VIII. User Interaction with an Automated Information Retrieval System

Eleanor Ide

Abstract

This study investigates relevance feedback, which is a procedure allowing user interaction with an automated information retrieval system. The user is given a small set of possibly relevant items, and is asked to judge each as relevant or non-relevant to his request. These user relevance judgments are then used for feedback to the information retrieval system, to produce a better subsequent set of retrieved items.

Several feedback strategies are evaluated for document retrieval, using a collection of 200 documents with 42 queries. The most promising method involves feeding back information contained in non-relevant documents retrieved, as well as that contained in relevant documents. Another useful procedure consists in feeding back document sets of different sizes to each user.

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 that of employing 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 non-relevant documents;

n_r = number of relevant documents in the collection;

n_s = number of non-relevant 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, $R' \in R$ and $S' \in S$, where $R' = \{r'_i\}$ is the subset of relevant documents retrieved and $S' = \{s'_i\}$ is the subset of non-relevant documents retrieved. Therefore, the following formula is used by Rocchio to construct a new query from the query of the previous iteration:

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

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

Rocchio investigated relevance feedback using the above formula and 17 queries, 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. Their formula is:

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

Three differences with Rocchio's formula are immediately apparent:

a) The descriptor vectors are not normalized by their length. In Rocchio's formula, the change to the weight of concept "a" in the query depends not only on the weight assigned to concept "a" in a retrieved document vector but also on the length of that document vector; that is, on the number of other concepts and on the magnitudes of weights in the document vector. This is not the case in the Riddle, Horwitz, and Dietz formula. When the latter formula is used, for instance, a document with generally highly weighted concepts changes the query more than does a document with generally lower weighted concepts, the number of concepts being equal. Weight magnitudes being roughly equal, a document with more concepts changes the query more than one with fewer. Rocchio's formula compensates for these effects.

b) The parameter α, which is the one variable in the above formula, is constant for all queries. Rocchio's formula uses a different multiplier for each query; the multiplier being dependent on the numbers of relevant and non-relevant documents retrieved (n_r' and n_s').

c) The non-relevant documents retrieved on the previous iterations are not used to update the query. However, Riddle, Horwitz, and Dietz tested a "negative heuristic strategy" which uses the two non-relevant documents first retrieved (the two which the system falsely judges most relevant to the query) to update those queries that retrieve no further relevant documents on the first feedback iteration. For such queries the formula becomes:

$$Q_{i+1} = Q_i + \alpha \sum_{1}^{n_r} r_i' - \sum_{1}^{2} s_i'$$

The feedback algorithm of Riddle, Horwitz, and Dietz produces an improvement in performance on most of the queries tested. The three experimenters recommend that the variable α in their formula be set to 1 for the first iteration and then increased by 1 for each subsequent iteration (called "increasing alpha strategy"). They also recommend their negative heuristic strategy.

The two studies here cited used a few queries for small document collections. The present study constitutes a further investigation of similar feedback algorithms in a somewhat more realistic environment (see section 2). An attempt is made to compare various feedback strategies based on average improvement in retrieval performance.

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. This collection is less "artificial" than the ADI collection for three reasons:

a) The Cranfield collection contains more documents and more queries.

b) The documents in the collection were chosen from a more typical environment. (The ADI collection consisted of <u>short</u> papers all published at the <u>same</u> time.)

c) The queries and relevance judgments were constructed by more and better qualified "users".

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 performs the same function as the α in the Riddle, Horwitz, and Dietz formula. The π permits the previous query to be increased in significance relative to the incoming documents. (This is impossible in the Riddle, Horwitz, and Dietz formula because all parameters are integers.) 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 non-relevant 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. These parameters are used in sections 3C and 3D. 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 non-relevant 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 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.

Three performance measures are used in this study. All are aggregate measures and none depends on a choice of a cutoff point; that is, all measure the retrieval performance over the entire document collection. The measure presented most often is the Quasi-Cleverdon curve of recall versus precision, calculated at each 5% of recall. [5] The other two are the normalized recall and precision measures proposed by Rocchio. [2]

3. Experimental Results

The following results are presented:

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.

b) A comparison of strategies which use only R', the set of relevant documents retrieved, to update the query. The different procedures are obtained by varying the parameters π, ω, and α in the query-update formula. The "increasing alpha strategy" of Riddle, Horwitz, and Dietz is included among the methods tested.

c) Results of an investigation of the effect of the number of documents given to the user on each iteration.

d) Results of an investigation of strategies for using set S', the non-relevant documents retrieved, to update the query.

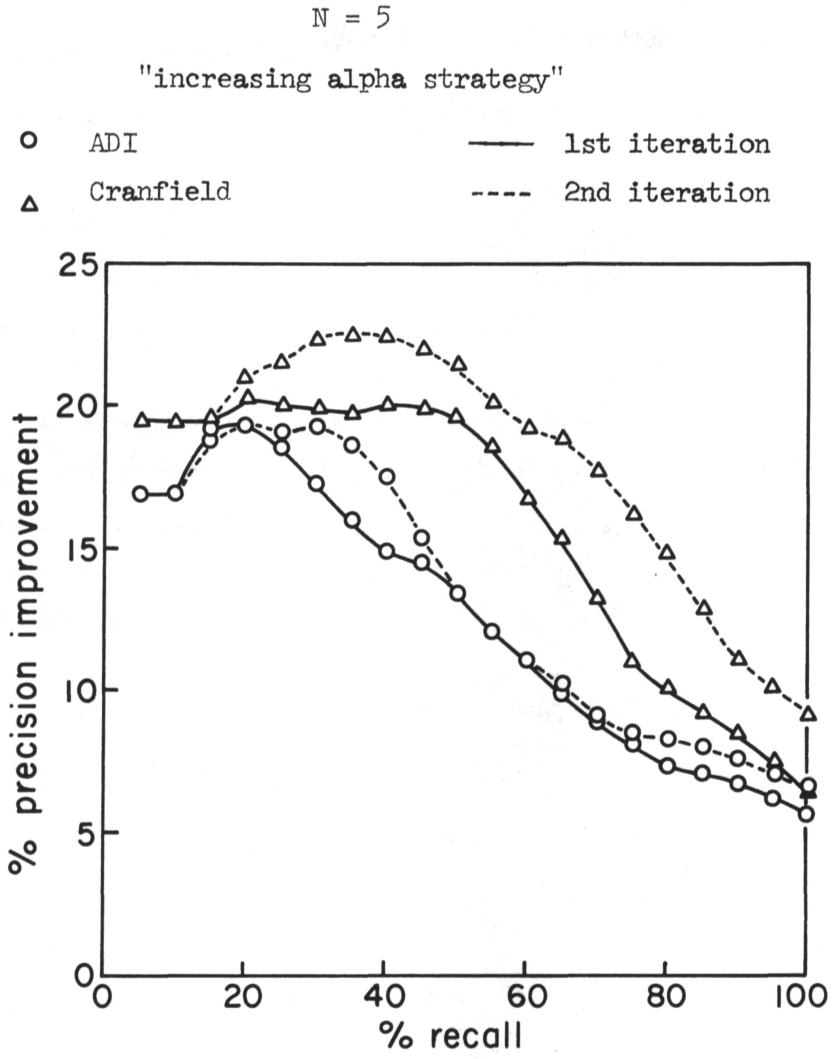
A) Comparison of the Cranfield and ADI Collections

The initial search results, before feedback, for the two collections are essentailly 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 the Cranfield performance 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. [6] The average generality number of the ADI collection is over twice that of the Cranfield collection.

Because the initial search results are somewhat at variance with each other, the improvement caused by feedback over the initial search 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.

Fig. 1 shows the differences in 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 collections would be expected to cause



Improvement Over Initial Search,
ADI and Cranfield Collections

Fig. 1

All experiments use a thesaurus dictionary and the cosine correlation function. All experiments are performed on the Cranfield 200 document collection unless otherwise stated.

less improvement in the larger collection. [6] 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.
b) The difference in collection scope. The ADI collection covers a wider subject area.
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.
d) The difference in query construction and relevance judgments.

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 are repeated for the Cranfield collection. To simulate their experiments with equation B, the parameter α is varied; π is kept equal to 1, and μ and ω equal to 0. Both the "increasing alpha" and "constant alpha" strategies are employed.

Rocchio used what amounts to different parameters for each query; that is, $\pi = n'_r\, n'_s$, $\alpha = n'_s$, $\mu = n'_r$ (see section 1). The experimental system is not now equpped to repeat Rocchio's algorithm.

Fig. 2 clarifies the effect of the "increasing alpha strategy" and the "constant alpha strategy" for the first, second, and third iterations of a feedback run. 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. Fig. 2 assumes that a document once retrieved is retrieved on all succeeding iteration; in the experimental system, the 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 the same significance to relevant documents whenever they are retrieved, it is necessary to substitute Q_o for Q_i in the formula; that is, to let $\pi = 0$ and $\omega = 1$ in equation B. This is called the "Q_o strategy" in Fig. 2.

Riddle, Horwitz, and Dietz 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 Fig. 2 give essentially the same results when $N = 5$. Using the Q_o strategy with different relative values of ω and α also does not change performance. Query update parameters (in equation B) for the six experiments performed are shown in Fig. 3. Among all six experiments, the differences in normalized precision and recall are less than 0.75% for all iterations.

	iter	Q_o	R^0	R^1	R^2
"increasing alpha strategy"	1	1	1	0	0
	2	1	3	2	0
	3	1	6	5	3
"constant alpha strategy"	1	1	1	0	0
	2	1	2	1	0
	3	1	3	2	1
"Q_o strategy"	1	1	1	0	0
	2	1	1	1	0
	3	1	1	1	1

Effects of Three Strategies on the Multipliers
of Documents Retrieved on Three Successive Iterations

Fig. 2

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

	π	ω	α
"increasing alpha"	1	0	1 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

Query Update Parameters for Six Strategies
Using Only Relevant Documents

Fig. 3

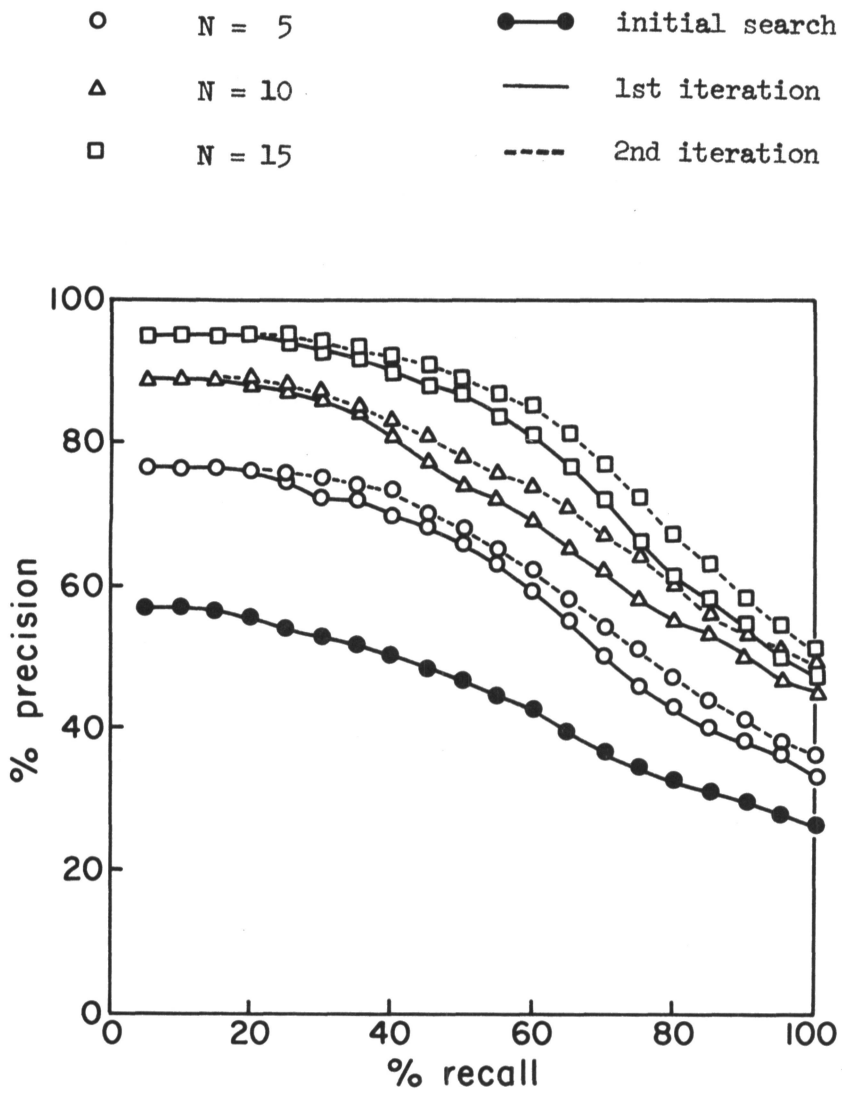
The 200 document collection seems quite insensitive to variations in the parameters π , ω, and α. The considerations mentioned in subsection A are probably relevant here also. This insensitivity indicates that perhaps the performance for the Cranfield collection is more stable in general than for the ADI collection. The effects of changes in the dictionary used to construct the descriptor vectors are also less pronounced for the Cranfield collection. [7]

C) Amount of Feedback

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.

Fig. 4 shows the performance of the "increasing alpha strategy" when 5, 10, and 15 documents are fed to the user for relevance judgments. The 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, especially at the high recall end of the curve.

It is important to note that certain users get no benefit from a 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 Fig. 4, in the table below the graph. This table probably explains



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

Fig. 4

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.

A variable feedback strategy is 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 (or perhaps two) relevant documents that he hasn't seen on previous iterations. 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 (or two) to improve the retrieval performance.

Fig. 5 shows the results of two iterations where the "user" is instructed to search the results until he finds one new relevant document or until he has seen 15 documents. The "X" curve in Fig. 5 shows what happens 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 Fig. 4 but by 50% recall has dropped to the second iteration N = 5 curve, which has been superimposed on Fig. 5. The table below the graph shows that the "average" user had to scan only four documents for

"Q_o strategy"

——— 1st iteration

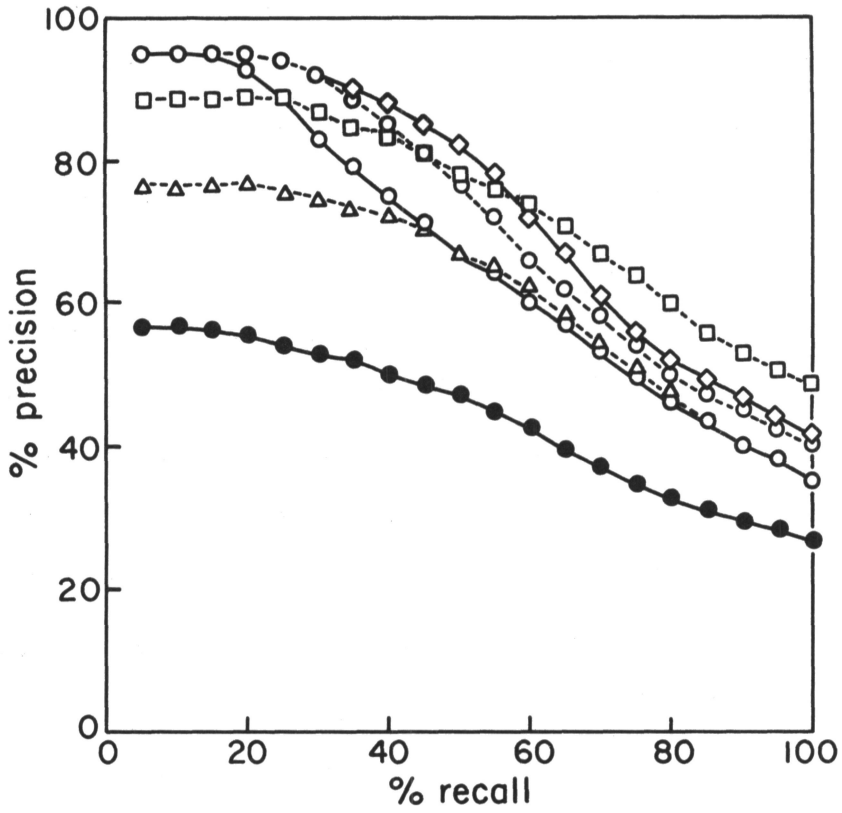
o feed back 1 new relevant

□ the 2nd iteration N = 10 curve (Fig. 4)

---- 2nd iteration

◇ feed back 2 new relevant

△ the 2nd iteration N = 5 curve (Fig. 4)



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

Feeding Back Only the First or First Two New Relevant Documents Found Within the First Fifteen Documents Retrieved

Fig. 5

feedback in order to achieve the performance displayed in the first iteration "o" curve. By contrast, each user looked at 10 documents to produce the second 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 second iteration N = 10 curve (20 documents scanned) at roughly 45% recall, and below the first iteration N = 10 curve (10 documents scanned) roughly 10% later (the second iteration N = 10 curve from Fig. 4 is superimposed on Fig. 5). 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. This fact 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 ("X") drops below the second iteration N = 10 curve at over 55% recall, and ends roughly 3% below the first iteration N = 10 curve. Alghouth 9 out of 42 users 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 ("X") 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. 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.

The experimental environment is not fully suited to study of variable feedback strategies, for one main reason. The programs for the experimental system are so designed that after the initial search, the "user" is given documents he has already seen. For example, the 10 documents fed back before the second iteration of the N = 10 curve are not necessarily ten new documents. Further study of this strategy should wait until the system has been reprogrammed with a viewpoint more appropriate for evaluating the experiments.

D) Strategies Using Non-relevant Documents

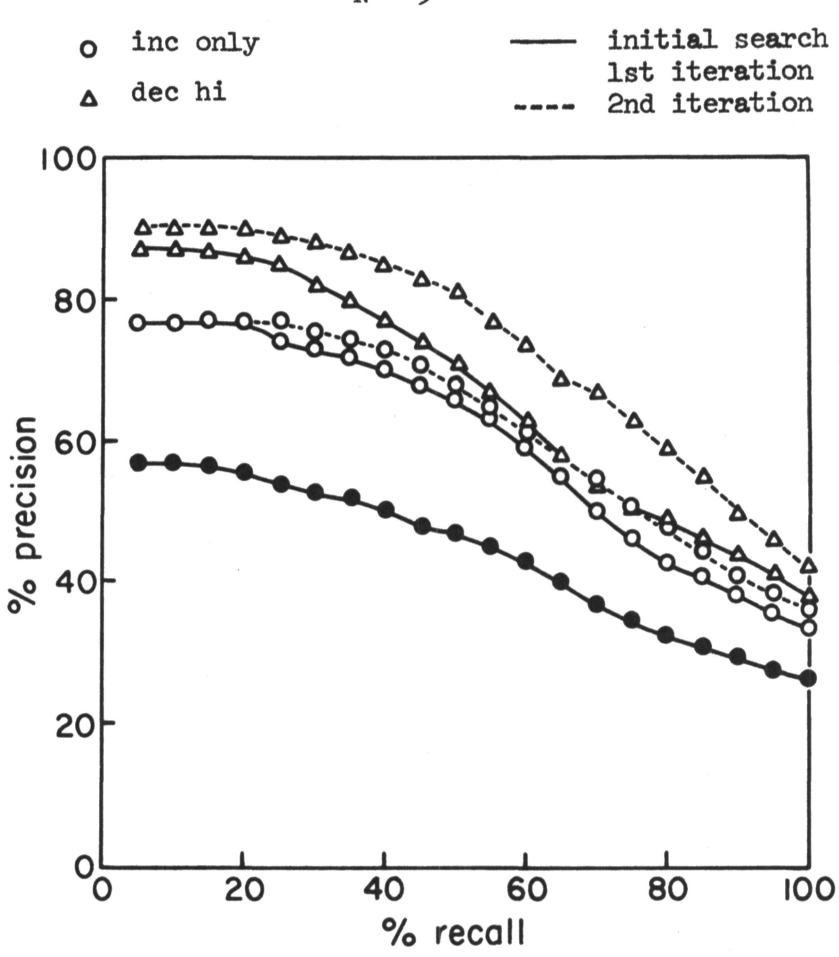
Rocchio's update formula (equation A) considers the information contained in the set of non-relevant documents retrieved (S') to be as important as that contained in the set of relevant documents retrieved (R').

If this is the case, the strategies so far examined are disregarding half the available feedback information! Further, information from non-relevant documents may help the user who finds no relevant documents in the first N on the initial search. It is seen in Fig. 4 that there are eleven such users out of 42 when N equals 5. The non-relevant document information therefore might be especially helpful in the N = 5 case.

However, problems arise in using the non-relevant documents in the experimental system under discussion. The effects of the description vector length and of the sizes of the sets R' and S' are compensated for in Rocchio's formula (see the discussion of the Riddle, Horwitz, and Dietz formula in section 1). The system used here (equation B) makes no such compensation, nor is there provision for negative weights in the query vector. Thus, if S' is large, and/or if the query length is small, there is danger that the query will be reduced to nearly the zero vector when the documents of S' are subtracted from it. The "negative heuristic strategy" of Riddle, Horwitz, and Dietz (see section 1) partly avoids this danger by using only the first two retrieved on the non-relevant documents. The present study compares results of using the first or the first two non-relevant documents retrieved.

Another consideration arises: whether the non-relevant documents should be used for all queries, or only for those not helped by the relevant documents. Rocchio did the former while Riddle, Horwitz, and Dietz did the latter (see section 1). The present study tries both.

Fig. 6 compares the strategy, called "dec hi", of decrementing each query be subtracting from it the first retrieved non-relevant document on each iteration, with the "Q_o strategy", which increments the query



Decrementing the Highest Non-relevant Document

Fig. 6

		Init	1	2
Normalized Recall	Inc Only	88%	90	90
	Dec High	88	87	90
	Dec 2 Hi	88	85	89
Normalized Precision	Inc Only	68	75	76
	Dec High	68	76	81
	Dec 2 Hi	68	75	80

Normalized Results for Non-relevant Document Strategies

Fig. 7

with the relevant documents but ignores the non-relevant (called "inc only" in this section). In the query update formula (equation B), the parameter values for "inc only" are $\pi = 0$, $\omega = 1$, $\alpha = 1$, $\mu = n_b = 0$, $n_a = N$. The values for "dec hi" are $\pi = 0$, $\omega = 1$, $\alpha = 1$, $\mu = -1$, $n_a = N$, $n_b = 1$. N is set equal to 5, which is, as noted earlier in this section, the number for which this strategy should cause the greatest improvement. Fig. 6 shows that the average results are consistently better for the "dec hi" strategy, especially on the second iteration.

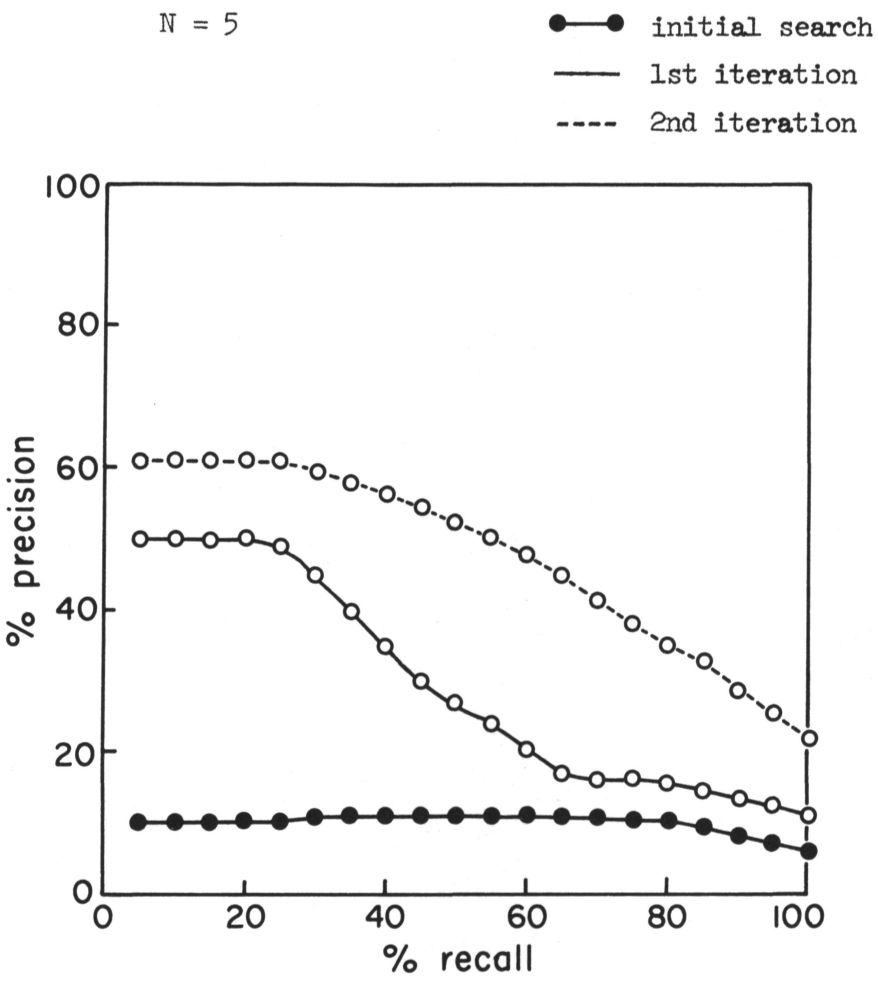
For this experiment, the implications of the normalized precision and recall results, given in Fig. 7, seem inconsistent with those of the recall-precision curves of Fig. 6. On the first iteration, the recall-precision curve for the "dec hi" strategy is above that for "inc only" at all recall levels. 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 which 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. In the individual query results, such behavior is in fact observed.

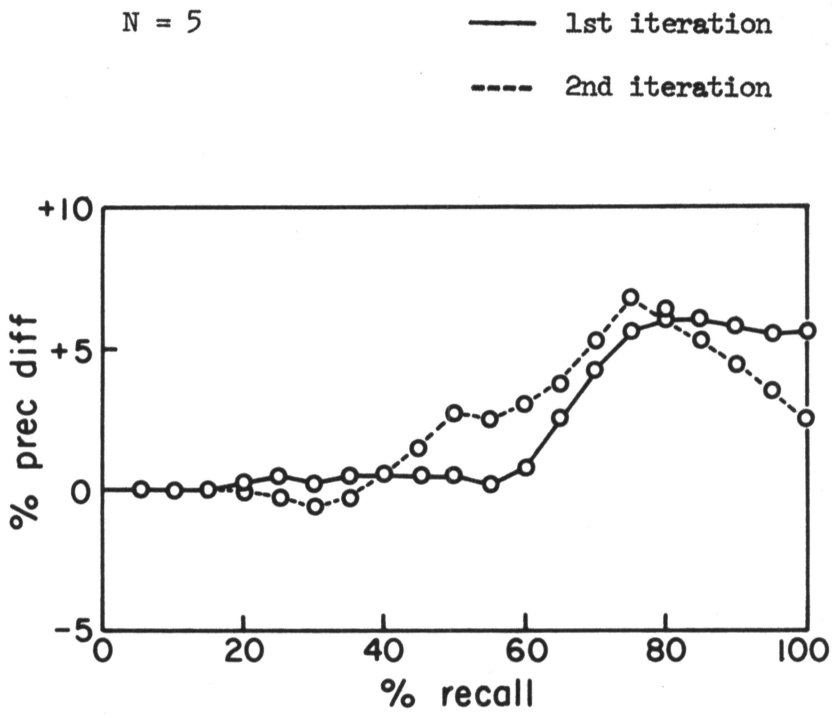
Fig. 8 shows how much the "dec hi" strategy helps the 11 users who receive no relevant documents in the first 5 on the initial search. For the "inc only" strategy, the initial search and all subsequent iterations are the same for these 11 users: the precision being about 10 percent. Feeding back one 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 for the 11 "bad" queries is so striking, it is natural to wonder whether this strategy is helping or hurting the other 31 users. Fig. 9 shows a difference curve for the "dec hi" and "inc only" strategies run on the 31 "good" queries only. A point above the zero line indicates that "dec hi" is better than "inc only" 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 is probably unnecessary and perhaps undesirable in this environment.



Decrementing the Highest Non-Relevant Document
on Eleven Bad Queries

Fig. 8

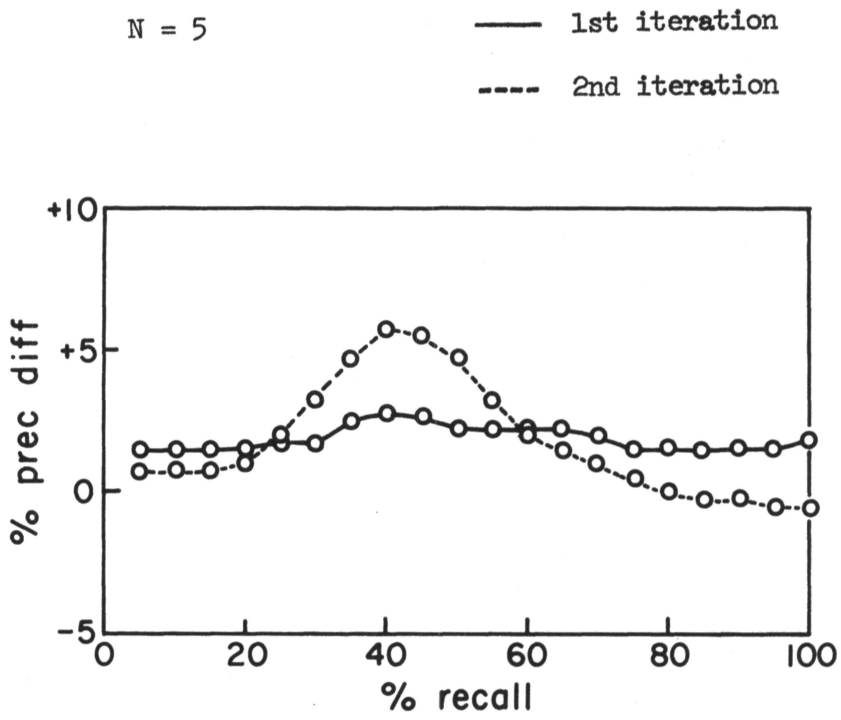


(Dec Hi) - (Inc Only)
For the 31 Good Queries, a Comparison Between
Decrementing the Highest Non-relevant
and Incrementing the Relevant Only

Fig. 9

Fig. 10 represents a difference curve comparing the "dec hi" strategy with the alternative of decrementing the query by subtracting the two highest non-relevant documents retrieved on each iteration (called "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 Fig. 7 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 is probably 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 "dec hi"). It might be possible to overcome this "disappearing query" phenomenon by juggling the parameters π, ω, α, and μ, without introducing the complications of Rocchio's normalizing method.

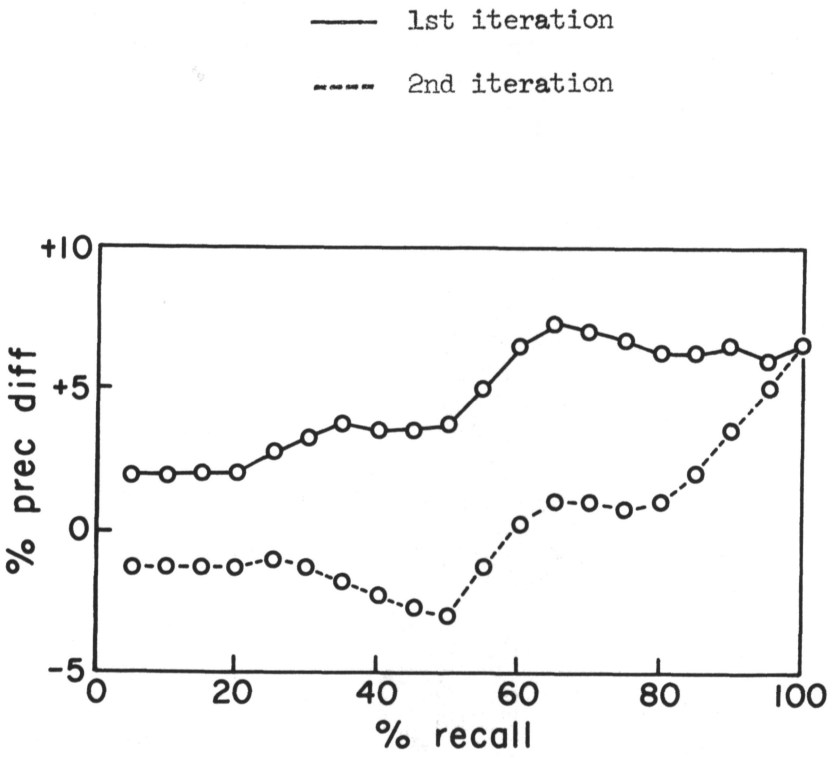
It was mentioned in Section 3B that much of the improvement between the N = 10 and the N = 5 curves of Fig. 4 might be caused by the improvement on the six queries that fetch a relevant document within the first 10 but not the first 5 on the initial search. Seven users out of the unlucky 11 are helped 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 performance of the N = 10 curve. Fig. 11 shows a difference curve between the N = 10 curve of Fig. 4, and the "dec hi" curve of Fig. 6. 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 iteration, discussed earlier in this section. For the second iteration, the "dec hi" curve is slightly better at the low recall end and only



(Dec Hi) - (Dec 2 Hi)

A Comparison of Decrementing
the Highest or the Two Highest Non-relevant Documents
on Each Iteration

Fig. 10



(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

Fig. 11

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), these results strongly favor this non-relevant document retrieval strategy. Feedback of non-relevant documents certainly deserves further investigation.

4. Conclusions

Four areas are investigated in the study:

A) Relevance feedback performance of the Cranfield and ADI collections are compared, using the query-update formula of Riddle, Horwitz, and Dietz. The Cranfield collection shows greater improvement in performance due to relevance feedback. When evaluating information retrieval experiments, some have argued that good results might be obtainable only in the small, artificial environments of such experiments. The present study provides a strong counter-argument in the case of relevance feedback, which performs better for the less artificial collection.

B) The "increasing alpha" and "constant alpha" strategies of Riddle, Horwitz, and Dietz are compared with several other strategies involving only the relevant documents retrieved. The differences among all six strategies investigated are insignificant. It is possible that the Cranfield collection is more stable in performance than is the ADI collection used by Riddle, Horwitz, and Dietz.

C) The effect of the number of documents fed back to the user on relevance feedback performance is investigated. Given a choice between scanning 5 or 10 documents on each iteration, a user might be willing to

double his effort to achieve the improvement in performance observed.

A variable feedback strategy to save user effort is studied: each "user" is "asked" to look at retrieved documents only until he finds one relevant document, which is at once used to update the query. In terms of user effort the first iteration of this strategy gives good results compared to the usual procedure of feeding back a fixed number of documents to each user. Results of the second iteration of the "feedback of one relevant" strategy are better at low recall but worse at high recall than those of the usual strategy. When each user is instructed to look for two relevant documents in the initial search output, results are better than the second iteration of the feedback of one relevant document. This experiment indicates that the second relevant document found on the initial search is more useful for retrieval than is the first new relevant found after feeding back one relevant to form the first iteration query. This result implies that the variable feedback strategy, which limits feedback to one or two new relevant for all queries, is not the way to achieve the best performance, especially at high recall levels. A combination strategy is proposed which may merge the advantages of the usual fixed feedback strategy with those of the variable feedback strategy. It is noted that further experiments in variable feedback should be conducted in a more appropriate experimental system.

D) The results of feedback strategies that use the non-relevant documents retrieved to update the query are encouraging. Because in the experimental system of the present study there is danger of the query being reduced to zero, the strategy (called "dec hi") that subtracts only the

first retrieved non-relevant document from the previous query gives the best performance. Especially after two feedback iterations, the results of the "dec hi" strategy when the user is given five documents for feedback are comparable to those of the strategy using only relevant documents, when the user is given ten feedback documents.

Two undesirable effects of the "dec hi" strategy are noted, the tendency for some low-ranking relevant documents to be lowered still more on the first iteration, and the possibility of erasing most of the query by subtracting large non-relevant document vectors. At least the latter effect might be overcome by using different parameters in the query update formula.

The present investigation supports relevance feedback as an information retrieval strategy (A above). 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 (B).

The most promising strategies investigated, variable feedback and non-relevant document feedback (C and D), 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 non-relevant feedback strategies should be studied in a system which permits Rocchio's normalizing. [1] Eventually, some combination of fixed and variable feedback may prove optimal in similar information retrieval environments.

Acknowledgments

The advice and encouragement of Michael Keen, and the programming support of Robert Williamson are gratefully acknowledged.

References

[1] J. J. Rocchio, Relevance Feedback in Information Retrieval, Report ISR-9 to the National Science Foundation, Chapt. 23.

[2] J. J. Rocchio, Document Retrieval System Optimization and Evaluation, Harvard University Doctoral Thesis, Report ISR-10 to the National Science Foundation, Chapters 3 and 5.

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

[4] O. W. Riddle, T. Horwitz, R. Dietz, Relevance Feedback in Information Retrieval Systems, Report ISR-11 to the National Science Foundation, Chapter 6.

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

[6] C. Cleverdon, E. M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 2, ASLIB Cranfield Research Project, 1966, Chapter 3.

[7] M. Lesk, G. Salton, Design Criteria for Automatic Information Systems, Report ISR-11 to the National Science Foundation, Chapter 5.