X. A Relevance Feedback System Based on Document Transformations

S. R. Friedman, J. A. Maceyak, and S. F. Weiss

Abstract

An information retrieval system using relevance feedback to modify the document space is proposed. A suitable algorithm is exhibited that achieves high precision and recall.

1. The Problem

An information retrieval system can be viewed as consisting of a document collection together with procedures for accessing and retrieving the documents in the collection. For evaluation purposes, the main goal of such a system can be taken to be the retrieval of all documents in the collection relevant to a request presented by a user. Hence, the problem of retrieval is to relate the user's need to a collection of information. How efficiently this relation can be generated depends in general on how the document collection is organized.

In traditional retrieval systems, i.e. libraries, books are grouped by subject matter according to some classification system such as the Dewey Decimal System or the Library of Congress System. Such classification systems divide human knowledge into categories such as physics, biology, etc. The virtue of all such classification systems is that books similar in content will often be located physically close to each other on the library shelves.

Libraries can be characterized as being static document spaces. That is, documents are assigned permanent positions in the document spaces.

Static document spaces suffer from two major weaknesses. First, given a classification system, it is possible to construct a request such that documents relevant to that request may not be located close to each other. For example, the topic "aerodynamics of birds" might be discussed in books which are primarily concerned with aerodynamics and only slightly concerned with birds, or, alternatively in books which are primarily concerned with birds and only tangentially with their aerodynamics.

Second, given a classification system, it is possible to imagine a document which is inherently difficult to classify. For example, should Church's Calculi of Lambda-conversion be placed with computing texts or with pure mathematics texts? The book relates to both areas.

Hence, static document spaces impose a classification system on the user which may not always suit his needs. The user's success with such a system varies as a function of how much his needs agree with the subject categories determined by the classification.

Most automated retrieval systems have inherited the static document space. These systems are normally more flexible than libraries, in that subject categories are permitted to overlap. However, the fact remains that documents are assigned permanent positions in the space. Hence, as with libraries, documents relevant to a given request might lie far away from each other in the space.

In the SMART system, the classification of a document is based on the importance of certain concepts in the document. A document is then represented by a vector whose elements are the relative weights of these concepts. A query is similarly represented. Hence, the query vector can

be thought of as positioned in the document space with access to all documents close to it. Since the user might not phrase his original query in the best possible manner, relevance feedback systems have been suggested to improve system operations. Most of these systems are based on successive modifications of the query vector using feedback information from the user. One such system described by Salton and Rocchio [4] has been implemented by Riddle, Horwitz, and Dietz. [2] The main weakness of the system lies in its inability to retrieve all documents relevant to a query when these documents do not lie close together in the document space.

The present study addresses itself to the question of whether a relevance feedback system can be devised that will retrieve such documents. The solution proposed is based on the use of a dynamic document space. Every user who submits a search request is given access to the standard document space. Subsequently, the document space is distorted in response to feedback information received from the user in an attempt to bring relevant documents closer to the query. Hence, the query is kept static, and the document space is altered. The standard document space is then no longer viewed as an absolute structure, but as an initial structure which can be altered to better suit the personal classification scheme of the user.

An implementation of such a relevance feedback system on the CDC 1604 is described here. The system has been tested for the 82-document ADI collection using queries for which a priori relevance judgments are available.

2. The Implementation

The initial task in implementing a relevance feedback system based

on document modification is to devise an algorithm that distorts the document space to bring relevant documents close to the query. The algorithm actually used is based on the assumption that concepts which are strong (highly weighted) in the documents judged relevant by the user and weak in the documents judged non-relevant tend to characterize relevant documents. Hence, the weights of these concepts should be raised in all documents in which they occur so as to increase the correlation between these documents and the query. Similarly, it is assumed that concepts which are strong in the non-relevant documents and weak in the relevant documents tend to characterize non-relevant documents and should be decreased in all documents in which they occur in order to lower the correlations.

The problem remains of how the concept weights should be raised or lowered. It seems natural that the amount of alteration for a concept should depend on how much that concept characterizes the relevant or non-relevant documents already seen by the user. That is, if a concept is judged to be a good positive discriminator (i.e. strong in relevant documents, weak in non-relevant documents), it should be raised by an amount proportional to how strong it is in the relevant documents. Similarly, concepts judged to be good negative discriminators (i.e. strong in non-relevant documents, weak in relevant documents), should be lowered by an amount proportional to how strong they are in the non-relevant documents. It also seems desirable to insist that the amount of alteration of a concept weight for a document should be proportional to that weight. That is, if a concept which is marked as a good discriminator is very weak for a particular document, then it should not be altered much since it doesn't

really characterize that document. On the other hand, if it is strong in the document, then it should be altered more since it provides a better characterization of the document.

These considerations were used to determine the final form of the algorithm:

a) Enter request. Retrieve the 10 documents having highest correlation with the query.

b) Use a priori relevance judgments to determine relevant and non-relevant documents.

c) Compute R_i, N_i, and D_i where

$$R_i = \frac{\text{The sum over all relevant documents of the weights of concept } i}{\text{The number of relevant documents}}$$

$$N_i = \frac{\text{Sum over all non-relevant documents of the weights of concept } i}{\text{The number of non-relevant documents}}$$

$D_i = R_i - N_i$ = The discrimination value of concept i.

d) Choose those concepts which seem to be good discriminators between relevant and non-relevant documents, i.e. those i for for which $|D_i| > \delta$. To these concepts, add all concepts which occur in the query.

e) For the i obtained in d), compute F_i^1 , F_i^2 , F_i^3 where

$$F_i^1 = \frac{\text{Weight of concept } i \text{ in query}}{\text{The sum of the weights of all concepts in query}}$$

$$F_i^2 = \frac{\text{Sum of the weights of concept } i \text{ in relevant documents}}{\text{Sum of the weights of all concepts in relevant documents}}$$

$$F_i^3 = \frac{\text{Sum of the weights of concept } i \text{ in non-relevant documents}}{\text{Sum of the weights of all concepts in non-relevant documents}}$$

These three numbers are measures of the importance of concept i in the query, the relevant documents, and the non-relevant documents respectively.

f) For the i obtained in d), compute T_i where

$T_i = \alpha_1 F_i^1 + \alpha_2 F_i^2$, if $D_i > \delta$, or if the query has concept i.

$T_i = -\alpha_2 F_i^3$, if $D_i < \delta$ and if the query does not have concept i.

g) Alter the document space by changing the weights of the concepts i obtained in d).

$w_i^j \leftarrow w_i^j + w_i^j T_i$ if document j has not been judged non-relevant, where w_i^j is the weight of concept i for document j and where the left arrow denotes assignment. If document j has been judged non-relevant, all its concept weights are set to zero. This step has the effect of moving documents judged non-relevant as far from the query as possible.

h) Return to a) using the same query and the altered document space. Actually g) is performed during the search in step a). This loop is continued until the user is satisfied with his results or until he gives up.

The algorithm makes use of three parameters: δ, α_1, α_2. δ is the cut-off point for discrimination values. High values for δ lower the number of good discriminators. α_1 and α_2 are the respective weights given to the query and to the relevant and non-relevant documents to denote their importance in adjusting the document space.

The algorithm was originally intended to use the cosine correlation function for correlating the query to documents. The cosine correlation

function is given by

$$\frac{\sum_{i=1}^{m} (q_i d_i)}{\sqrt{\left(\sum_{i=1}^{m} (q_i)^2\right)\left(\sum_{i=1}^{m} (d_i)^2\right)}}$$

where m is the number of indexing concepts, q_i and d_i are the weights of the i^{th} concept in the query and documents vectors respectively. However, early experiments with the algorithm gave better results with a modified version of the cosine function. In this version, the denominator is kept constant (i.e. equal to its value in the initial search) through all updates of the document space. This modification has the effect of magnifying all changes to the document vectors. It was decided to experiment with each of these correlation functions in the algorithm.

The algorithm was implemented by modifying existing relevance feedback programs which used query altering. Queries from the ADI collection were processed using various choices for α_1, α_2, δ and correlation function. After processing 11 queries, the program obtains precision-recall curves averaged over all 11 queries. These curves are given in Figs. 1 through 6.

3. Experimental Results

In general, modification of the document space using relevance feedback information leads to significantly better results than modification

of the query. The highest averages of normalized precision and normalized recall achieved by Riddle, Horwitz, and Dietz using query modification are .7554 and .8386 respectively. Corresponding averages using document modification are .8576 and .8871 for the modified cosine correlation function and .7987 and .8806 for the standard cosine correlation function. This improvement can also be seen by comparing Figs. 1 and 7. Fig. 1 gives recall-precision curves averaged over 11 requests using document space modification. Fig. 7 gives typical corresponding curves for query modification.

Tables 1 and 2 show the reason for this improvement. Table 1 lists the relevant documents retrieved by query modification and by document modification. For queries 11, 12, 14, and 15, document modification techniques retrieved more relevant documents. Table 2 gives the documents retrieved by both techniques for these queries and the number of concepts each such document has in common with the relevant documents retrieved by only document modification techniques. In most cases, this number is 1, indicating very little similarity between the documents retrieved by both techniques and those retrieved by only document modification. Hence, the improved results achieved by document modification are due to its ability to retrieve documents unlike each other but relevant to the query.

Figs. 1 through 6 show the effect of different values of δ on recall-precision curves averaged over 11 requests. The curves for the third iteration for δ equal to 3 and 5 are almost identical and are better than the third iteration curves for δ equal to 11. Also, the curves for the second iteration for δ equal to 3 are better than the curves for the second iteration for δ equal to 5. Hence, smaller values of δ cause the relevant documents to converge faster to the query.

Query No.	DOCUMENT NUMBERS QUERY MODIFICATION		DOC. MODIFICATION	
	Retrieved	Not Ret.	Retrieved	Not Ret.
9	50,82		50,82	
10	29,39,2	50,11	50,2,39	11,29
11	43	79,81,8	43,79	81,8
12	4	42,9,25,7	4,25,7,42,9	
13	1,14,37,27	80,65	1,37,65	14,27,80
14	20	33	20,33	
15	11,37	36,6,32,67	11,37,6	32,36,67
26	5		5	
28	69,9,48,70		69,9,48,70	

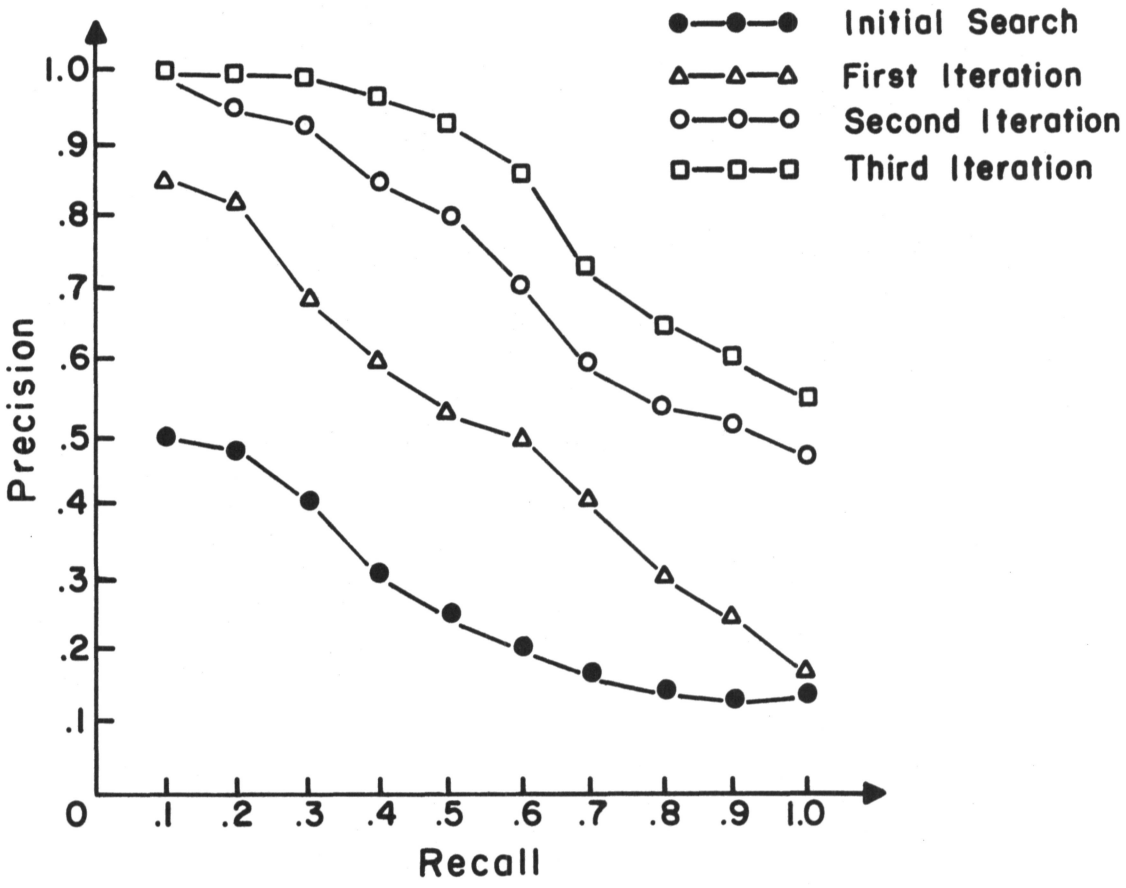
Comparison of relevant documents retrieved by query modification and document modification.

Table 1

Query No.	Retrieved by Query and Doc. Mod.	Retrieved only by Doc. Mod.	No. of Concepts in common
11	43	79	1
12	4	25	1
	4	7	1
	4	9	3
	4	42	2
14	20	33	1
15	11	6	1
	37	6	3

Relationship between documents retrieved by both techniques and those retrieved only by document modification.

Table 2

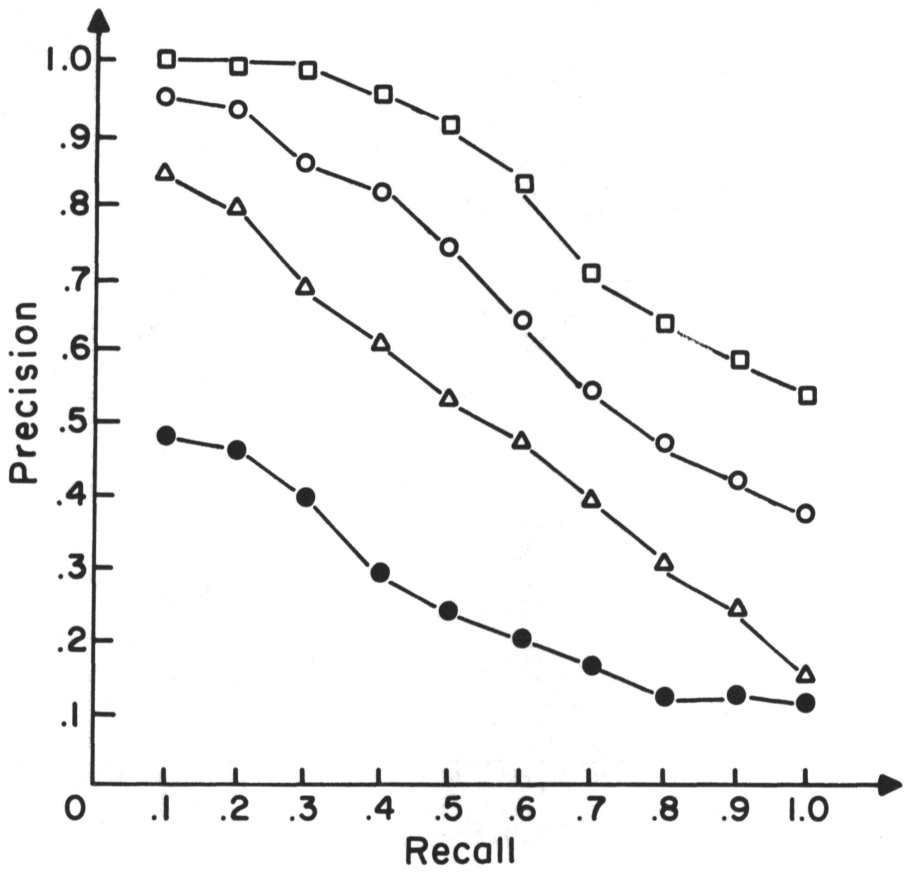


$\alpha_1 = \alpha_2 = 1$, $\delta = 3$, modified cosine

Document Space Modification

ADI Collection - Averages over 11 Requests

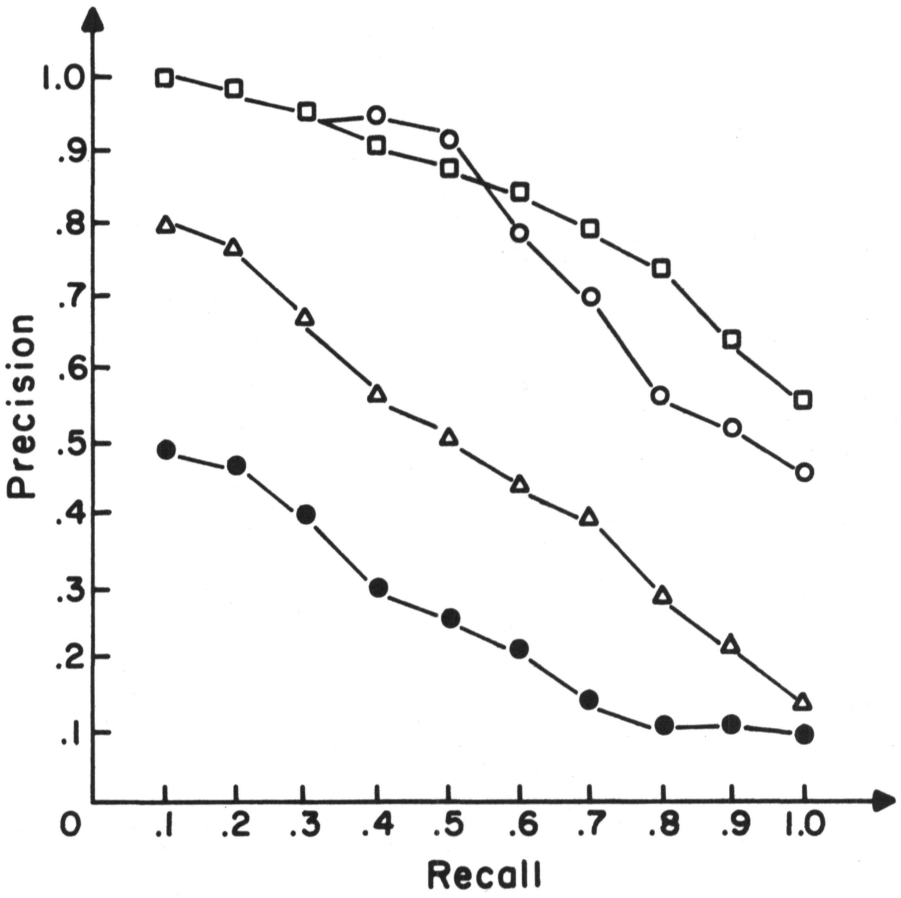
Fig. 1



$\alpha_1 = \alpha_2 = 1$, $\delta = 3$, modified cosine

Document Space Modification

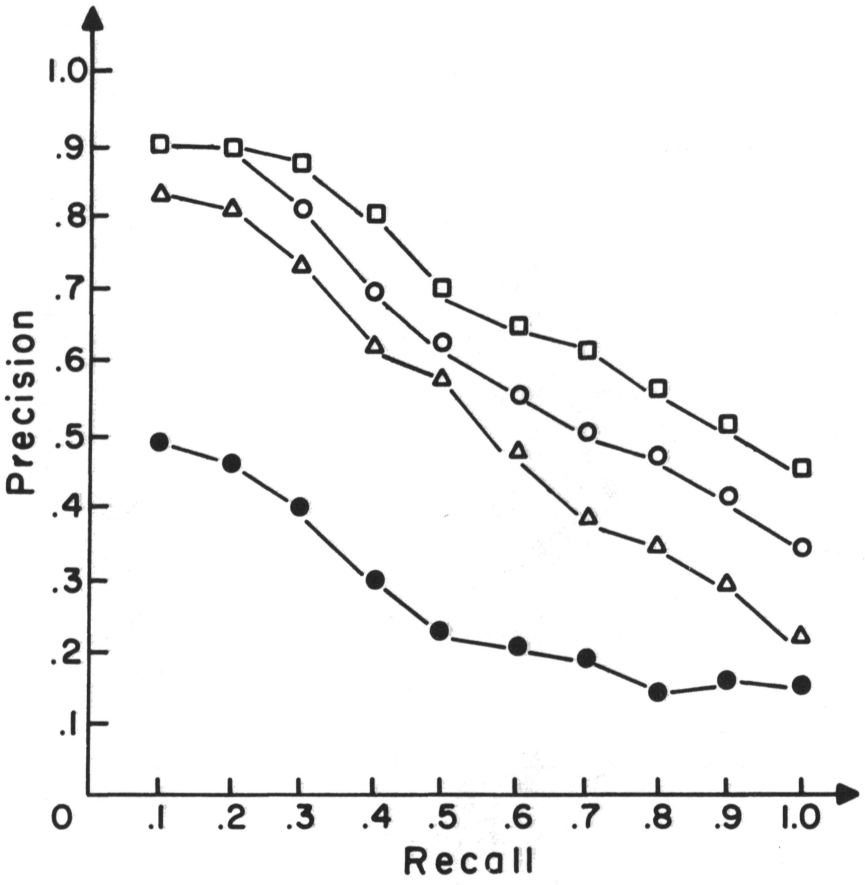
Fig. 2



$\alpha_1 = \alpha_2 = 1$, $\delta = 11$, modified cosine

Document Space Modification

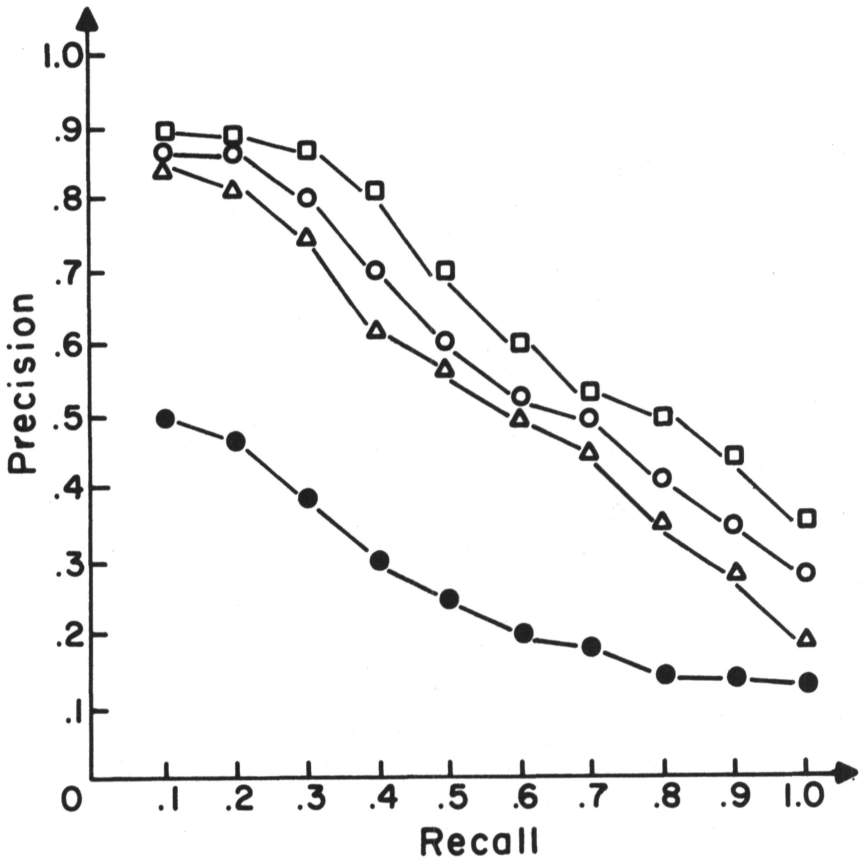
Fig. 3



$\alpha_1 = \alpha_2 = 1$, $\delta = 3$, standard cosine

Document Space Modification

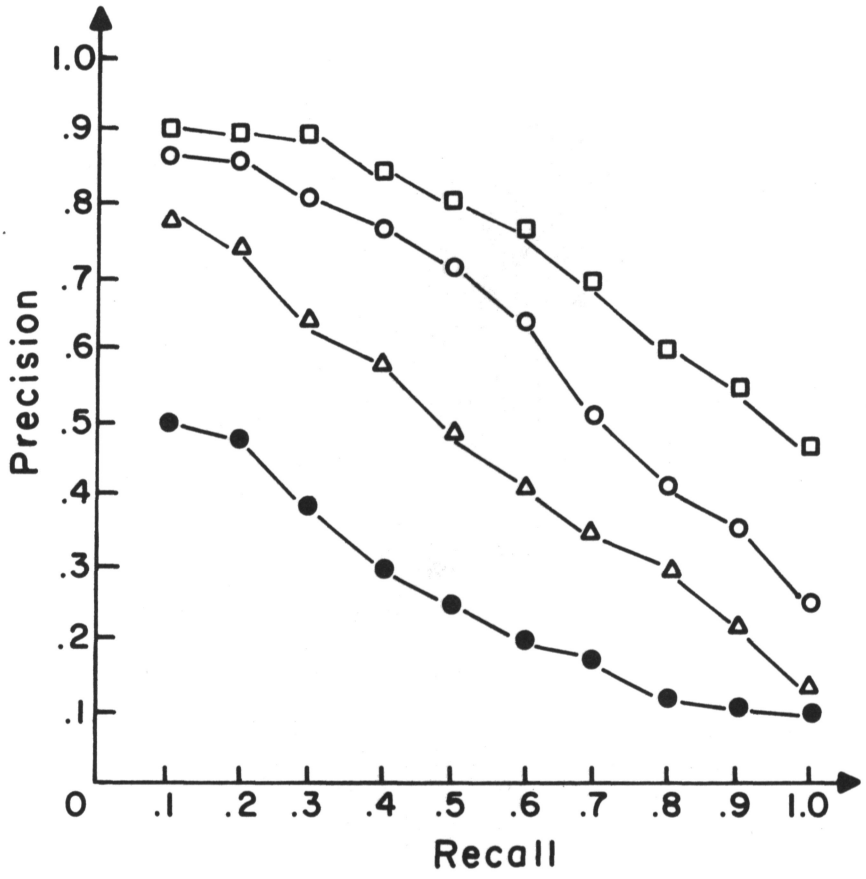
Fig. 4



$\alpha_1 = \alpha_2 = 1$, $\delta = 5$, standard cosine

Document Space Modification

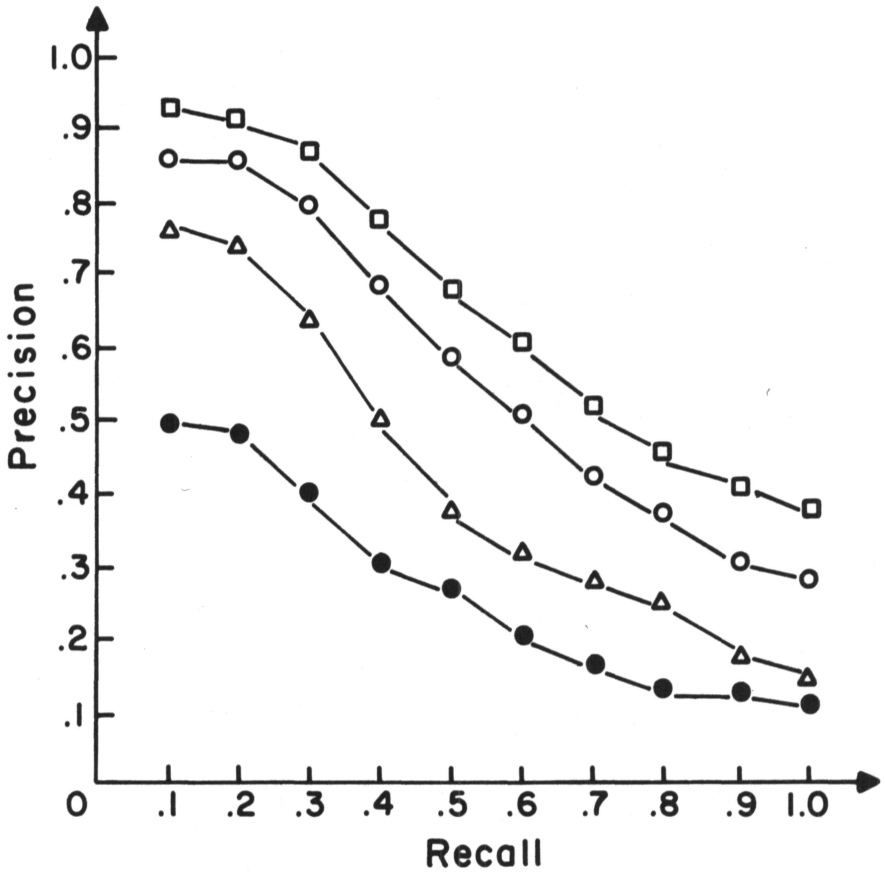
Fig. 5



$\alpha_1 = \alpha_2 = 1$, $\delta = 11$, standard cosine

Document Space Modification

Fig. 6



Typical results using query modification
ADI Collection - 11 requests

Fig. 7

Figs. 4 to 6 give average recall-precision curves achieved by using the standard cosine correlation function. Comparison with Figs. 1 to 3 shows that the modified cosine correlation function yields much better results. The modified cosine correlation function causes the relevant documents to converge faster to the query. In fact, as can be seen from the graphs, modified cosine achieves, on the average, better recall and precision after 2 iterations than standard cosine achieves after 3.

Experiments performed on the effect of varying α_1 and α_2 are quite disappointing. The results are extremely erratic and do not seem to yield any general conclusions about the effects of α_1 and α_2. In particular, the use of different values of α_1 and α_2 for each iteration causes undesirable instabilities such as relevant documents being very nearly retrieved on the second iteration and then being moved far from the query on the third iteration. The results for α_1 and α_2 equal to 1 are the best ones obtained in this investigation.

4. Conclusions

The effectiveness of relevance feedback is enhanced by the use of a dynamic document space. Such a space enables relevant documents to be brought close to each other and close to the query. Moreover, this study exhibits a relevance feedback algorithm based on document modification that consistently achieves better recall and precision than the algorithm based on query alteration suggested by Salton and Rocchio, and implemented by Riddle, Horwitz, and Dietz.

These conclusions suggest further areas for investigation. First, there is a definite need for a theoretical study of document modification. The present study presents a particular algorithm that works well. However, this algorithm is not necessarily optimal. Theoretical investigation might uncover such an algorithm.

Second, a study should be made of the efficiency of document modification. It is clearly a more time consuming method than query modification, since changes must be made to the entire document collection as opposed to being made only on the query. However, further investigation might show that document modification can be implemented without too great a loss in efficiency over query modification. In this case, the improved results achieved by document modification would justify the extra expenditure of computing time.

Third, an investigation should be made into the possibility of using both query and document modification. The results achieved in this study might be further improved by such a hybrid technique.

References

[1] E. M. Keen, "Evaluation of Relevance Feedback Methods", Report ISR-11 to the National Science Foundation, Cornell University, June 1966.

[2] T. Horwitz, R. Dietz, O. W. Riddle, "Relevance Feedback in an Information Retrieval System", Report ISR-11 to the National Science Foundation, June 1966.

[3] J. J. Rocchio, "Harvard Doctoral Thesis", Report ISR-10 to the National Science Foundation, Harvard Computation Laboratory, June 1965.

[4] G. Salton, J. J. Rocchio, "Information Search Optimization and Iterative Retrieval Techniques", AFIPS Conference Proceedings, Vol. 27, Part 1, FJCC, 1965.

[5] G. Salton, Automatic Information Retrieval, Class Notes, Computer Science 435, Cornell University.