ISR-9
August, 1965

XXIII. RELEVANCE FEEDBACK IN INFORMATION RETRIEVAL

J. J. Rocchio

SUMMARY

In evaluating the performance of a document retrieval system one must attempt to isolate the critical variables which determine system behavior. For this purpose a model is introduced which identifies indexing, search request formulation, and request-document matching as the three primary functions of an automatic retrieval system. Search request formulation, the responsibility of the users of the system, is considered to be the variable with the greatest potential variance. In view of this, the idea of the optimization of search requests is believed to constitute a primary possibility toward better control in evaluating indexing and request-document matching. Investigation into request optimality leads in turn to a method for improving search requests in an operational framework. For this purpose the notion of interaction between a user and an information retrieval system by means of relevance feedback is introduced. A process of request modifications is developed based on a sequence of retrieval operations, such that after each operation the user is allowed to communicate his evaluation to the system. This information is used as a basis for altering the user's query. The modification algorithm is developed below and some preliminary results are presented.

1. System Model

A document retrieval system is assumed in which the index transformation consists of mapping documents and requests from the natural language to multi-dimensional property vectors in an abstract space. The model is based on the SMART retrieval system, in which a text is represented as a numerical vector in a concept space. The index transformation is effected via a thesaurus mapping from word stems to concepts which may be augmented by several other text processing techniques, for example, phrase identification and hierarchical transformations. The end product of the index transformation is however a vector in the concept space in which the weight of concept i (that is, the ith component of the vector) is indicative of the frequency of this concept in the original document.

The retrieval or searching process consists in matching a request image against the set of document images which constitute the store. Various matching strategies are possible; we will assume here that the information contained in the index image of a test is represented by the orientation of the concept vector in the concept space (that is, the matching process is assumed independent of the magnitude of the concept vector). Retrieval therefore consists in locating the set of documents having an orientation (angular position) similar to that of the request image.

2. Request Formulation

The process of request formulation is a complex one and depends on particular attributes of the requestor, such as his knowledge of the contents of the store, his knowledge of the indexing and searching processes of the

system, his familiarity with the topic matter being searched, his personal preferences as to vocabulary and style, etc. In effect, the user must make a statistical decision based on his personal experience as to what request is most likely to produce results useful to him. Clearly the à priori likelihood of a request satisfying the user's needs varies over a wide range for any given retrieval system. For example, a user who needs to know whether a particular document is contained in the store can formulate a request which will satisfy this need with perfect certainty, for example, by submitting a request identical with the document in question. At the other end of the spectrum is the user who needs information on a topic unfamiliar to him. Clearly the probability of his being able to formulate a request which will retrieve the best set of documents satisfying his information needs is very small.

Operationally then, information requests submitted to a retrieval system vary over a wide spectrum of à priori likelihood of satisfying the user's needs. It is then pertinent to consider techniques of reducing this variance in two distinct contexts. First, in an operational sense one would like to process requests which are optimized with respect to the cost of retrieval, the cost of optimization and the value of the information to the user. Second, in the evaluation of information retrieval systems, it is desirable to isolate the effect of the indexing process on retrieval performance from the effects due to request formulation. Thus the gross results of retrieval experiments carried out with a sample set of requests can be used to evaluate the indexing discipline with respect to that particular sample of requests. If it were possible, however, to define an optimal

request corresponding to any given request (for some fixed index transformation), retrieval results based on optimal requests would provide a much clearer evaluation of the power of the indexing technique, since performance variations due to request malformation would be eliminated.

3. Request Optimization

To define an optimal request, it is necessary to operate with an explicit formulation of the request-document matching process. In accordance with the system model outlined above, an appropriate matching function is the cosine correlation of the request image and the set of document images in the store. The cosine correlation of two vectors is defined by

$$\rho(\bar{A},\bar{B}) = \frac{\bar{A}\cdot\bar{B}}{|\bar{A}|\,|\bar{B}|} = \cos(\bar{A},\bar{B}) \ .$$

Since the vector images produced by the index transformation are limited to having nonnegative concept weights, this correlation induces a ranking on elements of the store, equivalent to their angular distance from the request vector; (that is, $0 \leq \rho \leq 1$ corresponds to an angular separation of from 90 to 0 degrees).

Corresponding to any retrieval request $\bar{Q}$, we assume the existence of a subset D_R, $(D_R \subset D)$, of the set of documents contained in the store D. This set is the set of documents relevant to the request $\bar{Q}$ and must be specified outside the context of the retrieval system.

Having defined D_R, an <u>ideal</u> request may be defined as one which induces a ranking on the elements of D such that all members of D_R are ranked higher (that is, have a higher correlation) than all other elements of D.

Since relevance is a subjective attribute of a given request-document set pair, determined in theory by the individual requestor, there is no certainty that an ideal request (as defined above) in fact exists for a given request. In such a case one might say that the indexing is defined from the point of view of the particular user, since it does not allow distinctions equivalent to those he can make. This of course will be the norm rather than the exception, since the indexing process is designed to reduce rather than preserve information. For this reason an unambiguous, optimal request is defined as a function of D_R, D, and the index transformation, which is unique for every nonempty unique subset D_R of D.

An optimal request corresponding to a given subset D_R of a store D, under an index transformation T, is that request which maximizes the difference between the mean of the correlations of the relevant documents (members of D_R) and the mean of the correlations of the nonrelevant documents (members of D not in D_R).

In mathematical terms the optimal request vector $\bar{Q}_0$ corresponding to a set $D_R \subset D$ is defined as that vector $\bar{Q}$ for which

$$C = \frac{1}{n_0} \sum_{D_i \in D_R} \rho(\bar{Q},\bar{D}_i) - \frac{1}{n-n_0} \sum_{D_i \notin D_R} \rho(\bar{Q},\bar{D}_i)$$

is maximum, where $n_0 = n(D_R)$ the number of elements in D_R, and $n = n(D)$ the total number of elements in the store.

If we wish to consider only requests having nonnegative components (this corresponds to the assumption originally made about index images in the system under consideration), then the problem is modified to maximizing C subject to $\bar{Q}_i \geq 0$.

Substituting for $p(\bar{Q},\bar{D}_i)$, and using vector notation results in

$$C = \frac{1}{n_0}\left(\frac{\bar{Q}}{|\bar{Q}|}\right) \cdot \sum_{\bar{D}_i \in D_R} \frac{\bar{D}_i}{|\bar{D}_i|} - \frac{1}{n-n_0}\left(\frac{\bar{Q}}{|\bar{Q}|}\right) \cdot \sum_{\bar{D}_i \notin D_R} \frac{\bar{D}_i}{|\bar{D}_i|},$$

or

$$C = \frac{\bar{Q}}{|\bar{Q}|} \cdot \left[\frac{1}{n_0} \sum_{\bar{D}_i \in D_R} \frac{\bar{D}_i}{|\bar{D}_i|} - \frac{1}{n-n_0} \sum_{\bar{D}_i \notin D_R} \frac{\bar{D}_i}{|\bar{D}_i|}\right].$$

Since this is equivalent to $C = \bar{Q}^* \cdot \bar{A}$, where $\bar{Q}^*$ is a unit vector, clearly $Q_{opt} = k\bar{A}$ (k being an arbitrary scalar), or

$$\bar{Q}_{opt} = \frac{1}{n_0} \sum_{D_i \in D_R} \frac{\bar{D}_i}{|\bar{D}_i|} - \frac{1}{n-n_0} \sum_{D_i \notin D_R} \frac{\bar{D}_i}{|\bar{D}_i|}.$$

Further, a simple proof shows that C is maximized subject to $\bar{Q}_i \geq 0$ (where i ranges over all coordinates of $\bar{Q}$) for the vector

$$\bar{Q}'_{opt_i} = \begin{cases} \bar{Q}_{opt_i} & \text{for } Q_{opt_i} \geq 0 \\ 0 & \text{for } Q_{opt_i} < 0 \end{cases}.$$

Hence, under the assumptions made, an unambiguous optimal (for the criterion stated) query exists corresponding to any nonempty subset D_R of D. In evaluating the effectiveness of automatic information retrieval systems, this formulation of an optimal request provides the ability to isolate the

effects of indexing from variances due to request formulation. In effect, an optimal request measures the ability of the indexing transformation to differentiate a particular set of documents from all the others in the store, where the particular set in question is assumed to have some intrinsic association, specified independently of the system, that is, the set consists of the documents judged to be relevant to some particular topic.

4. Relevance Feedback

The formulation of the optimal query corresponding to a particular set of documents has no direct implication on operational information retrieval, since the set of documents in question is the object of the retrieval search. Thus there is no à priori way to make an optimal request, since having the ability to do so would eliminate the need for retrieval. This kind of circularity suggests a strong analogy to feedback control theory. Thus if we consider a sequence of retrieval operations starting with an initial query $\bar{Q}_0$, which is then modified on the basis of the output produced by the retrieval system, using $\bar{Q}_0$ as input, in such a way that the modified query $\bar{Q}_1$ is closer to the optimal query for this user, a precise analogy to a sequential feedback system can be drawn. Let the user specify which of the retrieved documents (resulting from the search using $\bar{Q}_0$) are relevant and which are not. This information constitutes an error signal to the retrieval system. On the basis of the error and the original input, it is then possible to produce a modified query (new command input) such that the retrieval output will be closer to what the user desires, or such

that the modified query will in effect be closer to the optimal query for this user's needs. The effectiveness of this process will depend on how good the initial query is, and on how fast the process of iteration converges to the optimal request.

On the basis of the formulation of request optimality, we then seek a procedure for using the relevance feedback from an initial retrieval operation to produce an improved query. Let $\bar{Q}_0$ be the original retrieval request, and let the results of the retrieval operation be a list in correlation order of the documents whose images are most closely related to $\bar{Q}_0$. The user examines this list and specifies which of the documents in it are relevant and which are not. Since the modification is to be based only on a sample of the relevant documents (assuming that some are missing from the retrieved list associated with $\bar{Q}_0$), the modified request will be formed by adding to the original query $\bar{Q}_0$ an optimal query vector based on the feedback information. The resultant vector (the new query) should thus be a better approximation to the optimal query than $\bar{Q}_0$, and should therefore produce better retrieval when resubmitted.

Hence we seek a relation of the form

$$\bar{Q}_1 = f(\bar{Q}_0, R, S),$$

where $\bar{Q}_0$ is the original query, R is the subset of the retrieved set which the user deems relevant, and S is the subset of the retrieved set (based on Q_0) which the user deems nonrelevant. The form suggested immediately by the above is

$$\bar{Q}_1 = \bar{Q}_0 + \frac{1}{n_1}\sum_{i=1}^{n_1}\bar{R}_i - \frac{1}{n_2}\sum_{i=1}^{n_2}\bar{S}_i,$$

where $n_1 = n(R)$, $n_2 = (S)$, $R = \left\{\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_{n_1}\right\}$, $S = \left\{\bar{S}_1, \bar{S}_2, \ldots, \bar{S}_{n_2}\right\}$, and all vectors are unit vectors. Thus $\bar{Q}_1$ is the vector sum of the original query vector plus the optimal vector to differentiate the members of the set R from those of the set S. In other words, $\bar{Q}_1$ is the vector sum of $\bar{Q}_0$ plus the optimal vector for the subset of the store for which the user has provided relevance information.

The above equation may be rewritten in the form

$$\bar{Q}_1 = n_1 n_2 \bar{Q}_0 + n_2 \sum_{\bar{R}_i \in R} \bar{R}_i - n_1 \sum_{\bar{S}_i \in S} \bar{S}_i .$$

$\bar{Q}_1$ may be restricted to having nonnegative components by setting

$$\bar{Q}'_{1_i} = \begin{cases} \bar{Q}_{1_i} & \text{for } \bar{Q}_{1_i} \geq 0 \\ 0 & \text{for } \bar{Q}_{1_i} < 0 \end{cases} .$$

The above represents the basic relation for request modification using relevance feedback. This relation can be modified in various ways by imposing additional constraints. For example, the weighting of the original query ($\bar{Q}_0$) could be a function of the amount of feedback, such that with large amounts of feedback the original query has less effect on the

resultant than with small amounts of feedback. Another constraint, for example, might be to regulate the number of nonzero components of the modified query on the basis of the degree of overlap of a component among the relevant set which is feedback. Clearly there are a large number of variations to this basic relation which might be tried.

The modification process described is of course amenable to iteration and hence can be written in the general form

$$\bar{Q}_{i+1} = f(\bar{Q}_i, R_i, S_i),$$

where $\bar{Q}_i$ is the ith query of a sequence, and R_i and S_i are the relevant and nonrelevant subsets respectively, identified in response to retrieval with query $\bar{Q}_i$. It is expected that the rate of convergence of such a sequence to a near optimal query will be rapid enough to make the process economical; however, this will be investigated experimentally. In any case, the convergence rate can be estimated by the user, since it is reflected in the stability of the retrieved output.

The user's original query serves to identify a region in the index space which should contain relevant documents. Since he has no detailed knowledge about the characteristics of the document images in the store, it is unlikely that the vector image of his query is optimally located. By identifying relevant documents in the region, the user provides the system with sufficient information to attempt to produce a modified query which is positioned centrally with respect to the relevant documents, while maintaining maximum distance from the nonrelevant documents. This is possible, however, only insofar as the index images of the relevant set are differentiable from those of the nonrelevant set.

In a theoretical framework, the request optimization process focuses on the power of the index transformation to distinguish sets of associated documents within the store by eliminating variances due to particular query formulation. In an operating context, relevance feedback provides a technique whereby the system user can extract the full power of the index transformation to his retrieval problem, at the cost of iteration (possibly on a sample collection from a large store).

5. Initial Experimental Results

To test the effectiveness of the technique of relevance feedback as outlined above, some experiments have been conducted using the SMART retrieval system. A set of 17 requests which had been run through various processing options of the SMART system was used as a test sample. For each request the output resulting from a cosine correlation run using the SMART thesaurus (version 2) was examined. From the initial portion of the retrieved list (source documents having high correlations), two sets of documents were selected, one containing relevant documents and one containing nonrelevant documents. The vector images of the documents (given by the SMART thesaurus) together with the image of the request were used as input to a FORTRAN program which produced a modified vector suitable for input to SMART. The request modification algorithm used was as follows. The first step of the process applied the optimal request algorithm directly to produce

$$\bar{Q}_1 = n_1 n_2 \bar{Q}_0 + n_2 \sum_{i=1}^{n_1} \bar{R}_i - n_1 \sum_{i=1}^{n_2} \bar{S}_i ,$$

where $\bar{R}_i (i=1,n_1)$ constituted the relevant subset identified, and $\bar{S}_i (i=1,n_2)$ the nonrelevant subset, (the $\bar{R}_i$'s and $\bar{S}_i$'s being normalized vector images of the documents in question under the index transformation). Since the SMART system is designed to operate on document and query images having nonnegative components only, the vector $\bar{Q}_1$ was modified by setting

$$\bar{Q}'_{1_i} = \begin{cases} \bar{Q}_{1_i} & \text{for } \bar{Q}_{1_i} \geq 0 \\ 0 & \text{for } \bar{Q}_{1_i} < 0 \end{cases} .$$

A further modification was introduced to keep the modified query from becoming too specialized to the relevant set which was fed back. This modification was incorporated since relatively small amounts of feedback were to be tested.

The problem was therefore designed to allow a nonzero component in the resultant query image if and only if it occurred in $\bar{Q}_1$ and was either (a) in $\bar{Q}_0$ or (b) occurred in at least $n_1/2$ of the images of the relevant set, and in more relevant vectors than nonrelevant vectors.

The amount of feedback, that is, the number of relevant and nonrelevant documents returned was varied from request to request with the two sets kept roughly equivalent. In some cases, where there were only a few relevant documents to a particular request, only the relevant documents were

Document Rank	Document Number	Correlation	User Feedback
1	167	.46	not relevant
2	166	.43	not relevant
3	188	.40	-
4	221	.38	relevant
5	314	.38	-
6	55	.37	-
7	79	.36	relevant

(a) Retrieval results using original query "I-R Indexing" including user feedback

Retrieval Results Using Original Query			Results Using Query Modified by User Feedback		
Ranks of Relevant Documents	Document Number	Correlation	Ranks of Relevant Documents	Document Number	Correlation
4	221	.38	1	79	.54
7	79	.36	2	221	.47
13	3	.26	4	3	.33
15	80	.26	5	126	.31
17	48	.25	6	80	.30
23	126	.21	25	48	.17
Recall .976 Precision .728			Recall .991 Precision .928		

(b) Comparison of search results using original and modified queries

Query Processing Using Relevance Feedback

Figure 1

Document Rank	Document Number	Correlation	User Feedback
1	351	.65	relevant
2	353	.42	relevant
3	350	.41	relevant
4	163	.36	-
5	82	.35	-
6	1	.32	-
7	208	.27	not relevant
8	225	.25	not relevant
9	54	.24	-
10	335	.21	not relevant

(a) Retrieval results using original query for "Pattern Recognition" including user feedback

Retrieval Results Using Original Query			Results Using Query Modified by User Feedback		
Ranks of Relevant Documents	Document Number	Correlation	Ranks of Relevant Documents	Document Number	Correlation
1	351	.65	1	351	.66
2	353	.42	2	350	.60
3	350	.41	3	353	.55
4	163	.36	5	163	.37
6	1	.32	6	1	.32
9	54	.24	7	54	.29
26	205	.17	11	314	.23
27	224	.17	16	205	.19
33	314	.16	17	39	.19
34	39	.12	30	224	.16
Recall .972 Precision .864			Recall .989 Precision .923		

(b) Comparison of search results using original and modified queries

Query Processing Using Relevance Feedback

Figure 2

Document Rank	Document Number	Correlation	User Feedback
1	157	.42	relevant
2	165	.40	relevant
3	362	.39	not relevant
4	296	.37	relevant
5	308	.37	not relevant
6	307	.37	not relevant
7	226	.36	-
8	88	.36	-

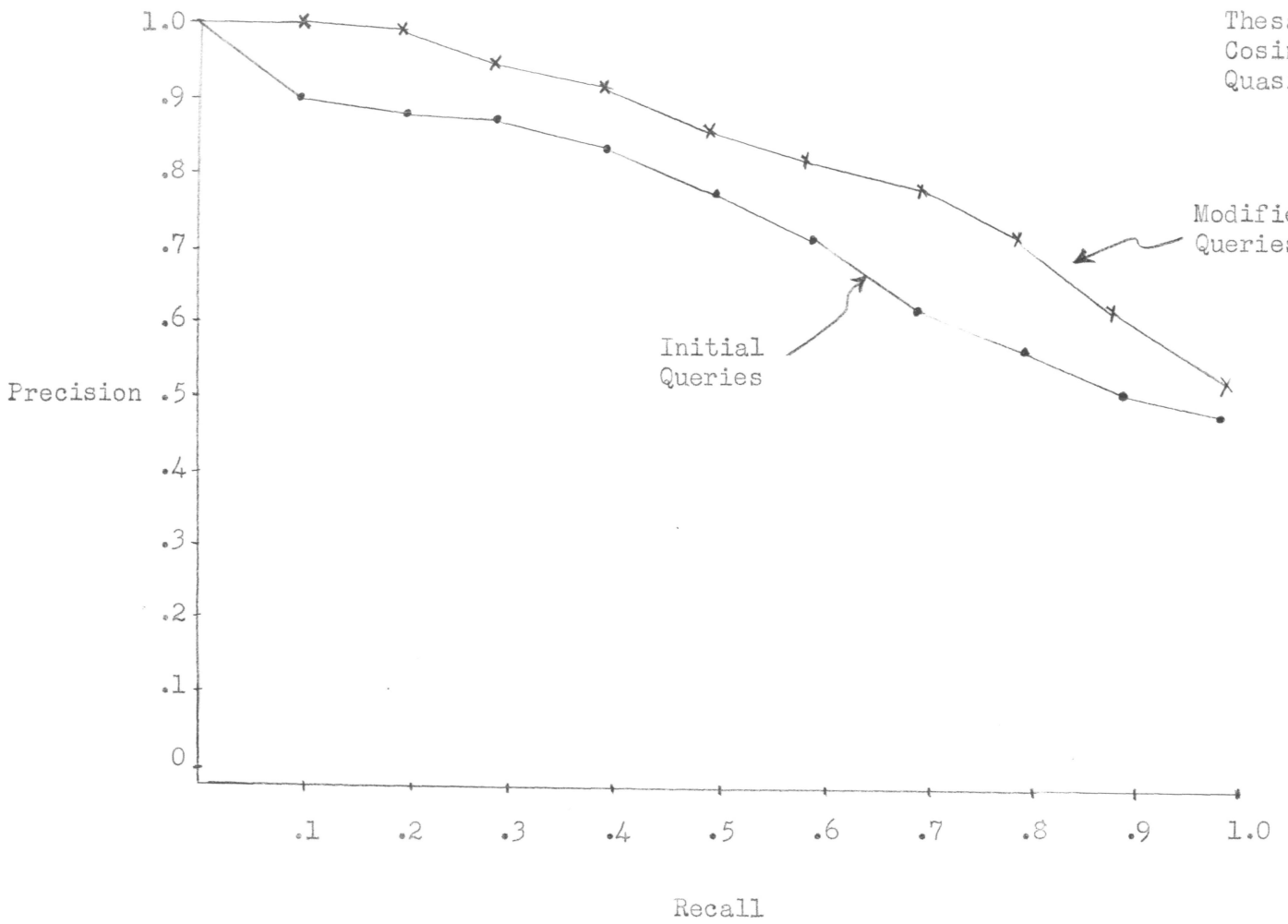
(a) Retrieval results using original query "Analog-Digital" including relevance feedback

Retrieval Results Using Original Query			Results Using Query Modified by User Feedback		
Ranks of Relevant Documents	Document Number	Correlation	Ranks of Relevant Documents	Document Number	Correlation
1	157	.42	1	296	.58
2	165	.40	2	157	.56
4	296	.37	3	165	.53
19	42	.27	4	42	.42
21	46	.26	40	46	.20
Recall .984 Precision .870			Recall .983 Precision .918		

(b) Comparison of search results using original and modified queries

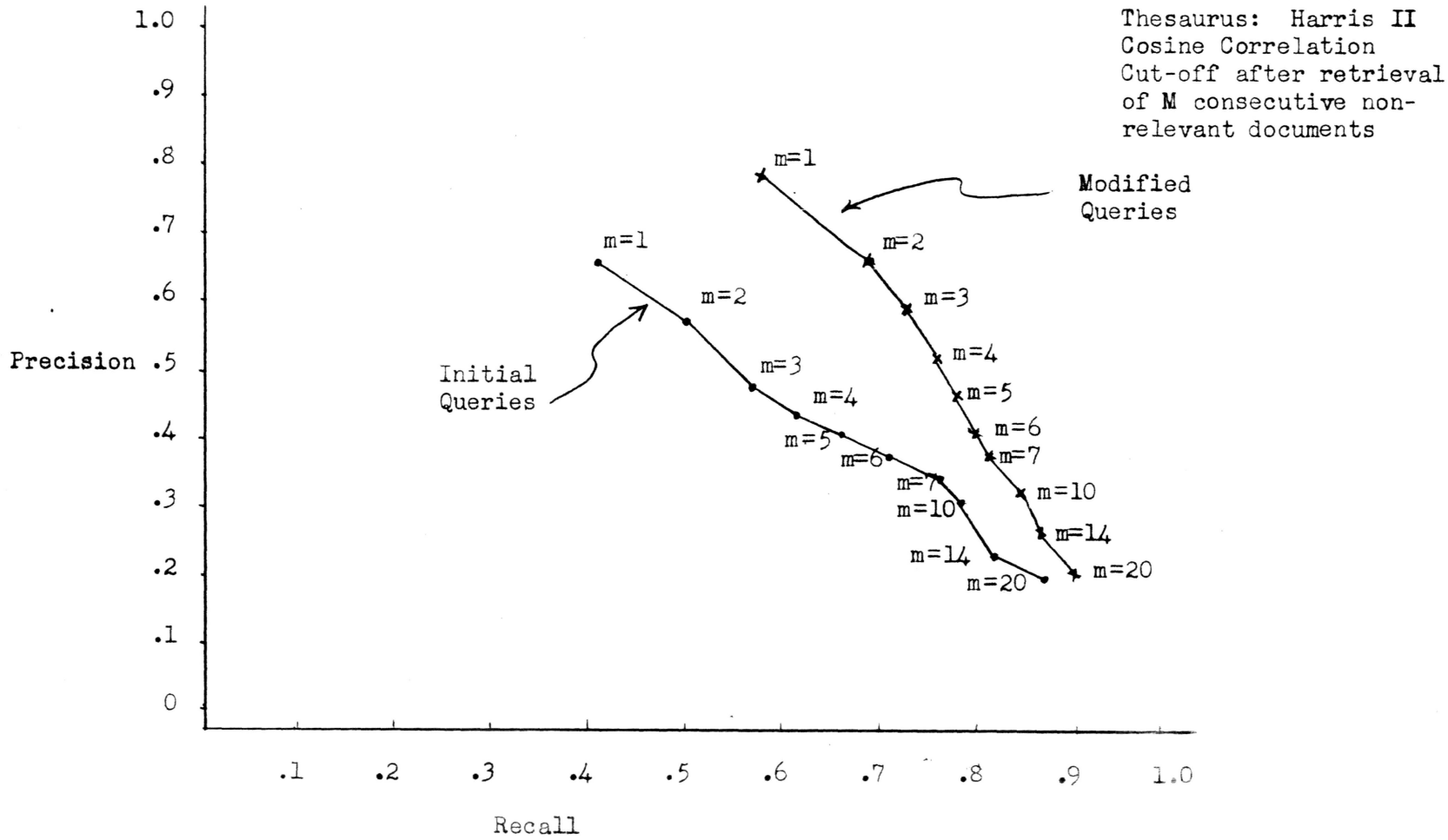
Query Processing Using Relevance Feedback

Figure 3



Precision Versus Recall for Initial Queries
and Queries Modified by Relevance Feedback
(averaged over 17 search requests)

Figure 4



Precision Versus Recall For Initial Queries
and Queries Modified by Relevance Feedback
(averaged over 17 search requests)

Figure 5

identified to see what would happen under these circumstances. Figures 1, 2, and 3 show the results of one iteration of the search request modification process for three different queries. In the first two cases there is substantial improvement in both the recall and precision evaluation measures (shown for the original and modified query), while in the third case only the precision measure is increased. It is important to note that while the query modification process improves the performance with respect to the relevant documents identified by the user, as one would expect, it also in general improves performance with respect to other relevant documents. This is illustrated in Figs. 4 and 5, which show precision versus recall averaged over a sample of 17 search requests. Figure 4 is the result of averaging the interpolated precision for each request at recall values of 0.1 k, for integral values of k from 1 to 10. Figure 5 is a precision versus recall curve for the same set of 17 queries produced in a different way. This plot results from choosing a cut-off after m consecutive nonrelevant documents (m ranging from 1 to 20), and averaging the precision and recall values for each request at each cut-off point to get a single average point which is plotted.

The general conclusion which can be tentatively drawn at present is that relevance feedback provides a powerful tool for improving performance at the cost of iteration of a request through the system.