XII. The Evaluation Problem in Relevance Feedback Systems

by

Harold A. Hall and Nelson H. Weiderman

Abstract

The problem of evaluating relevance feedback systems is discussed in the context of the operating SMART document retrieval system. A ranking effect and a feedback effect are defined and are related to the changes that occur in precision and recall during successive feedback iterations. An evaluation scheme which isolates the feedback effect is presented.

1. Introduction

The SMART document retrieval system is a user-oriented system in which a list of documents is presented to the user in response to a natural language query.[1] In any such system the practical assumption is made that the user will be unable on his first try to formulate a query which is optimal for his particular needs. In light of this, a document retrieval system must take advantage of computer-user interaction in order to reformulate the original query to obtain more desirable results. In an operating system the specific search strategy is initially unspecified so that it may be adapted to the needs of the user. [1]

One strategy is to use relevance feedback. [2,3,4] That is, the user is presented with a number of documents and is asked to specify those which are pertinent to his interests. The terms in the documents judged relevant are then added to the terms of his original query to move the query in the document

[1] For detailed information on the SMART document retrieval system, see [1]

space to an area where additional relevant documents are present. This iterative procedure may be continued until the type and amount of information desired is retrieved. The assumption is made that documents of a similar nature lie in close proximity to one another in the document space. [2]

Although the user's job is simply to make relevance judgments, a relevance feedback system may be varied by changing many parameters, e.g. the number of documents fed back, the relevance weighting factors used, the use of negative feedback (deletion of concepts appearing in documents not deemed relevant), etc. It is the purpose of this study to examine the criteria with which a relevance feedback system (using a given set of parameters) may be evaluated.

2. The Present Evaluation System

A "perfect" document retrieval system would retrieve all those documents which are relevant and none which are not relevant. To measure the extent to which a system approaches this ideal situation, the following two measures have been devised:

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

$$\text{Precision} = \frac{\text{Number of relevant retrieved}}{\text{Total number retrieved}} \; : [5]$$

Using a given ranking of documents together with their relevance judgments, one may calculate recall and precision after each n documents retrieved. These points are plotted on a graph with recall on the abscissa and precision on the ordinate as in Fig. 1. For a given document ranking the precision and recall remain at zero until a relevant document is found, at which time recall becomes 1/r and precision becomes 1/n, where r is the total number of

Rank	Relevant	Recall	Precision
1		0.0	.0000
2	X	0.2	.5000
3		0.2	.3333
4.		0.2	.2500
5	X	0.4	.4000
6		0.4	.3333
7		0.4	.2857
8	X	0.6	.3750
9	X	0.8	.4444
10		0.8	.4000
11		0.8	.3636
12		0.8	.3333
13		0.8	.3076
14		0.8	.2857
15	X	1.0	.3333
16		1.0	.3125
17		1.0	.2941
18		1.0	.2777
19		1.0	.2684
20		1.0	.2500

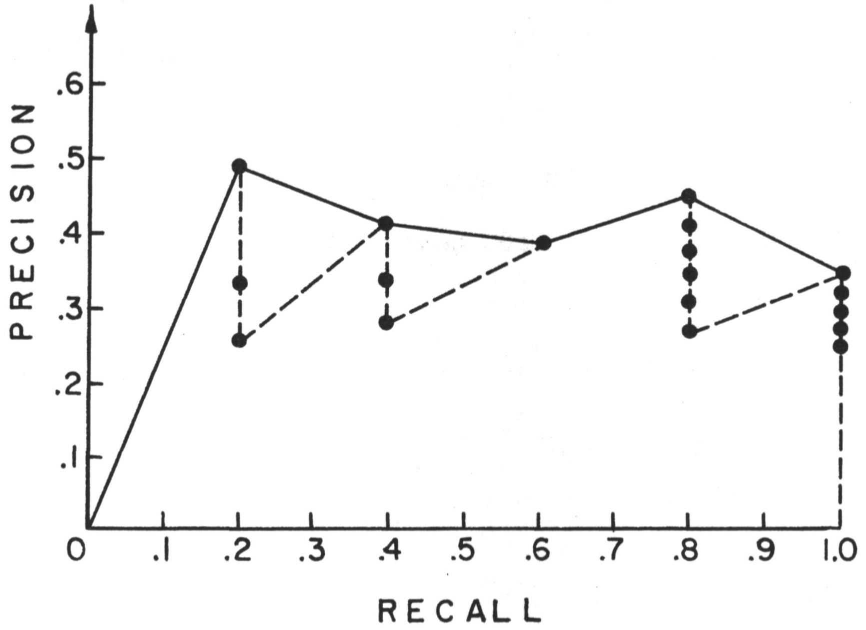


Fig. 1: Peak-Interpolated Recall-Precision Graph

relevant documents in the collection and n is the number of documents retrieved. Recall now remains constant while precision decreases until another relevant document is found, at which time the recall jumps to 2/r and precision becomes 2/n', and so forth. The result is a set of points arranged on r vertical lines equally spaced along the abscissa.

A smooth curve is obtained by interpolating between the r points which represent the relevant documents (the peaks of the graph if the points are joined), and the point (0,1) if the first document is not relevant or the point (0,0) if the first document is not relevant. Graphs of this type are generally averaged over a number of queries at twenty points of recall at intervals of 0.05.

Two points should be made about these graphs. First, only r+1 points are used in their construction. This may cause serious difficulties when the number of relevant documents is small. For example, if there is only one relevant document, it is meaningless to ask what the precision is at any value of recall other than zero or one; hence at all other points of recall, the value for precision is interpolated. Second, it should be noted that graphs constructed in this manner do not represent absolute recall and precision. In many cases this fact causes little or no difficulty for evaluation, since only relative differences between various plots are being considered. It is restrictive only in cases in which one would like to know what the precision and recall are after a certain number of documents have been retrieved. This particular information is not available from these graphs.

In computing standard recall and standard precision it is necessary to pick a cut-off value to distinguish between documents retrieved and documents not retrieved. Since it is desirable to use a measure of retrieval efficiency which is independent of the distinction between retrieved and nonretrieved documents, the present system uses recall-like and precision-like measures, termed normalized precision and normalized recall, which take values between zero and one, and can be interpreted as approximations to average standard recall and average standard precision for large document collections. Specifically, if R_j is standard recall and P_j is standard precision after j documents retrieved, we have

$$R_{norm} \sim \frac{1}{N} \sum_{j=1}^{N} R_j \quad \text{and} \quad P_{norm} \sim \frac{1}{N} \sum_{j=1}^{N} P_j$$

where N is the total number of documents in the collection. [6,7]

3. The Problem

In reappraising an evaluation technique the basis for the reappraisal consists in a thorough examination of the problem being considered--in this case, relevance feedback. The aim of a document retrieval system is to maximize the number of relevant documents retrieved while at the same time minimizing the number of non-relevant documents retrieved; i.e., to maximize recall and precision. Therefore, the aim of a relevance feedback system should be to maximize the number of relevant documents found using relevance feedback <u>over and above</u> the number of relevant documents found without such a system. In other words, for a given number of documents retrieved using relevance feedback one wishes to retrieve more relevant documents than would have been retrieved had the user looked at the same number of documents after an initial ranking.

To this end it is the job of an evaluation scheme to compute precision and recall for the initial search and the feedback iterations in such a way that meaningful comparisons can be made between them. In particular, one must distinguish between two phenomena occurring simultaneously to change precision and recall. First, there is the relative change in recall and precision due to a shift in document rankings from one iteration to the next, herein called the ranking effect. Second, there is the relative change in recall and precision which occurs when new relevant documents are found in successive iterations, herein called the feedback effect. Unfortunately, there is no way of distinguishing these two effects using the peak-interpolated precision-recall curves. In fact, the ranking effect completely swamps the feedback effect.

Evidence in support of this assertion is available in the data collected by Messrs. Riddle, Horwitz, and Dietz in a prototype study of relevance feedback. [4] After each query perturbation, documents were ranked in correlation order without regard to whether any given document was retrieved on a previous iteration or not. Fifteen documents were retrieved on each iteration. For the particular set of parameters--namely increasing alpha strategy with cosine correlation--which were judged optimal in the study using an initial search and three feedback iterations, it was found that of thirty-four queries tested, eleven of them retrieved all of the relevant documents on the initial search, twelve others retrieved no new relevant documents after the initial search, and for only eleven were new relevant documents retrieved as a result of relevance feedback. From the same data it was found that relevant documents retrieved on the initial and successive searches increased an average of 5.5 positions in the rankings by the third iteration. On the other hand, for the same thirty-four queries, only sixteen new relevant documents, out of ninety-two relevant documents retrieved, were retrieved as a result of relevance feedback.

Before embarking on a plan for an alternate evaluation technique, the distinction between the ranking effect and the feedback effect and the reasons for making any distinction at all should be made clear. The ranking effect is due to a change in the ranks of relevant documents already retrieved. Since the retrieved documents are the very documents used to perturb the query in the first place, their correlations with the perturbed query will undoubtedly increase for successive iteration, and these documents will tend to move up in ranking from one iteration to the next. Assume, in order to isolate the ranking effect, that no new documents are retrieved for successive iterations, and that all relevant documents are weighted equally in each iteration. Then, in the limit, the iterated query will approach the centroid of these relevant documents, since the original query, unless weighted each time, would become negligible in its effect. The ranking effect is of significant interest in the study of Rocchio's notion of an optimum query and its generation by a request optimization algorithm. [2] Hence the ranking effect measures the extent to which a query has been perturbed toward the centroid of the set of previously retrieved documents.

To isolate the feedback effect, one must hold rankings of previously retrieved relevant documents constant while retrieving documents not previously evaluated by the user. Hence the feedback effect measures the effectiveness of the perturbed query in bringing new relevant documents to the attention of the user.

The fact that these two effects should be separated is clear. Certainly one should distinguish between the case in which successive iterations do present new relevant documents to the user, and the case in which they do not. It is therefore obvious that determination of the feedback effect is essential in the evaluation of a general purpose relevance feedback system for document retrieval. For this reason, an evaluation scheme which isolates the feedback effect from the more prominent ranking effect is presented.

4. The Solution

Although the retrieval system is not so restricted, assume for purposes of illustration that the number of documents retrieved at any one time is equal to some constant k. The proposed system then behaves as follows:

The user submits his request. In general, the system matches the request against a subset of the document collection and ranks the documents in correlation order. It returns to the user those k documents which correlate most highly with the query. The user now examines the retrieved documents and assesses their relevance to his request. Using these relevance judgments, the system generates a new query and then ranks the documents with respect to this new query. The system now presents to the user those k most highly ranked documents which the user has not already seen. Thus, after again making his relevance judgments, the user will have considered 2k different documents. The number of iterations allowed is a parameter of the system.

For evaluation purposes, the documents retrieved on the initial search are ranked 1 to k throughout, those retrieved on the first iteration are ranked k+1 to 2k for each iteration beginning with the first, and so forth. Thus a set of rankings might appear as in Fig. 2. A stylized recall-precision graph for this system appears in Fig. 3 in which recall and precision are plotted after k, 2k, 3k,..., nk documents, where n is the total number of searches.

Consider the following formalism. Define R_{ij} to be the recall on the ith iteration after j documents retrieved. Similarly P_{ij} is the precision on the ith iteration after j documents retrieved. The objective of a relevance feedback system as stated earlier is now the maximization of the differences, after j documents retrieved, between the number of relevant documents retrieved on the ith iteration and the number retrieved on the initial search. In terms of recall and precision the objective is to find

$$\max_{\text{s.p.c.}} (R_{ij} - R_{0j}) \quad \text{and} \quad \max_{\text{s.p.c.}} (P_{ij} - P_{0j})$$

for each j as i ranges from 1 to n , where n is the number of iterations. R_{0j} and P_{0j} are recall and precision, respectively, after j documents retrieved on the initial search; and the abbreviation s.p.c. indicates that the maximums are to be taken over all system parameter changes. If R is the set of all recall differences, a graph of the points of R might appear as in Fig. 4. (Points are joined to produce continuous curves.) A similar graph is constructed in Fig. 5 for P, the set of precision differences. The objective is therefore, in terms of the graphs, to maximize distances between the R-j and P-j curves.

In terms of Fig. 3, the differences between numbers of relevant documents found are taken along the dashed lines, with vertical displacement indicating increased precision and displacement to the right indicating increased recall. It should be noted that although it is hoped that both recall and precision will increase from one iteration to the next, it is possible for one or both to decrease.

Since new documents are retrieved each time, rather than old documents in new order, the ranking effect is eliminated, and all the improvement between curves on the graphs of Figs. 3, 4, and 5 is due to relevance feedback. Normalized recall and normalized precision are computed and based on the rankings indicated above.

5. Preliminary Results

The evaluation scheme was tested on two document collections, the eighty-two document American Documentation Institute (ADI) collection derived from abstracts of papers on information retrieval, and the two hundred document Cranfield collection derived from abstracts of papers on aerodynamics. Identical input parameters were used for the two collections; that is, the user is presented five new documents each time and on each of three iterations the weights of the concepts in the original query and the weights of concepts in newly retrieved relevant documents were added to the query of the previous iteration.

For this particular choice of parameters, the results show a significant difference in relevance feedback performance between the ADI and the Cranfield collections. This difference is illustrated in the R-j and P-j

Rank	Initial Search		1st Iter.		2nd Iter.		3rd Iter.	
	Doc. No.	Rel.	Doc.	Rel.	Doc.	Rel.	Doc.	Rel.
1	4		4		4		4	
2	8	X	8	X	8	X	8	X
3	16	X	16	X	16	X	16	X
4	37		9	X	9	X	9	X
5	1		27		27		27	
6	22		22		22		22	
7	3		1		55	X	55	X
8	2		2		14		14	
9	9	X	3		3		3	
10	86		14		13		13	
11	42		13		1		74	X
12	13		37		74	X	1	

Fig. 2

Hypothetical Document Rankings for k=3.
(Cross-hatched areas designate documents retrieved.)

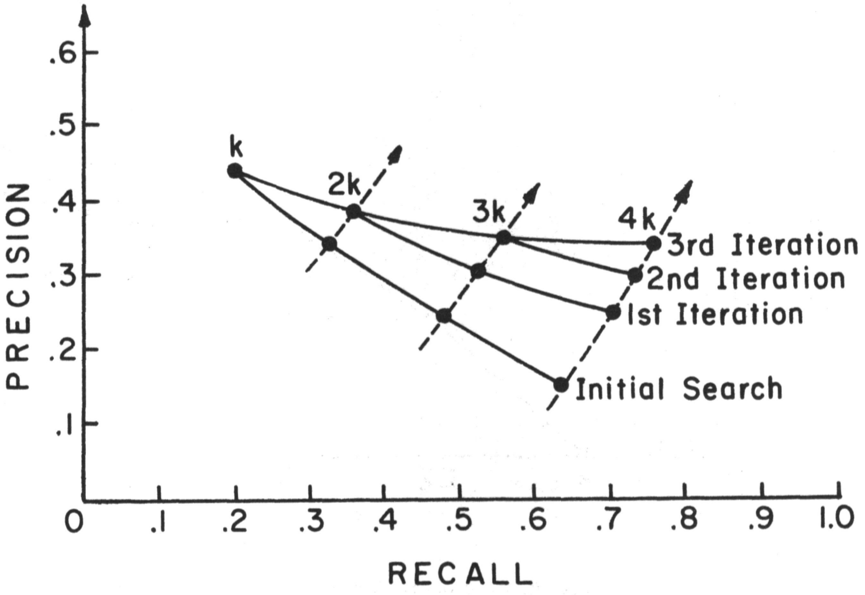
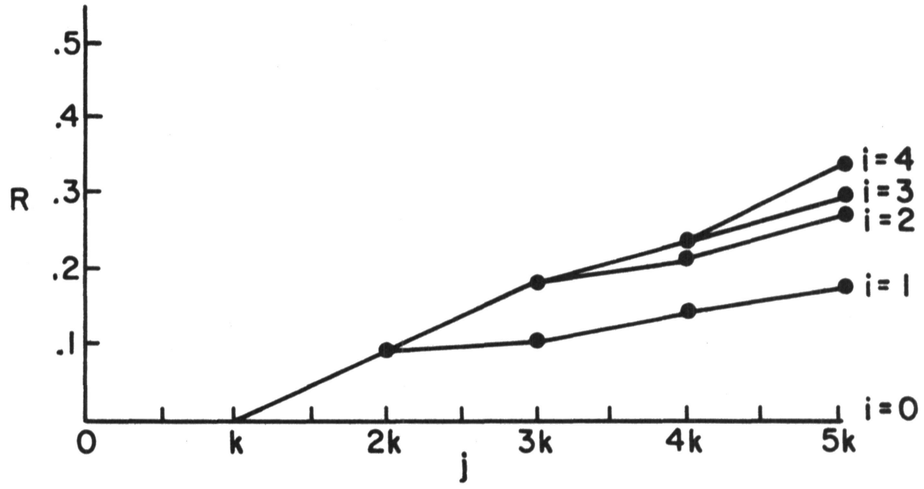
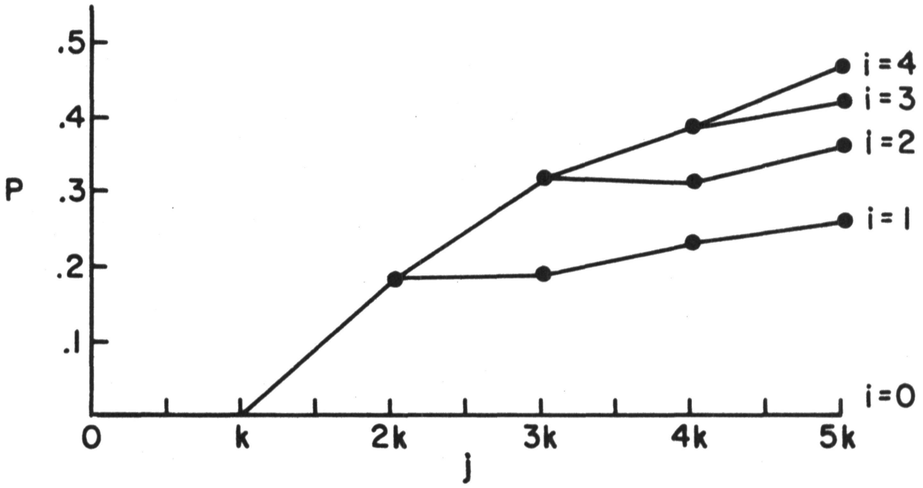


Fig. 3: Stylized Recall-Precision Graph



Stylized R-j Graph of Recall Differences

Fig. 4



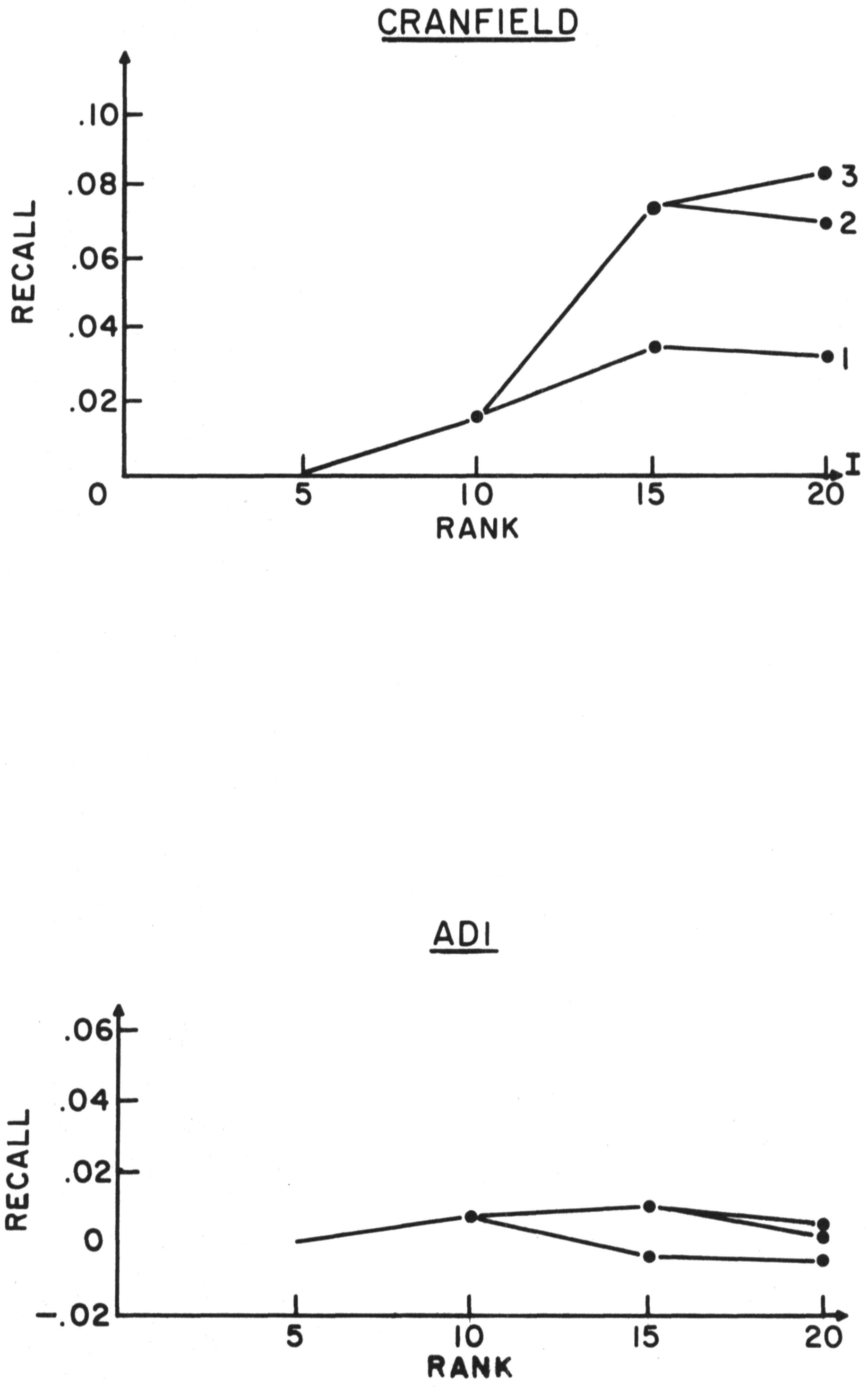
Stylized P-j Graph of Precision Differences

Fig. 5

curves of Figs. 6 and 7. In the Cranfield collection, relevance feedback accounts for an increase in standard recall, averaged over forty-two queries, of .0857 at a retrieval level of twenty documents. The corresponding increase for the ADI collection, averaged over 35 queries, is only .0053. In the Cranfield collection, average precision increased by .0274 at a retrieval level of twenty documents, while in the ADI collection, it increases by only .0014. Thus, for this particular set of parameters, the Cranfield collection is ten to twenty times more effective in increasing precision and recall due to relevance feedback.

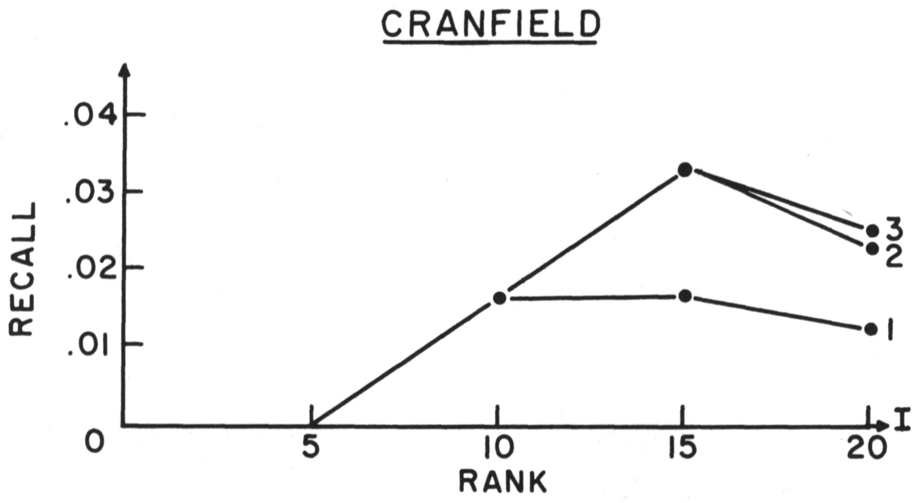
Several reasons may be suggested for this difference in performance. Since the document representations of the two collections were constructed by the same technique (document vectors consisting of thesaurus concept classes), it appears that the difference in performance is a result of the document collections themselves. In particular, the Cranfield collection has, in general, both longer queries and longer documents than the ADI collection. This may indicate that relevance feedback performance is, in part, a function of how detailed are the representations of documents and queries with respect to the actual documents and queries. The difference in performance may also be due to a closer clustering of relevant documents in the Cranfield collection than in the ADI collection.

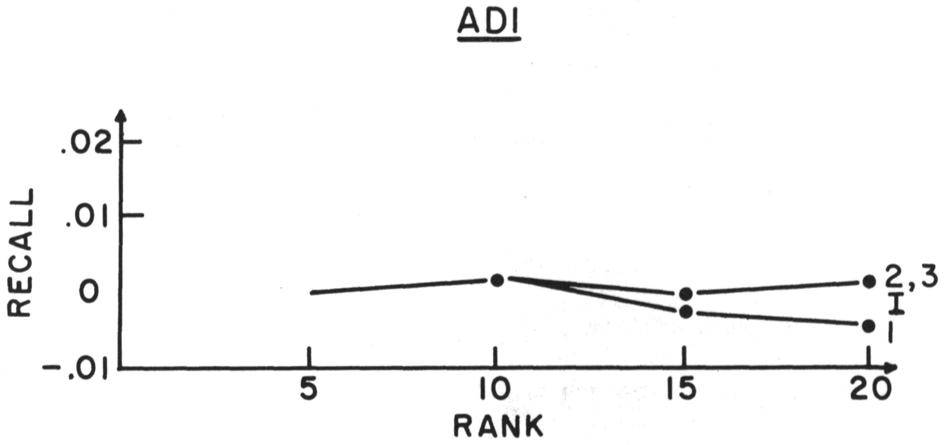
Differences in document collections and the determination of a set of parameters which yield optimum retrieval using relevance feedback are areas for further investigation, but are not the object of this study. The applications of the evaluation scheme are presented merely to illustrate its effectiveness in distinguishing the feedback and ranking effects. Pre-



Increase in recall due to the feedback effect in the Cranfield and ADI Collections

Fig. 6





Increase in precision due to the feedback effect in the Cranfield and ADI Collections

Fig. 7

vious results have indicated that the combined ranking and feedback effects produce significant increases in recall and precision. [6] The results of Figs. 6 and 7 indicate that the feedback effect was negligible in producing increased recall and precision in the ADI collection, whereas in the Cranfield collection, the feedback effect is definitely evident.

6. Conclusion

In some iterative or comparative procedures in document retrieval--such as searching for an optimum query or comparing various correlation coefficients--the measure of effectiveness of the procedures is the extent to which the rank of the documents increases.

The point of view taken herein is that for a relevance feedback system the measure of its effectiveness should be a measure of how many new relevant documents are retrieved as a result of feedback.

From examination of results previously obtained in relevance feedback, it is clear that the present evaluation scheme does not include this important measure. The measures of precision and recall do not separate the effect of increasing the ranking of documents from the effect of retrieving new relevant documents.

The evaluation scheme presented herein has several advantages. First, it isolates the feedback effect from the more prominent ranking effect by keeping retrieved document rankings constant. As a consequence, one is forced to use the convention that once a document is retrieved, it remains retrieved in the evaluation. This is consistent with the actual environment in which such a system would be used. Second, the system offers the advantage of presenting absolute measures of precision and recall, rather than possibly unreliable interpolated values. That is, instead of using r+1 points (where r is the number, usually a relatively small one, of documents relevant to a particular query) to plot a graph of precision and recall as is presently done, it is possible to use a large number of points by computing precision and recall at each level of retrieval. This technique enables greater discrimination between similar cases. Third, the system clearly illustrated the precision-recall trade-off by separating these two measures as in Figs. 4 and 5. Finally, the evaluation permits comparisons of precision and recall at a fixed number of documents retrieved.

References

[1] G. Salton and M. E. Lesk, "The SMART Automatic Document Retrieval System--An Illustration," Communications of the ACM, Vol. VIII, No. 6, June, 1964.

[2] J. J. Rocchio, Harvard University Doctoral Thesis, Report ISR-10 to the National Science Foundation, Chapters 3, 5.

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

[4] W. Riddle, T. Horwitz, and R. Dietz, "Relevance Feedback in an Information Retrieval System," Report ISR-11 to NSF, Chapter 6.

[5] C. W. Cleverdon, "The Testing of Index Language Devices," ASLIB Proceedings, Vol. XV, No. 4, April, 1963.

[6] G. Salton, "The Evaluation of Automatic Retrieval Procedures--Selected Test Results Using the SMART System," American Documentation, Vol. XVI, No. 3, July, 1965.

[7] G. Salton, Automatic Information Retrieval, unpublished manuscript (class notes), Chapters 1, 9.