended strategy (without normalized vectors) using all retrieved documents.

$Q_0+$ : The strategy described in Section VII-B, that gives added weight to the original query and uses only relevant documents.

The differences in feedback effect recall-precision curves among these strategies are not significant. The largest differences found were 1.3% in the first iteration,

significant at the 30% level, and 2.0% in the second iteration, significant at the 11% level. The largest difference between $Q_0+$ and any other strategy was 1.2%, significant at the 64% level. These significance figures were obtained using the less conservative T-test. Figure 17 shows the differences among the three strategies in normalized recall and normalized precision. The feedback effect results agree with the total performance results in showing a drop in normalized precision for the two non-relevant document strategies on the first iteration. The five percent difference between $Q_0+$ and feedback Dec Hi is significant at the 6% level, and the six percent difference between $Q_0+$ 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 $Q_0+$ and feedback Dec Hi is 46%, and that comparing $Q_0+$ and Rocchio is 95%.

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 $Q_0+$ and differences favoring the non-relevant document strategy are mixed in rank, producing insignificant results on the WSR test. Yet, some of the results favoring $Q_0+$ (not <u>all</u>, because the ranks are mixed) must be very large in magnitude, to give significant indications on the T-test. Thus, for

| | | Normalized Recall | | | Normalized Precision | |
|---|---|---|---|---|---|---|
| | | % Difference | T Test | WSR Test | % Difference | T Test |
| $Q_0$+ strategy minus Rocchio strategy | Iter 1 | 6.1 | 3.4 | 24.1 | 2.9 | 15.4 |
| | Iter 2 | 4.4 | 17.7 | 48.9 | 2.0 | 43.0 |
| $Q_0$+ strategy minus Dec Hi strategy | Iter 1 | 5.4 | 5.7 | 98.5 | 2.1 | 31.4 |
| | Iter 2 | 7.0 | 9.4 | 45.6 | 3.0 | 29.6 |

Statistical Comparison of Feedback Effect

Relevant and Non-Relevant Document Strategies

Normalized Recall and Precision

Figure 17

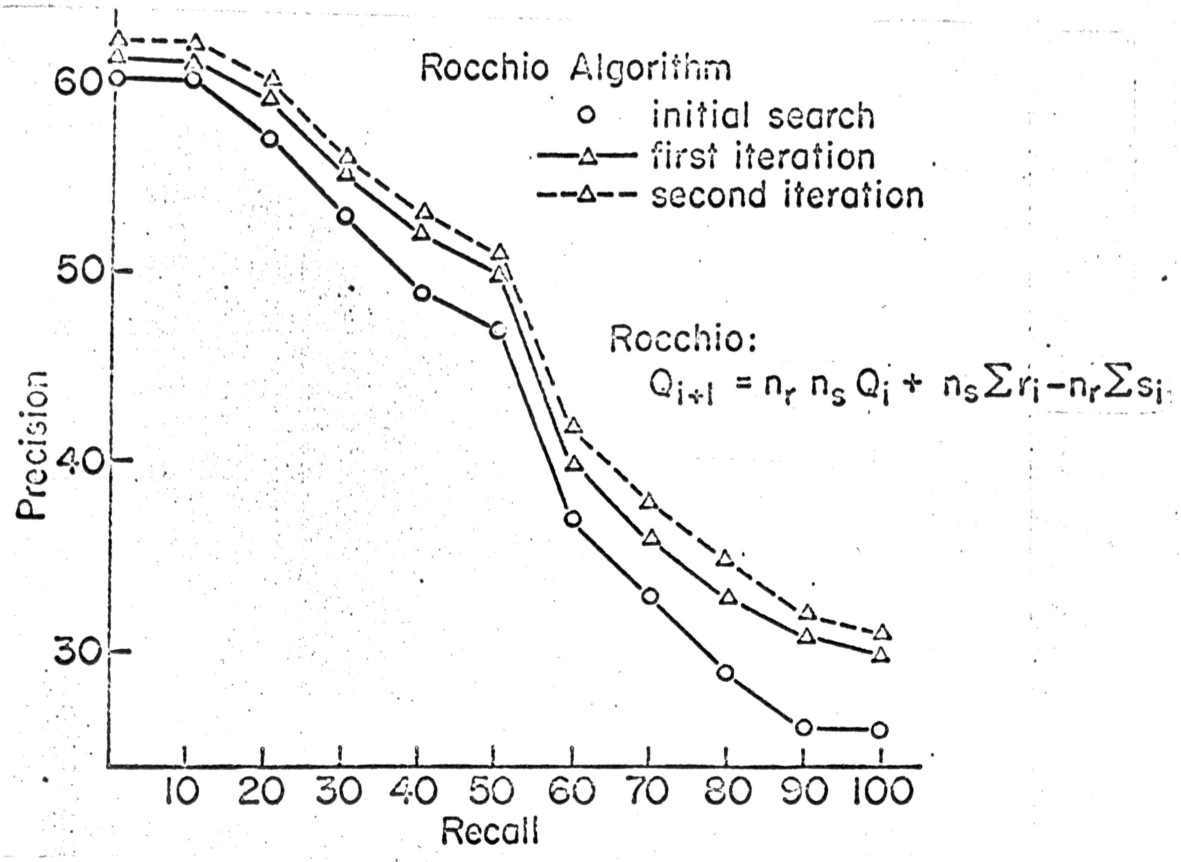
some queries, the Rocchio and Feedback Dec Hi algorithms must be much less effective than $Q_0+$ as measured by normalized recall, while remaining effective as measured by normalized precision.

The normalized recall obtained by feedback effect evaluation shows the same behavior as the total performance normalized recall on the first iteration. Both evaluation methods lead to the conclusion that the use of non-relevant documents for feedback apparently raises the ranks of fairly high-ranking relevant documents, and at the same time lowers the ranks of some low-ranking relevant documents on the first iteration.

The significance levels obtained by comparing the first and second iteration results to the initial search result within the strategies are very informative. Figures 18 and 19 show the performance of algorithms $Q_0+$ and Rocchio 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 seem quite similar. In fact, the differences in average performance are not significant. Yet, the significance levels displayed in Figure 18 differ greatly from those displayed in Figure 19.

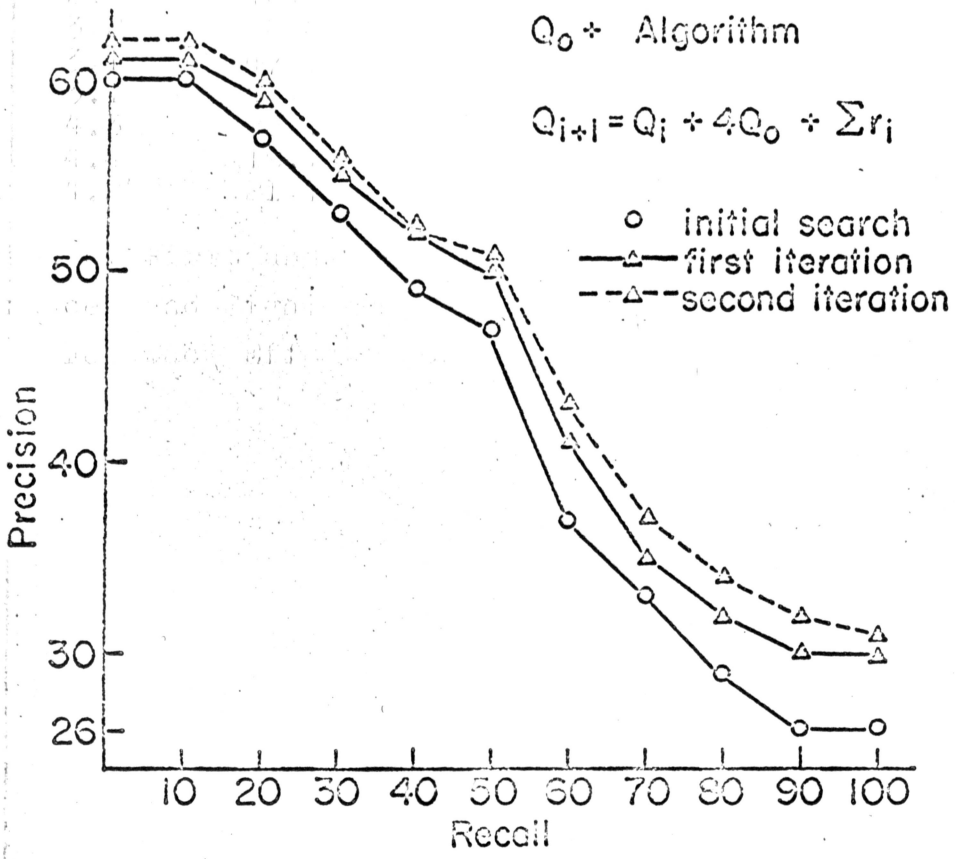
For the $Q_0+$ strategy, the differences between the initial search and each feedback iteration are significant. On the first iteration, the two overall measures and the precision differences from 20% through 50% recall are significant at the 5% level or less, and only at 70 and




| | First Iter Minus Initial Search | | Second Iter Minus Initial Search | | Second Iter Minus First Iter | |
|---|---|---|---|---|---|---|
| | % Difference | % Significance | % Difference | % Significance | % Difference | % Significance |
| Normalized Recall | -3.0 | 60.0 | -0.8 | 23.5 | 2.2 | 17.4 |
| Normalized Precision | -0.1 | 55.5 | 1.6 | 20.8 | 1.8 | 8.0 |
| Recall Level 10% | 1.0 | 51.8 | 1.6 | 30.7 | 0.6 | 3.6 |
| 20% | 2.1 | 25.8 | 2.8 | 12.4 | 0.7 | 3.1 |
| 30% | 2.3 | 14.9 | 3.2 | 5.3 | 0.9 | 1.9 |
| 40% | 2.9 | 10.4 | 4.1 | 2.2 | 1.2 | 5.1 |
| 50% | 3.2 | 8.8 | 4.3 | 3.0 | 1.1 | 10.9 |
| 60% | 3.1 | 47.7 | 4.5 | 28.7 | 1.4 | 5.7 |
| 70% | 3.1 | 43.2 | 5.4 | 12.0 | 2.3 | 3.1 |
| 80% | 4.3 | 22.1 | 5.8 | 8.5 | 1.5 | 11.8 |
| 90% | 4.4 | 19.1 | 5.1 | 5.7 | 1.3 | 8.5 |
| 100% | 4.2 | 21.1 | 5.3 | 7.4 | 1.1 | 17.1 |

Comparison of First and Second Iterations to Initial Search Differences and Significance Levels - Rocchio Algorithm Feedback Effect Recall-Precision Curves

Figure 18




| | First Iter Minus Initial Search | | Second Iter Minus Initial Search | | Second Iter Minus First Iter | |
|---|---|---|---|---|---|---|
| | % Difference | % Significance | % Difference | % Significance | % Difference | % Significance |
| Normalized Recall | 3.1 | 0.6 | 3.7 | 0.1 | 0.6 | 24.9 |
| Normalized Precision | 2.7 | 1.8 | 3.6 | 0.5 | 0.9 | 16.5 |
| Recall Level 10% | 1.0 | 6.8 | 1.8 | 1.3 | 0.8 | 11.2 |
| 20% | 1.8 | 4.3 | 2.5 | 0.8 | 0.7 | 17.7 |
| 30% | 2.2 | 1.5 | 3.0 | 0.9 | 0.8 | 25.0 |
| 40% | 3.0 | 0.5 | 3.7 | 0.4 | 0.7 | 26.4 |
| 50% | 3.1 | 0.8 | 3.7 | 1.1 | 0.6 | 40.7 |
| 60% | 3.5 | 5.7 | 5.7 | 2.8 | 2.1 | 7.0 |
| 70% | 2.4 | 29.2 | 4.4 | 19.5 | 2.0 | 7.6 |
| 80% | 3.2 | 18.3 | 4.9 | 4.5 | 1.7 | 4.1 |
| 90% | 3.6 | 6.1 | 5.6 | 2.4 | 2.0 | 3.5 |
| 100% | 3.6 | 6.1 | 5.2 | 2.6 | 1.7 | 9.8 |

**Comparison of First and Second Iterations to Initial Search Differences and Significance Levels - $Q_0$+ Algorithm Feedback Effect Recall-Precision Curves**

**Figure 19**

80% 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 six of the twelve differences are significant at the 10% level or less, two at the 5% level or less. Significance results for the feedback Dec Hi strategy are similar.

When comparing first and second iterations, the $Q_0$+ results are no longer more significant than the Rocchio results. In fact, the Rocchio results are significant (10% level or less*) for eight of the twelve measures; the $Q_0$+ results for only five. The significant improvement between first and second iterations occurs at the high recall end of the $Q_0$+ curve, while the improvement for the Rocchio strategy is more evenly distributed.

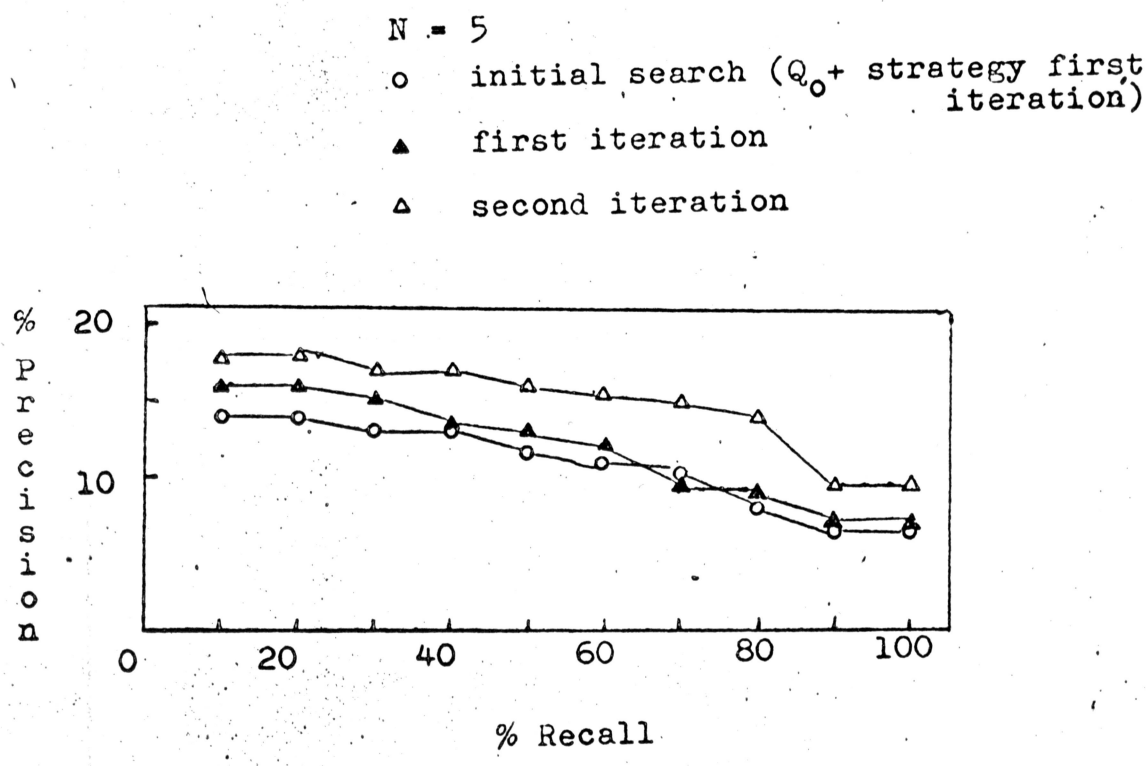
This difference between strategies in the significance of the improvement over the initial search leads to a general conclusion: Performance on all measures is less consistent for the non-relevant document strategies than for the $Q_0$+ strategy. However, since the average magnitude of

*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.

this improvement is equal for the three algorithms (from the significance results presented in Figure 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 consistant $Q_0+$ strategy.

The total performance results of Figure 13 indicate that the queries that retrieve no relevant documents on the initial search are helped by the non-relevant feedback strategies. Figure 20 supports this conclusion with evidence that even using feedback effect evaluation, the Rocchio strategy provides better performance on these eleven queries. Figure 14 adds the information that on the average, the total performance on the remaining 31 queries is not hurt by negative feedback. The preceding paragraph leads to the conclusion that in feedback effect the Rocchio strategy gives worse performance on some of these queries. This conflict between total performance and feedback effect results requires further investigation of subgroups of queries.

The document curve differences presented in Figure 21 provide new information about the performance of the negative feedback strategies. The $Q_0+$ strategy is taken as the norm, and the Rocchio and Feedback Dec Hi differences from $Q_0+$ are graphed. For the first fifteen or so documents retrieved, both Rocchio and Feedback Dec Hi are superior in feedback effect performance to $Q_0+$. After 40 documents have been retrieved, both are <u>much</u> worse than $Q_0+$

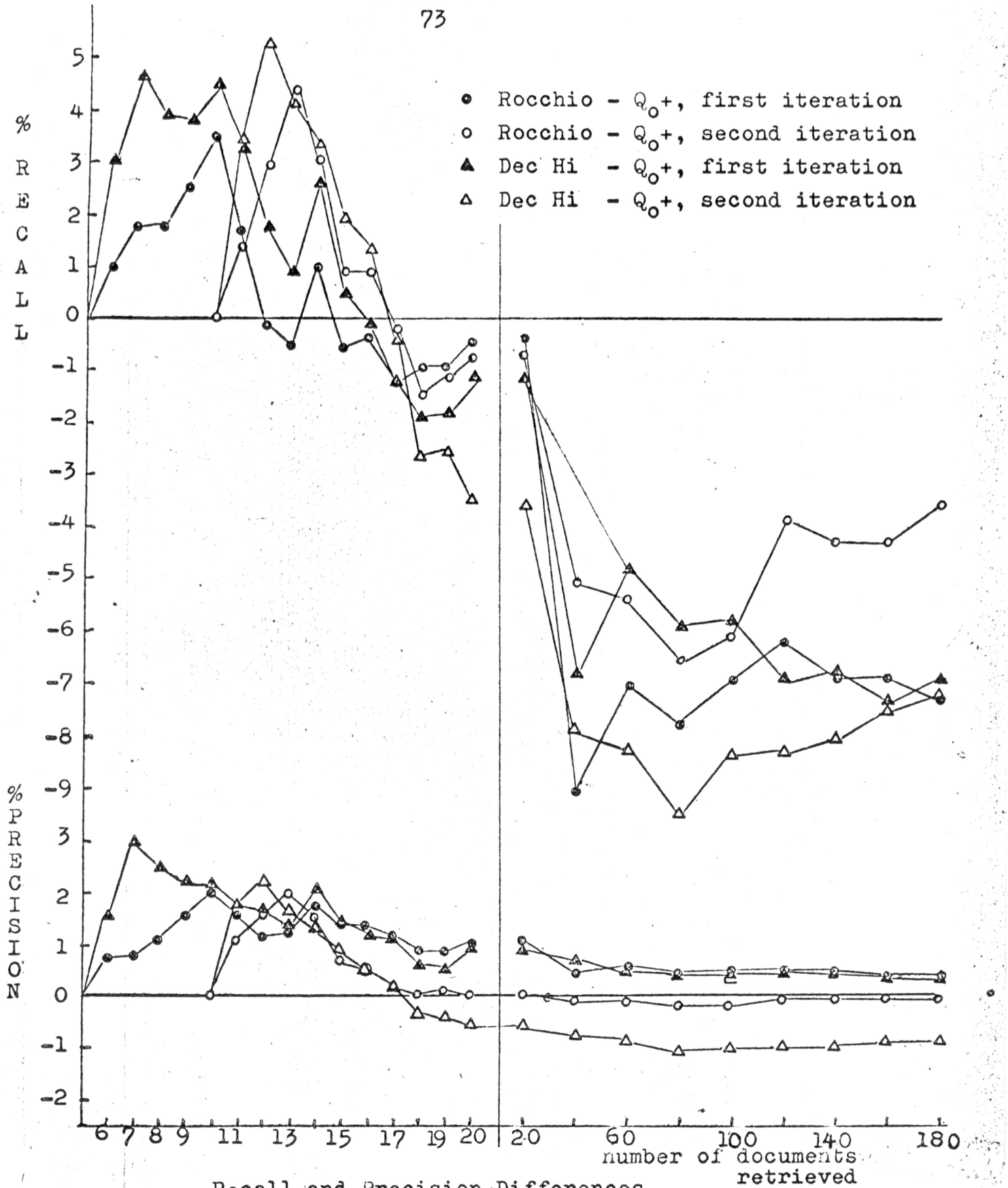



Rocchio Strategy

On Eleven Bad Queries

Feedback Effect Recall-Precision Curves

Figure 20




Recall and Precision Differences
Comparing Rocchio and Dec Hi Strategies
to $Q_0$+ norm
Feedback Effect Document Curves

Figure 21

in recall, though about the same in precision. The recall-precision curves of Figures 18 and 19 average out these two extremes, and lose the significant information.

Figure 20 strongly supports the conclusion implied by normalized recall, that on the first iteration non-relevant document strategies tend to raise some relevant documents, but to lower others that are already low in rank. The average advantage of the non-relevant document strategies appears early in the retrieval process. After 20% of the collection has been scanned, the $Q_0+$ strategy is clearly superior in recall. The rank-freezing of the feedback effect evaluation affects the ranks of the earliest documents retrieved, so the recall-precision curves of the non-relevant document strategies appear superior in total performance but only equal in feedback effect. Normalized recall expresses the later large drop in recall which overwhelms the earlier advantage of negative feedback. Thus, the document curves support and clarify the tentative deductions made from the less detailed measures presented earlier.

Total performance comparisons encourage the use of algorithms that employ negative feedback of non-relevant documents. However, the feedback effect results indicate that the performance of negative feedback algorithms is highly variable. These findings encourage a search for a means of predicting the appropriate strategy for a given query. For this reason the characteristics of selected subgroups of queries are explored in the following section.

E. Characteristics of Query Subgroups

To investigate the performance of positive and negative feedback in more detail, the available queries are split several ways into pairs of query subgroups. Each subgroup pair represents a contrast based on one or two characteristics. For example, all queries with four or fewer relevant documents might form one subgroup of a pair, and all queries with five or more relevant documents might constitute the contrasting subgroup of that pair. The six queries that retrieve all relevant documents with rank 5 or less on the initial search are omitted from analysis because relevance feedback cannot improve the feedback effect performance on these queries. Figure 22 lists the characteristics used for selection and describes some subgroups for which comparisons are reported.

Each subgroup is statistically compared to the contrasting subgroup using the Wilcoxon Rank Sum test (see Section V). All findings of less than 10% significance are reported in this section. If the word 'null' is used in the WRS column of a figure, the significance level of the indicated comparison is greater than ten percent, and the 'null hypothesis' of no difference between subgroups is supported. Although significance level from five to ten percent are not normally considered meaningful, they are reported here for two reasons. First, the WRS test is conservative when too many ties in rank occur, and the data contains ties. Second, the numbers of queries in each

| Selection Characteristic | Group Name (Size) | Group Description |
|---|---|---|
| Initial Search Retrieval | Eleven Bad (11) | No relevant documents are retrieved with rank 5 or less on the initial search. |
| | Twenty-Five (25) | Some but not all relevant documents are retrieved with rank 5 or less on the initial search. |
| Relevance Feedback Performance | Good Performance (16) | At least one feedback strategy retrieves all relevant documents with rank 15 or less within three iterations. |
| | Bad Performance (20) | No feedback strategy retrieves all relevant documents with rank 15 or less within three iterations. |
| Correlation of Modified Query with Original Query | | Six subgroups are chosen. The number of queries in each is given in the following table: |
| Relevance Feedback Strategy | $Q_o$+ (13) | The $Q_o$+ strategy retrievesmore documents with rank 15 or less in three iterations. |
| | Rocchio (15) | The Rocchio strategy retrieves more documents with rank 15 or less in three iterations. |
| Number of Concepts in Original Query and Number of Relevant Documents | High-Low or Low-High (17) | Queries having relatively many relevant documents and relatively few concepts or vice versa:<br>From 2-3 relevant and 7+ concepts (2)<br>From 4-5 relevant and 10+ concepts (5)<br>From 5-6 relevant and 3-6 concepts (5)<br>7+ relevant and 3-7 concepts (5) |
| | Similar (19) | Queries having a number of concepts and a number of relevant documents similar in magnitude:<br>From 2-4 relevant and 3-6 concepts (8)<br>From 4-6 relevant and 7-9 concepts (6)<br>6+ relevant and 8+ concepts (5) |

| Strategy | $Q_o$+ | | | Rocchio | | |
|---|---|---|---|---|---|---|
| Iteration | 1 | 2 | 1-2 | 1 | 2 | 1-2 |
| Low | 17 | 16 | 13 | 16 | 19 | 17 |
| High | 19 | 20 | 17 | 20 | 17 | 19 |

Some Query Subgroups Investigated

Figure 22

subgroup is low, and statistical significance is difficult to prove for small samples. For these reasons significance levels from five to ten percent may indicate areas for productive investigation using larger query collections and perhaps more sophisticated statistical techniques. Twenty-two variables are used for WRS comparisons within each subgroup pair. Two are not generated by the search process; the number of concepts in the initial query and the number of relevant documents (2 vars.). The three search-related measures used are correlation of the modified query with the original query, feedback effect normalized recall, and feedback effect normalized precision. Normalized recall and precision are calculated for the initial search (2 vars.) and all three measures are calculated for two iterations of two feedback strategies, the positive feedback $Q_0+$ strategy and the Rocchio algorithm which uses negative feedback (12 vars.). For normalized recall and precision, the improvement caused by feedback over the initial search is used for comparison to remove the effect of initial search differences between subgroups. To provide a direct comparison between positive and negative feedback, the differences between the $Q_0+$ and Rocchio strategies in normalized recall and precision for two iterations are used (4 vars.). Finally, the difference between the first and second iteration correlation of the modified query with the original query is calculated for each strategy (2 vars.). Obviously significant relationships such as the difference

in number of relevant documents between queries with four or fewer and queries with five or more relevant documents are not reported.

Normalized recall and normalized precision were chosen for the subgroup comparisons because they are overall measures of retrieval. However, the analysis in the previous section indicates that the normalized figures are not representative of overall performance as indicated by the recall-precision curves and the document curves. In particular, normalized recall shows a large drop for the Rocchio strategy, and neither recall nor precision reflects the initial advantage of the Rocchio strategy displayed in Figure 21. Therefore, the normalized measures may not be the best choice for meaningful comparison of positive and negative feedback.

Figures 24, 28, and 29 in this section display recall-precision curves. Unfortunately, significance tests between subgroups for recall-precision curves are not available. However, in three subgroup pairs, selected by strategy, performance, and number of relevant documents, Wilcoxon Signed Rank tests of the difference between the $Q_0+$ and Rocchio strategies within each subgroup were made. All differences were significant in both strategy subgroups; no differences were significant in any other subgroup.

In the figures of this section, the average values of variables as well as the WRS probabilities are presented. It should be noted that the WRS probabilities do <u>not</u> indicate

the significance of differences in average value, but the significance in differences in rank sum when all queries in both subgroups are ranked. The average value is reported because it is a more familiar figure and conveys more intuitive meaning than the rank sum.

The first subgroup pair listed in Figure 22 is familiar from the previous section. Figure 13 and 20 present total performance and feedback effect recall-precision curves for the 'eleven bad' group, both showing an advantage for the Rocchio strategy. Figure 23 presents some of the significant WRS findings for this group. The average number of relevant documents for the eleven bad queries is 4.3, contrasting with 5.6 for the remaining twenty-five queries. The WRS probability that these subgroups represent populations that have the same distribution of number of relevant documents is less than ten percent, so the difference is of doubtful significance. When the $Q_0+$ improvement is compared to the Rocchio improvement, the normalized recall and precision indicate an advantage for the $Q_0+$ strategy in both subgroups. This finding contradicts the recall-precision curves presented earlier, and is misleading for the reasons stated early in this section. The meaningful conclusion to be drawn from Figure 23 is that the differences in feedback improvement between subgroups are not significant except for the first iteration of the $Q_0+$ strategy. That is, the performance of the Rocchio strategy does not depend on whether or not relevant documents are

'Eleven Bad' Group: Eleven queries that retrieve no relevant documents with rank 5 on the initial search.

'Twenty-Five Group: Twenty-Five queries that retrieve some but not all relevant documents within rank 5 on the initial search.

| | | | Eleven Bad | Twenty-Five | WRS Probability |
|---|---|---|---|---|---|
| Number of Relevant Documents | | | 4.3 | 5.6 | <10% |
| Initial Search | | NR | 76.4 | 88.0 | <02 |
| | | NP | 42.8 | 72.0 | <01 |
| First Iter. | Improvement $Q_o$+ Strategy | NR | 0.0 | 5.2 | <05 |
| | | NP | -0.2 | 4.6 | <02 |
| | Improvement Rocchio | NR | -18.7 | 3.3 | <10 |
| | | NP | -0.5 | 3.2 | null |
| Second Iter | Improvement $Q_o$+ | NR | 1.7 | 5.3 | null |
| | | NP | -0.3 | 5.0 | null |
| | Improvement Rocchio | NR | -10.9 | 3.5 | null |
| | | NP | -2.1 | 3.7 | null |
| Correlation of modified query with original query | $Q_o$+ Strategy | Iter 1 | 100.0 | 78.5 | <01 |
| | | Iter 2 | 95.9 | 78.8 | <01 |
| | | Iter 1-2 | 4.1 | -0.4 | <02 |
| | Rocchio Strategy | Iter 1 | 59.7 | 43.8 | <01 |
| | | Iter 2 | 50.2 | 45.8 | null |
| | | Iter 1-2 | 9.4 | -0.1 | <01 |

Characteristics of

Subgroups Selected By Initial Search Retrieval

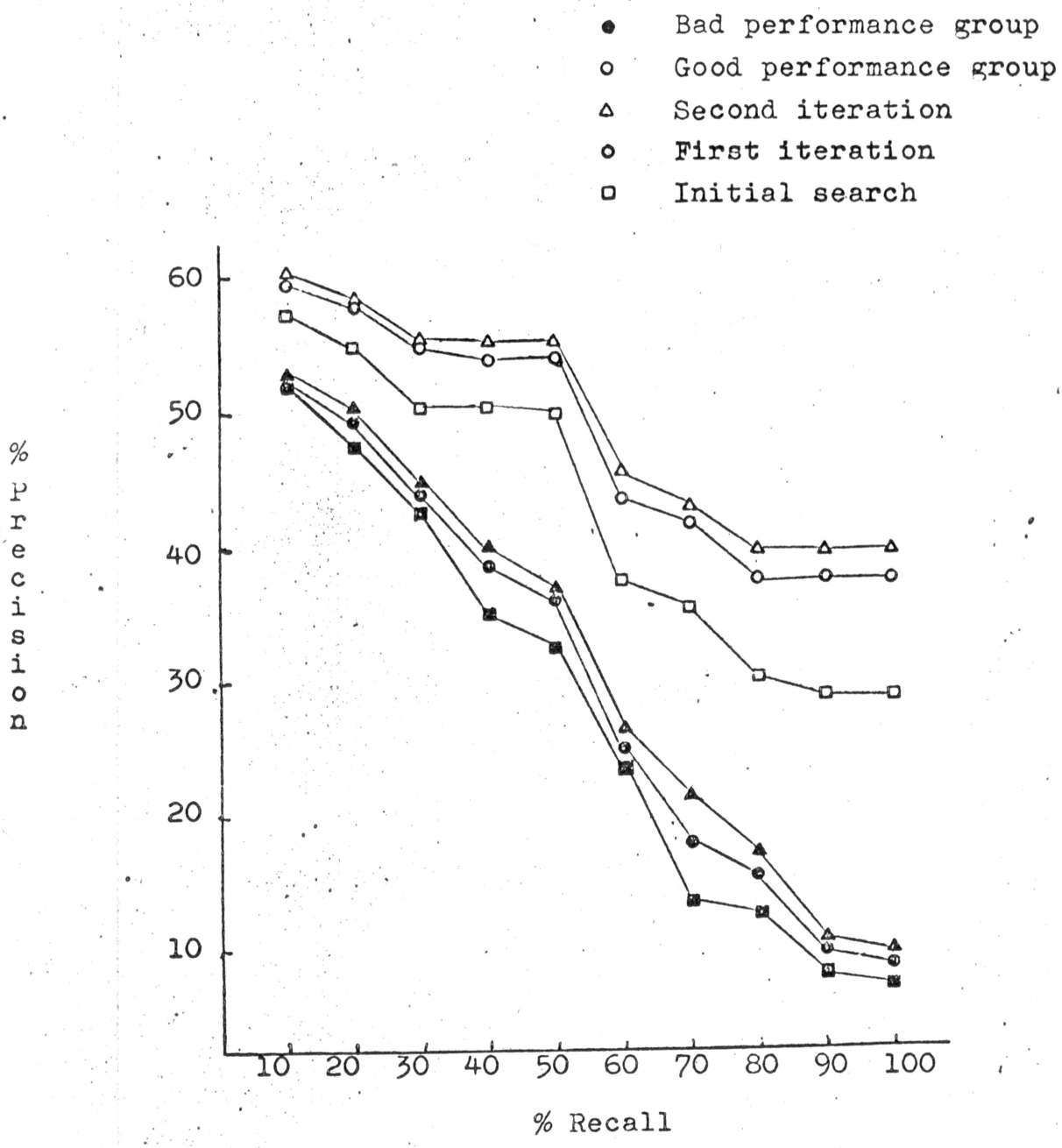
Figure 23

available for feedback. This conclusion agrees with the recall-precision curves for this subgroup. It should be noted that the low average normalized figures for the eleven bad queries are due to a single query, query 34, that is destroyed by all negative feedback strategies. Since magnitude is reflected in the averages but not in the WRS test, query 34 has a disproportionate effect on the average figures but does not similarly bias the probabilities.

The comparisons using correlation of the modified query with the original query show stronger differences between subgroups than the performance comparisons. The Rocchio strategy changes the query more in both subgroups, as expected. The $Q_0+$ strategy changes the eleven queries not at all on the first iteration and very little on the second. The figure for the first iteration correlation minus the second iteration correlation indicates that the eleven bad queries tend to move further away from the original query on the second iteration, but the twenty-five queries tend to stay about the same distance from the original query. The direction of the Rocchio strategy comparisons is the same as that of the $Q_0+$ comparisons, but all correlations are much weaker. The eleven bad queries change significantly less than the twenty-five on the first iteration, but on the second the amount of change no longer differs. The tendency for the eleven bad queries to move further away from the original query on the second iteration is stronger for the Rocchio strategy, so that the second iteration change compensates for the lack of change in the first iteration.

Figure 24 presents feedback effect recall-precision curves for the subgroups selected by performance. These curves seem to indicate not only better initial search performance, but also greater first iteration improvement for the good performance group. The precision level of the good performance group drops less as recall increases than that of the bad performance group. The normalized recall and precision reported in Figure 25 indicate better initial search but slightly worse first iteration feedback for the good performance group. The first and second iteration feedback improvement differences are not significant. It is interesting to note that the eleven bad queries on initial retrieval had fewer relevant documents than the remaining twenty-five, but the bad performance group tends to have more relevant documents than the good performance group. Also noteworthy is the significant tendency of the second iteration Rocchio query to move further away from the original query in the bad performance group, as though searching farther afield for relevant documents. This tendency is not observed for the $Q_0+$ strategy.

Figure 26 describes the general behavior of the modified queries in relation to the initial query. For both strategies the first iteration and second iteration queries tend to be similar in correlation with the original query. For the $Q_0+$ strategy, queries that don't move far from the original query on the first iteration tend to move farther on the second; this tendency is much weaker for the Rocchio




Subgroups Selected By Performance
Feedback Effect Recall-Precision Curves
Rocchio Strategy

Figure 24

**Good Performance Group:** Sixteen queries that retrieved all relevant documents with rank greater than 16 in three iterations of at least one feedback strategy.

**Bad Performance Group:** Twenty queries that did not retrieve all relevant documents with rank greater than 16 in three iterations of any feedback strategy.

| | | | Good Performance | Bad Performance | WRS Probability |
|---|---|---|---|---|---|
| Number of Relevant Documents | | | 4.5 | 5.7 | <10% |
| Correlation of Modified query with original, Rocchio Strategy | | Iter 1 | 50.9 | 46.8 | null |
| | | Iter 2 | 51.1 | 41.6 | null |
| | | Iter 1-2 | -.2 | 5.2 | <02 |
| Initial Search | | NR | 90.6 | 79.5 | <01 |
| | | NP | 70.5 | 57.1 | <10 |
| First Iter. | Improvement, $Q_o$+ Strategy | NR | 2.2 | 4.8 | null |
| | | NP | 2.9 | 3.4 | null |
| | Improvement, Rocchio | NR | -5.3 | -1.9 | null |
| | | NP | -0.5 | 0.1 | null |
| Second Iter. | Improvement, $Q_o$+ Strategy | NR | 3.8 | 4.7 | null |
| | | NP | 5.2 | 3.4 | null |
| | Improvement, Rocchio | NR | -0.9 | -0.9 | null |
| | | NP | 2.5 | 1.4 | null |

Characteristics of

Subgroups Selected By Performance

Figure 25

| | | WRS Probability |
|---|---|---|
| Correlation of Modified Query with Original $Q_o+$ | First Iter./Second Iter | <01% |
| | First Iter./Iter 1-Iter 2 | <01 |
| | Second Iter/Iter 1-Iter 2 | null |
| Correlation of Modified Query with Original Rocchio | First Iter./Second Iter | <01 |
| | First Iter./Iter 1-Iter 2 | <10 |
| | Second Iter/Iter 1-Iter 2 | null |
| Correlation of Modified Query with Original $Q_o+$ / Rocchio | First Iter | <05 |
| | Second Iter | null |
| | Iter 1 - Iter 2 | <01 |

Cross-probabilities for

Correlation of Modified Query

With Original Query

Figure 26

strategy. The first iteration correlations for the $Q_0$+ strategy and Rocchio strategy tend to vary similarly, the second iteration queries are no longer related; but the movement between first and second iterations strongly tends to be in the same direction.

Figure 27 reports the characteristics of five of the six subgroups chosen by correlation of the modified query with the original query. For the $Q_0$+ strategy on both iterations, queries that are more correlated with the original query tend to have fewer relevant documents and inferior performance. These findings can be explained by the behavior of the eleven queries that do not retrieve relevant documents initially. For the Rocchio strategy, there is a slight counter-tendency for queries that remain more correlated with the original on the second iteration to have more rather than fewer relevant documents. The significant findings for the Rocchio strategy concern the direction of query change between first and second iterations. Queries that move further from the original query tend to have more relevant documents and poorer performance. This tendency agrees with the earlier finding in Figure 25. The subgroups chosen on the basis of $Q_0$+ query change between first and second iterations are not shown because for all variables the differences between subgroups support the null hypothesis.

Thus far no relationships explaining the differences in performance between positive and negative feedback have

| | | | Initial Search | | |
|---|---|---|---|---|---|
| | | | Number Relevant | Normalized Recall | Normalized Precision |
| Correlation of Modified Query with Original $Q_o$+ | First Iter | Low | 5.9 | 88.2 | 73.7 |
| | | High | 4.5 | 81.1 | 53.5 |
| | | WRS | <10% | <10% | <01% |
| | Second Iter | Low | 6.3 | 90.4 | 76.5 |
| | | High | 4.3 | 79.7 | 52.3 |
| | | WRS | <01 | <05 | <01 |
| Correlation of Modified Query with Original Rocchio | First Iter | Low | 5.1 | 84.6 | 65.7 |
| | | High | 5.3 | 84.3 | 60.9 |
| | | WRS | null | null | null |
| | Second Iter | Low | 4.9 | 80.3 | 58.4 |
| | | High | 5.4 | 89.1 | 68.2 |
| | | WRS | <10 | null | null |
| | Iter I minus Iter 2 | Low | 4.3 | 92.5 | 74.2 |
| | | High | 5.9 | 77.2 | 53.1 |
| | | WRS | <05 | <01 | <01 |

Characteristics Of

Subgroups Selected By

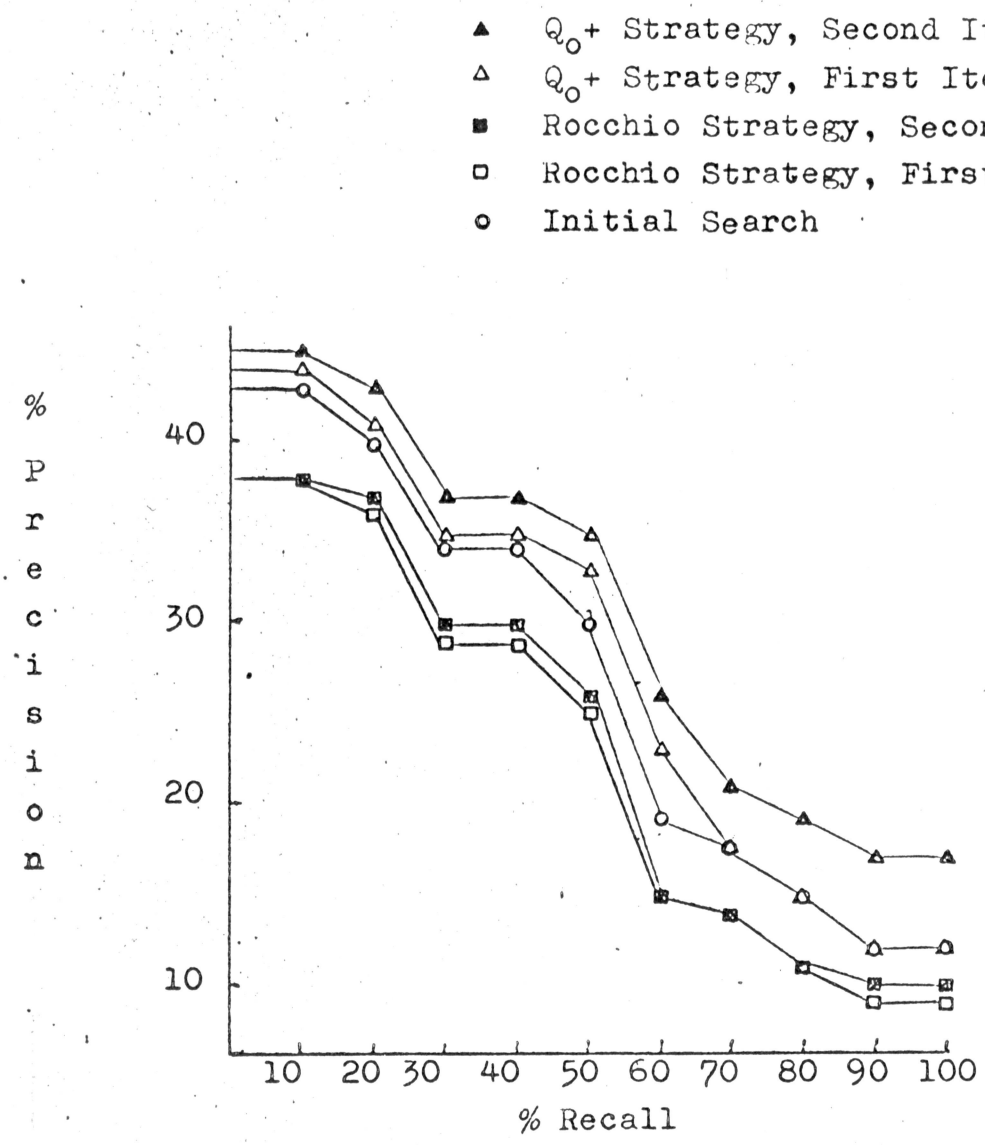
Correlation of Modified Query

With Original Query

Figure 27

been observed. No subgroup pair has shown significant differences on the four final variables; difference between $Q_0+$ and Rocchio strategies for first and second iterations. The strategy subgroups were chosen in an attempt to explore these differences from the opposite direction, to select the queries that display differences between positive and negative feedback and see if they also show other differences. Thirteen queries showing superior performance with $Q_0+$ and fifteen showing better performance with Rocchio were selected; the remaining eight queries showed no difference between strategies.

Figures 28 and 29 show the feedback effect recall-precision curves for each strategy in each group. In the $Q_0+$ group, the $Q_0+$ strategy causes slight improvement on the first iteration and more improvement on the second over the initial search, but the Rocchio strategy degrades performance. The initial search performance of the Rocchio group is higher than that of the $Q_0+$ group until 70% recall. Both the $Q_0+$ and Rocchio strategies improve performance in the Rocchio group, but the Rocchio improvement is greater. In Figure 29 the initial search on the remaining queries is graphed, showing that initial performance is far superior for those queries that have equivalent performance on both strategies. Figure 30 shows a similar pattern in the normalized recall and precision. In all cases, the improvement caused by the $Q_0+$ strategy is statistically equivalent in the two groups, but the Rocchio strategy



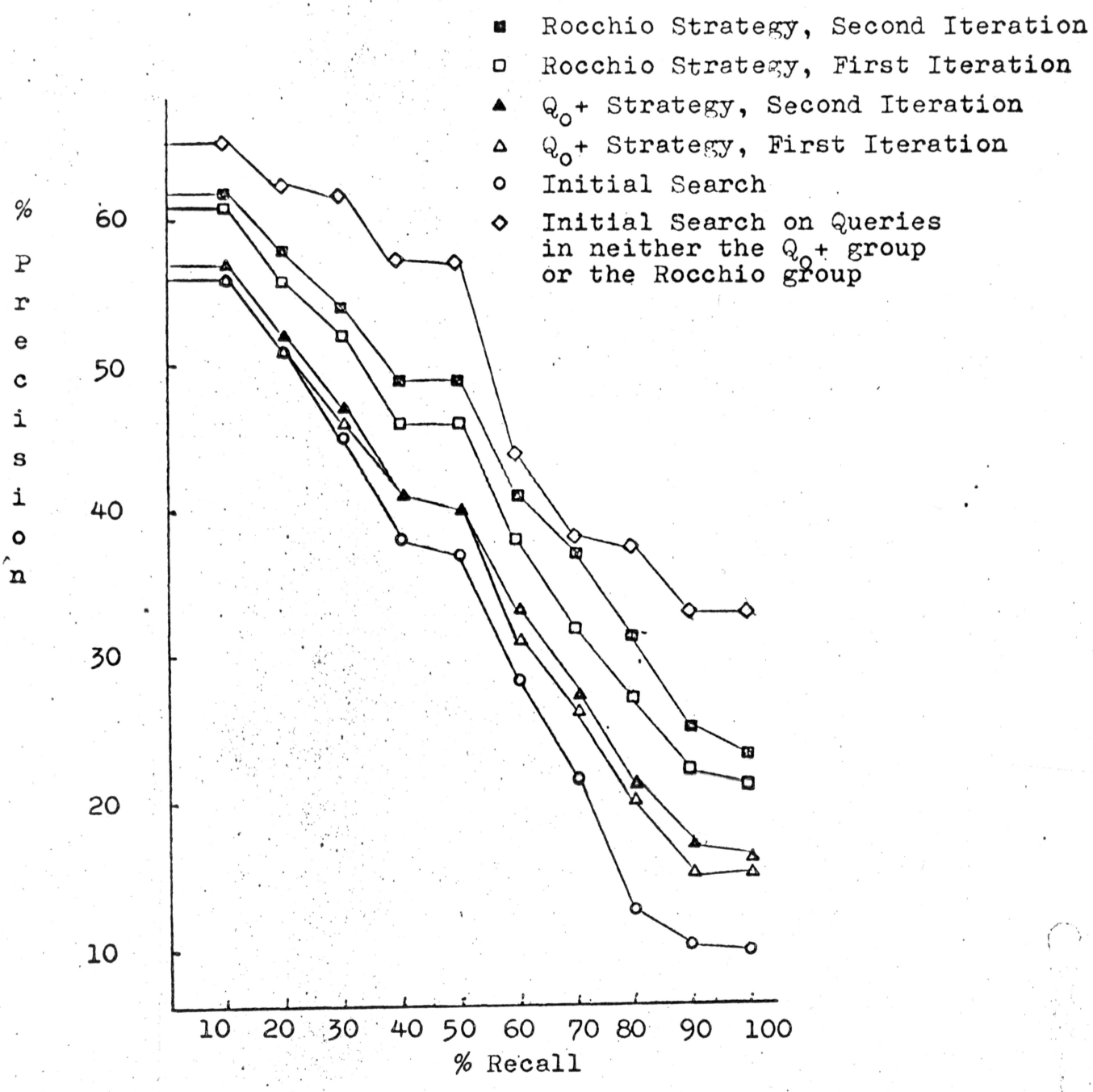

Subgroups Selected By Strategy

$Q_0$+ Group

Feedback Effect Recall-Precision Curves

Comparing Positive and Negative Feedback

Figure 28




Subgroups Selected By Strategy

Rocchio Group

Feedback Effect Recall-Precision Curves

Comparing Positive and Negative Feedback

Figure 29

| | |
|---|---|
| Q: | $Q_o$+ strategy |
| R: | Rocchio strategy |
| Q-R: | $Q_o$+ strategy minus Rocchio strategy |
| $Q_o$+ Group: | Thirteen queries that have better performance with the $Q_o$+ strategy. |
| Rocchio Group: | Fifteen queries that have better performance with the Rocchio strategy. |

| | | | $Q_o^+$ Group | Rocchio Group | WRS Probability |
|---|---|---|---|---|---|
| Initial Search | Normalized Recall | | 83.9 | 82.1 | null |
| | Normalized Precision | | 58.6 | 60.3 | null |
| First Iter | Normalized Recall | Q | 2.9 | 3.3 | null |
| | | R | -16.3 | 2.8 | <01% |
| | | Q-R | 19.2 | 0.5 | <01 |
| | Normalized Precision | Q | 1.9 | 1.5 | null |
| | | R | -12.0 | 7.2 | <01 |
| | | Q-R | 13.9 | -5.7 | <01 |
| Second Iter | Normalized Recall | Q | 4.5 | 3.2 | null |
| | | R | -14.7 | 7.5 | <01 |
| | | Q-R | 19.2 | -4.3 | <01 |
| | Normalized Precision | Q | 4.2 | 1.4 | null |
| | | R | -10.6 | 10.7 | <01 |
| | | Q-R | 14.8 | -9.3 | <01 |

Comparison of Positive and Negative Feedback

In Subgroups Selected by Feedback Strategy

Figure 30

degrades performance in the $Q_0+$ group. Except for normalized recall in the first iteration, the Rocchio strategy improves performance in the Rocchio group more than the $Q_0+$ strategy does. The hypothesis of greater variability in performance with the Rocchio strategy is again reinforced.

Unfortunately, the WRS tests show no differences between the $Q_0+$ and Rocchio groups except in feedback performance. The indication of the recall-precision curves that the Rocchio group is superior on the initial search is not supported by the normalized measures. No differences in number of concepts, in number of relevant documents, or in query correlations are found.

To further investigate strategy differences, three subgroup pairs are chosen based on feedback improvement in the normalized measures. One subgroup includes all queries that show feedback improvement over the initial search for all measures; normalized recall and precision for two iterations of both strategies. There are only thirteen queries in this group, because of the zero change in the 'eleven bad' queries with the $Q_0+$ strategy and the first iteration normalized recall plunge often encountered with the Rocchio strategy. The contrasting subgroup of the 'All Measures' pair contains the twenty-three queries that have zero or negative improvement on any measure. Two other subgroup pairs are chosen similarly, one by feedback in improvement on all measures of the performance of the Rocchio strategy (17 queries improved on all Rocchio measures, 19 did not),

and the other by feedback improvement for the $Q_0+$ strategy (16 queries improved on all $Q_0+$ measures, 20 did not).

Figure 31 displays the significant differences for the three subgroup pairs. A tendency for the improved queries to be less correlated with the original query is seen in the All Measures pair. This tendency is even more significant in the pair based on $Q_0+$ measures, but it disappears in the Rocchio Measures pair. The 'Eleven Bad' queries that retrieve no relevant documents on the initial search account in part for these findings. None of the queries in the 'Eleven Bad' group are in the 'All $Q_0+$ Measures' group. However, three of the eleven bad queries improve on all Rocchio measures. The eleven bad queries have high correlations with the original query, especially on the first iteration of the $Q_0+$ strategy when the correlation equals 1 for all eleven queries. The differences in correlation in the $Q_0+$ Measures subgroup pair cannot be entirely explained by the eleven queries, however. The eleven queries do not improve in both the All Measures and the $Q_0+$ Measures pairs, yet the $Q_0+$ differences in correlation are more significant than the All Measures differences.

Again an a priori choice of subgroup pairs forces significant differences in the performance of negative and positive feedback strategies. Figure 31 shows no significant

**All Measures Group:** Improved on all normalized measures for two iterations of both Rocchio and $Q_0$+ strategies.(13 queries)

**Not All Group:** Zero or negative improvement on any measure, any strategy, any iteration. (23 queries)

**All $Q_0$+ Measures Group:** Improved on all normalized measures for two iterations of the $Q_0$+ strategy. (16 queries)

**Not All $Q_0$+ Group:** Zero or negative improvement on any $Q_0$+ measure. (20 queries)

**All Rocchio Measures Group:** Improved on all normalized measures for two iterations of the Rocchio strategy. (17 queries)

**Not All Rocchio Group:** Zero or negative improvement on any Rocchio measure. (19 queries)

| | Iter | All Measures | Not All | WRS | All $Q_0$+ Measures | Not All $Q_0$+ | WRS | All Rocchio Measures | Not All Rocchio | WRS |
|---|---|---|---|---|---|---|---|---|---|---|
| $Q_0$+ | 1 | 77.8 | 89.1 | =1% | 77.9 | 90.8 | <1% | 82.5 | 87.3 | null |
| | 2 | 75.9 | 88.7 | <1 | 76.0 | 90.5 | <1 | 80.6 | 87.1 | null |
| Rocchio | 1 | 42.2 | 52.2 | <5 | 41.1 | 54.6 | <1 | 45.7 | 51.2 | null |
| | 2 | 42.2 | 47.8 | null | 40.5 | 50.0 | <10 | 42.8 | 48.5 | null |
| **Correlation of Modified Query With Original Query** | | | | | | | | | | |
| Iter 1 | NR | -0.4 | 11.3 | null | 2.6 | 12.5 | null | -2.0 | 15.2 | <1% |
| | NP | -1.2 | 4.8 | <5 | 0.1 | 4.7 | null | -5.6 | 10.0 | <1 |
| Iter 2 | NR | -0.8 | 8.4 | null | -0.2 | 9.3 | null | -4.0 | 13.1 | <1 |
| | NP | -1.6 | 3.2 | null | -0.3 | 2.9 | null | -6.7 | 8.8 | <1 |
| **$Q_0$+ Strategy Minus Rocchio Strategy** | | | | | | | | | | |

**Comparison of Negative and Positive Feedback**

**In Subgroups Chosen by Feedback Improvement**

**Figure 31**

relationship in the $Q_0+$ Measures pair with the differences between the $Q_0+$ and Rocchio strategies. However, in the Rocchio Measures pair all strategy difference measures show relationships significant at the one percent level. The relationships in these two subgroup pairs support the conclusion drawn from Figures 28 through 30 that the Rocchio performance variability creates the performance differences between strategies. A tendency for the thirteen queries that improve on all measures to favor the Rocchio strategy is significant only for first iteration precision improvement. This tendency supports the difference in recall-precision curves observed in Figures 28 and 29. In these figures the Rocchio strategy improves the Rocchio group more than the $Q_0+$ strategy improves the $Q_0+$ group on the first iteration.

Except for a tendency explained by initial search retrieval, no relationships have been found that can predict feedback improvement for the $Q_0+$ strategy or the Rocchio strategy. However, the lack of a relationship between feedback improvement and initial search performance is encouraging, since it indicates that relevance feedback causes as much improvement in original queries providing inadequate information as it causes in initially well-phrased queries.

Neither the experimental nor the analytical approach isolates a single variable that predicts performance differences between negative and positive feedback. At this

point, although several aspects of retrieval behavior have been detailed, initial search performance seems to be the

only effective predictor of final results. However, it seems anomalous that neither of the search-independent variables is related to any performance variable. No subgroup pair shows any difference in number of concepts in the original query, and differences in number of relevant documents are significant at the ten percent level or insignificant. Subgroups based on the number of concepts or on the number of relevant show no relationship with any variable.

Number of concepts is a measure related to the length of the query, and indicates the amount of detail with which the user has specified his needs. The number of relevant documents is an indication of how wide a subject area the user's query is intended to specify within the given document collection. Although these two variables are theoretically important to retrieval, each has no individual relationship to performance, and they are not related to each other. Therefore, it seems probable that the number of concepts in the original query and the number of relevant documents have some joint relationship to retrieval behavior. In fact, Figure 32 shows that these two variables combined are the desired predictor of performance differences between negative and positive feedback.

Two contrasting subgroups are chosen based on the relationship of the number of concepts to the number of relevant. In the 'Similar' group the two numbers are

either both low, both high, or both in mid-range. The contrasting 'High-low, Low-high' group contains those queries with few relevant and many concepts or with few concepts and many relevant. The Similar group attains significantly better performance with the Rocchio strategy than does the High-low, Low-high group. The differences between the $Q_0+$ and Rocchio strategies favor the $Q_0+$ strategy in the High-low, Low-high group and the Rocchio strategy in the Similar group. In short, every significant relationship in Figure 30 is echoed in Figure 32. The fact that some Figure 32 relationships are weaker can be attributed in part to the eight 'neutral' queries omitted from Figure 30 but included

'Similar' Group: Number of concepts in original query and number of relevant documents are similar in magnitude;

From 2-4 relevant and from 3-6 concepts, from 4-6 relevant and from 7-9 concepts, 6 or more relevant and 8 or more concepts.

'High-low, Low-high' Group: Few concepts and many relevant or few relevant and many concepts. Not meeting the criteria of the 'similar' group.

| | | | High-low Low-high | Similar | WRS Probability |
|---|---|---|---|---|---|
| Initial Search | Normalized Recall | | 82.2 | 86.4 | null |
| | Normalized Precision | | 60.0 | 65.8 | null |
| First Iter | Normalized Recall | Q | 4.8 | 2.5 | null |
| | | R | -10.7 | 3.1 | < 10% |
| | | Q-R | 15.6 | -0.5 | < 01 |
| | Normalized Precision | Q | 3.5 | 1.6 | null |
| | | R | -6.8 | 5.8 | < 05 |
| | | Q-R | 10.3 | -4.2 | < 01 |
| Second Iter | Normalized Recall | Q | 6.1 | 2.5 | null |
| | | R | -10.4 | 7.7 | < 05 |
| | | Q-R | 16.5 | -5.2 | < 01 |
| | Normalized Precision | Q | 5.4 | 1.6 | null |
| | | R | -6.1 | 9.0 | < 02 |
| | | Q-R | 11.5 | -7.5 | < 01 |

Comparing Positive and Negative Feedback
In Subgroups Selected By
Number of Concepts in Original Query
and Number of Relevant Documents

Figure 32

in Figure 32. Three of these fall in the Similar group; the remaining five in the contrasting group. The average differences in Figure 32 compare favorably to those in Figure 31. Except for first iteration normalized recall, the differences between the Similar and High-low, Low-high groups are as great or greater than the differences between queries that improve on all Rocchio measures and those that don't.

The joint relationship of query size and number of relevant documents is of little use for prediction in an operating retrieval system, since the number of relevant documents in the collection is not known at the beginning of the search. However, some estimator of the number of relevant documents might be available to the system before feedback. The user could be asked to state whether he intends his query to be specific or general, and some users might even be able to estimate the number of relevant documents available. In a larger collection the number of relevant documents retrieved on the initial search might be useful for prediction as the number of relevant documents available. In this collection the number of relevant retrieved by the original query when N equals 5 correlates highly with the number of relevant documents in the collection. Spearman's coefficient of rank correlation is significant at the one percent level [21]. However, the number of relevant documents retrieved can range from 0 to 5 only, and this range does not provide sufficient information for prediction of differences in performance of negative and positive feedback strategies.

When the number of relevant retrieved and the number of concepts are used to predict strategy differences, the WRS test results support the null hypothesis. Nevertheless, a search for a predictive relationship between query size and some estimator of the number of relevant documents might well be profitable in a larger collection.

The results in Figure 32 indicate the possibility of taking advantage of the performance differences between negative and positive feedback by choosing in advance the appropriate strategy for each query. Another approach is to develop a single algorithm that causes feedback improvement on all queries. With this possibility in mind, the factors causing the failure of the Rocchio algorithm on some queries in the High-low, Low-high group should be investigated. It is evident from earlier results that the inferior Rocchio performance on some queries is not caused by a failure to retrieve relevant documents on the initial search. In fact, the possible obliteration of the initial query by subtraction of non-relevant documents does not appear to be a general problem. Only query 34 is reduced to zero by the Rocchio strategy. All other queries gain in length on the first iteration. Of the ten queries that lose some concepts, seven gain in performance from the change.

The data presented in this section does not directly indicate the causes of the variability of the Rocchio strategy. In Section VII-C a hypothesis consistent with all experimental results is advanced to explain the contrasting behavior of positive and negative feedback.