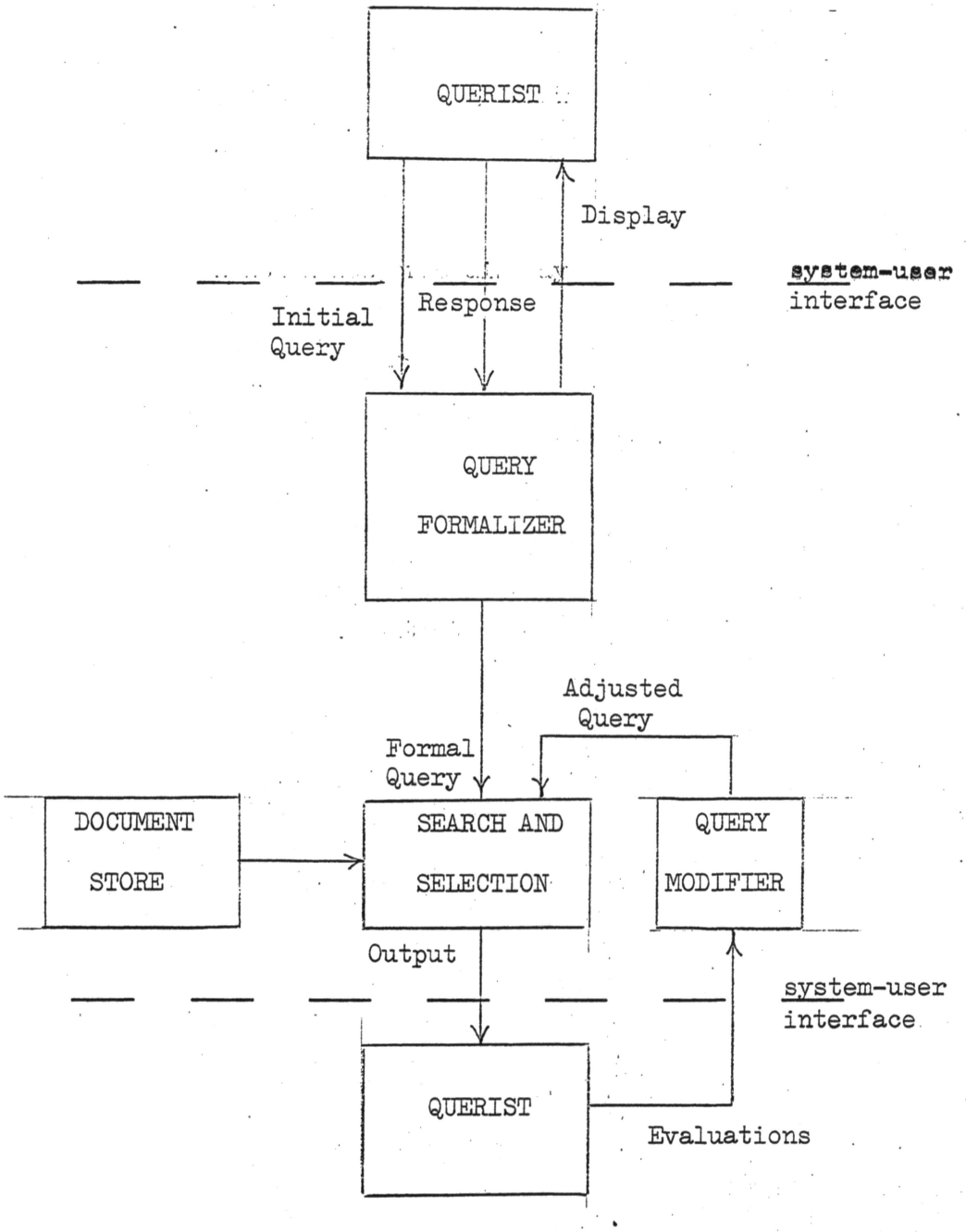
CHAPTER 3

SEARCH REQUEST FORMULATION

1. Introduction

The dialogue initiated by user-generated inputs to a document retrieval system is a dynamic communication process that needs to be optimized if the system is to provide effective service. A general block diagram of a retrieval system delineating possible user-system interfaces is shown in Figure 3.1. It will be assumed that the operational environment of the system allows for real-time interaction between the user and the system, and that this facility can be exploited in the implementation of the overall request formulation process. The block labeled "query formalizer" may be interpreted as an initial preprocessing stage designed to buffer the user from the internal structure of the system. Simple processes such as error detection (spelling, etc.), as well as more complex ones (vocabulary aids in the form of suggested synonyms and term relations), may be incorporated here.[1,2] The present section, however, deals with the implementation of the block labeled "query modifier". In particular, it is assumed that the system allows for iterative searching, and is capable of automatically modifying the user's original query on the basis of the results of previous iterations and of the user's evaluation of these results. To this end the concept of search request optimality is introduced, and the



Query Formulation- The User-System Interfaces

Figure 3.1

role of the query modifier is specifically identified with automatic query optimization.

In evaluating the performance of a document retrieval system, one must attempt to isolate the effect of the critical variables which determine system behavior. The formalization of the notion of request optimality leads, then, to the ability to measure accurately the behavior of the indexing function under a given query-document matching relation, since performance variations due to the malformation of a search request can be eliminated. In this context, then, request optimization can be a powerful evaluation tool for testing index language devices.

2. Request Formulation

The process of search request formulation in a document retrieval system is a complex one and depends on particular attributes of the requestor, such as his acquaintance with the reference collection, his knowledge of the indexing and searching functions of the system, his familiarity with the topic matter being searched, his personal preferences as to vocabulary, style, etc. In effect, each user must must make a statistical decision based on his individual experience as to what formal request statement is most likely to produce results useful to him. It can be assumed, therefore, that the a priori likelihood that a user's search request actually satisfies his information needs varies over a wide range for a typical user population. For example, some users will be able to

express quite explicit formulations of their information needs and can be expected, therefore, to obtain relevant source documents with a high probability of success. At the other end of the spectrum, however, there are those users with less explicit information needs or with only vague familiarity with the subject area being searched. Clearly, this class of users is less likely to be able to formulate search requests which will retrieve useful references.

Under these assumptions it is then pertinent to consider techniques for reducing search request variance in two distinct contexts. First, from an operational viewpoint, one would like to process search 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 context of retrieval evaluation, it is desirable to isolate explicitly the effect of the request formulation from the effects of indexing and request-document matching. Normally, in testing indexing devices one compares gross retrieval results obtained for some sample set of search requests. The comparisons so obtained reflect the joint behavior of the test queries and the indexing scheme, but do not provide an explicit comparison of the indexing methods alone. If it were possible, however, to define an optimal search request (for a fixed indexing technique), corresponding to any given test query, the comparative retrieval results for the optimal requests would provide a much clearer evaluation of the power of the indexing methods, since performance variations due to request malformation would be eliminated.

3. Request Optimization

To define an optimal search request, it is necessary to start with an explicit formulation of the model which specifies the retrieval system. In particular it will be shown that a reasonable definition of request optimality is directly related to the retrieval or query-document matching function. In the model outlined in Chapter 1, it was assumed that the matching criterion for selecting reference documents in response to input queries is the magnitude of the angular distance between the query vector image and the vector images of the source documents. It is now convenient to introduce a query-document correlation function which is a monotonic function of angular distance in the vector space (over the range $|\theta| \leq 180°$). Assuming, therefore, that the output of a retrieval operation is a partial ordering of all source document representations in the collection, D, derived on the basis of the angular distance from the input query image, the cosine correlation function,

$$\rho(\bar{a},\bar{b}) = \frac{\bar{a}\cdot\bar{b}}{|\bar{a}|\,|\bar{b}|} = \cos\theta_{\bar{a},\bar{b}} \qquad (3.1)$$

can be used to induce the same ordering. Note that the correlation function is inverse to the angular distance in that $\theta = 0°$ maps into $\rho = 1$, and $|\theta| = 180°$ into $\rho = -1$. Thus the range $0° \leq |\theta| \leq 180°$ maps into the range $-1 \leq \rho \leq +1$, so that increasing angular separation corresponds to decreasing correlation. It may also be noted that the restriction of the vector images to nonnegative components (as is the

case under the index representation assumed) results in a restriction of the range of the angular distance to $0° \leq |\theta| \leq 90°$, corresponding to a correlation range of $1 \geq \rho \geq 0$.

It is now necessary to postulate an explicit objective for any given retrieval operation. And corresponding to any search request representation, $\bar{q}$, it is assumed that there exists a non-empty subset D_R ($D_R \subset D$) of the set of reference document tokens D. This subset, D_R, is that set of document index images which corresponds to documents in the collection $\mathcal{L}$, relevant to the search request. As relevance must be subjectively defined, the specification of the subset, D_R, must be made outside the context of the retrieval system. It is then assumed that the information needs of the user can be satisfied by the content of those documents whose index images are contained in D_R. The case in which D_R is empty, i.e. when there are no useful references in the source collection, will be considered separately.

The identification of the subset D_R is the goal of retrieval. Since the query-document matching function results in an ordering of the source collection, an ideal search request can be defined as one which induces a ranking on the elements of D such that all members of the set D_R are ranked above (have a higher correlation) all other elements of D. Note that in this definition any degree of relevance or ordering among the members of the subset D_R with respect to their value to the user is ignored.

Since relevance is a subjective attribute of a given search request-document collection pair, determined in theory by the individual

user, there is no certainty that an ideal search request (under some fixed index transformation) in fact exists. Relevance, the relation between the query and the subset D_R is a function of the user's information needs and his interpretation of the text of reference documents; while the retrieval ordering produced by the system is a function of the query and document index representations. In the case where there is no ideal request corresponding to a given subset D_R, one might say that the index transformation is deficient from the point of view of the particular user (who specified D_R), since it does not allow distinctions equivalent to those he can make. In general one must assume that this will be the norm rather than the exception, since the indexing process is designed to reduce rather than preserve information. It is therefore useful to define an unambiguous, optimal search request as a function of D_R, D, and implicitly, therefore, of the index transformation, such that for every non-empty unique subset D_R of D, this optimal search request both exists and is unique.

An optimal search request index image corresponding to a given subset D_R of a collection of document images produced by the index transformation T is that request image $\bar{q}$ which maximizes the difference between the mean of its correlations with the relevant documents (members of D_R) and the mean of its correlations with the nonrelevant documents (members of D not in D_R).

The subjective relevance relation which specifies a subset D_R corresponding to each input query induces an effective partition

of the reference collection. The definition of the optimal request reflects the partition in terms of the statistical properties of the correlations of the query and source document index images.

Let $\bar{q}$ represent the index image of a search request and $\bar{d}_i$ the index image of a reference document ($\bar{d}_i = T(D_i)$, $D_i \in \mathcal{D}$). In mathematical terms, the optimal request vector $\bar{q}_o$ corresponding to a subset D_R of D is defined as that vector $\bar{q}$ which maximizes:

$$C = \frac{1}{n_o} \sum_{\bar{d}_i \in D_R} \rho(\bar{q},\bar{d}_i) - \frac{1}{m-n_o} \sum_{\bar{d}_i \notin D_R} \rho(\bar{q},\bar{d}_i) \qquad (3.2)$$

where $n_o = n(D_R)$ the number of elements in D_R, and $m = n(D)$ the total number of elements in the reference collection.

Substituting for $\rho(\bar{q},\bar{d}_i)$ and using vector notation results in:

$$C = \frac{1}{n_o} \sum_{\bar{d}_i \in D_R} \frac{\bar{q}\cdot\bar{d}_i}{|\bar{q}|\,|\bar{d}_i|} - \frac{1}{m-n_o} \sum_{\bar{d}_i \notin D_R} \frac{\bar{q}\cdot\bar{d}_i}{|\bar{q}|\,|\bar{d}_i|} \qquad (3.3)$$

or:

$$C = \frac{\bar{q}}{|\bar{q}|} \cdot \left[\frac{1}{n_o} \sum_{\bar{d}_i \in D_R} \frac{\bar{d}_i}{|\bar{d}_i|} - \frac{1}{m-n_o} \sum_{\bar{d}_i \notin D_R} \frac{\bar{d}_i}{|\bar{d}_i|} \right] \qquad (3.4)$$

From this last equation it is clear that C is just the dot product of a unit vector along the direction of $\bar{q}$ and a vector which

is a function of the relevant and nonrelevant partition classes of D. Thus C may be written as:

$$C = \bar{q} \cdot \bar{a}$$

and therefore the vector $\bar{q}_o$ which maximizes C is:

$$\bar{q}_o = k\bar{a}$$

or:

$$\bar{q}_o = k \left[\frac{1}{n_o} \sum_{\bar{d}_i \in D_R} \frac{\bar{d}_i}{|\bar{d}_i|} - \frac{1}{m-n_o} \sum_{\bar{d}_i \notin D_R} \frac{\bar{d}_i}{|\bar{d}_i|} \right] \qquad (3.5)$$

with k being an arbitrary scalar.

Two observations can be drawn from the result. First, if the vector sumations are taken over two arbitrary subsets, say R and S, resulting in:

$$\bar{q}_o = k \left[\frac{1}{n_1} \sum_{\bar{r}_i \in R} \frac{\bar{r}_i}{|\bar{r}_i|} - \frac{1}{n_2} \sum_{\bar{s}_i \in S} \frac{\bar{s}_i}{|\bar{s}_i|} \right] \qquad (3.6)$$

where $n_1 = n(R)$ and $n_2 = n(S)$; then $\bar{q}_o$ is that vector which maximizes the difference between the mean of its correlations with the members

of R and the mean of its correlations with the members of S. Second, it can easily be shown from the definition of the vector dot product that C is maximized by the vector $\bar{q} = \bar{q}'_o$ subject to the condition that the components of $\bar{q}$ be nonnegative. The components of $\bar{q}'_o$ are given by:

$$\bar{q}'_{o_j} = \begin{cases} \bar{q}_{o_j} & \text{if } \bar{q}_{o_j} \geq 0 \\ 0 & \text{if } \bar{q}_{o_j} < 0 \end{cases} \quad (j = 1,N) \qquad (3.7)$$

Hence, under the assumptions made, an unambiguous optimal (for the criteria stated) query image exists corresponding to any non-empty subset D_R of D. Further, the equation 3.5 provides an effective means of generating such a query from knowledge of the relevant subset D_R. In the evaluation of information retrieval systems and in particular in the evaluation of the indexing function of such systems, this formulation of an optimal search request provides the ability to isolate the effects of indexing from variances due to request formulation. An optimal search request measures the ability of the index transformation to differentiate a particular set of documents from all the others of a collection. In an evaluation situation, where one assumes prior knowledge of the document subset relevant to each test query, the retrieval performance of the optimal query corresponding to the relevant subset provides a direct measure of the ability of the system to extract from the index representations of documents the same kind of information the user can extract from the natural language.

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 a priori way to generate an optimal request, since the ability to do so would eliminate the need for retrieval. This kind of circularity suggests a strong analogy to feedback control theory. Consider therefore a sequence of retrieval operations which start with an initial query $\bar{q}_0$. A modified query $\bar{q}_1$ is to be produced based on the original output, such that $\bar{q}_1$ is a better approximation to the optimal query for this user than $\bar{q}_0$. 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 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 $\bar{q}_o$. 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 $\bar{q}_o$), the modified request will be formed by adding to the original query, $\bar{q}_o$, 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 $\bar{q}_o$, and should, therefore, produce better retrieval when resubmitted.

Hence we seek a relation of the form:

$$\bar{q}_1 = f(q_o, R, S)$$

where $\bar{q}_o$ 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 $\bar{q}_o$) which the user deems nonrelevant. The form suggested immediately by the above is:

$$\bar{q}_1 = \alpha_1 \bar{q}_o + \alpha_2 \left(\frac{1}{n_1} \sum_{i=1}^{n_1} \bar{r}_i - \frac{1}{n_2} \sum_{i=1}^{n_2} \bar{s}_i \right) \qquad (3.8)$$

where $n_1 = n(R)$, $n_2 = n(S)$, $R = \{\bar{r}_1, \bar{r}_2, \ldots \bar{r}_{n_1}\}$, $S = \{\bar{s}_1, \bar{s}_2, \ldots \bar{s}_{n_2}\}$, where all vectors have been normalized to unit length, and α_1 and α_2 are arbitrary weighting coefficients.

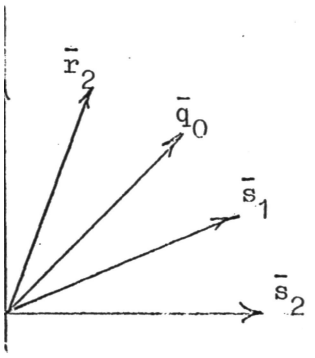
The modified query $\bar{q}_1$, then, is a weighted vector sum of the original query vector plus the optimal vector to differentiate the members of the set R from those of the set S. In other words, $\bar{q}_1$ is the vector sum of $\bar{q}_o$ plus the optimal vector for the subset of the reference collection for which the user has provided relevance information. If equal weight is given to the original query and the optimal vector based on the feedback information, equation (3.7) may be written in the form:

$$\bar{q}_1 = n_1 n_2 \bar{q}_o + n_2 \sum_{i=1}^{n_1} \bar{r}_i - n_1 \sum_{i=1}^{n_2} \bar{s}_i \qquad (3.9)$$

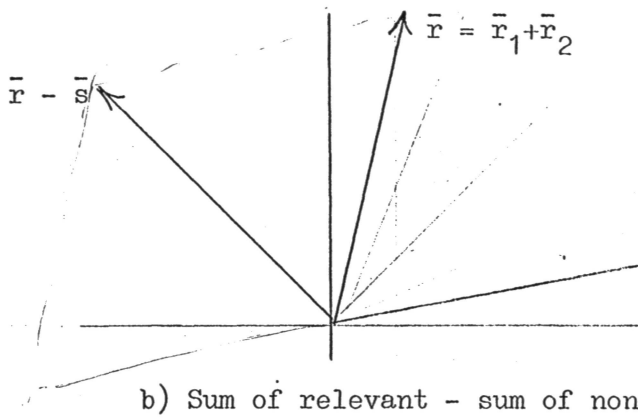
If $\bar{q}_1$ is to be restricted to a vector with only nonnegative components, the following may be used:

$$\bar{q}'_{1_j} = \begin{cases} \bar{q}'_{1_j} & \text{for } \bar{q}_{1_j} \geq 0 \\ 0 & \text{for } \bar{q}_{1_j} < 0 \end{cases}$$

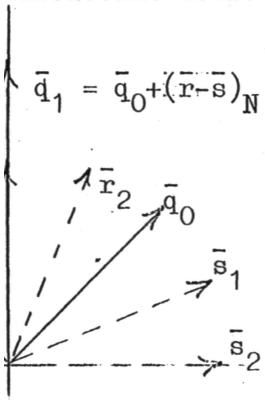
Figure 3.2 provides a two-dimensional geometrical interpretation of the relevance feedback request modification process. Part (a) shows the initial query $\bar{q}_o$ located between the relevant and non-relevant document vectors. The vector $\bar{r}$-$\bar{s}$ shown in part (b) is the optimal vector (i.e. the vector which maximizes the function C of equation (3.2)) for differentiating the subset $R = \{\bar{r}_1, \bar{r}_2\}$, from $S = \{\bar{s}_1, \bar{s}_2\}$. Part (c) shows the resultant of adding $\bar{q}_o$ to the normalized vector sum $\bar{r}$-$\bar{s}$, which results in the new



al query ($\bar{q}_0$), relevant docs. ($\bar{r}_1$,$\bar{r}_2$) onrelevant docs. ($\bar{s}_1$,$\bar{s}_2$).



b) Sum of relevant - sum of non document vectors.



tant of $\bar{q}_0$ + normalized (sum levant - sum of nonrelevant).

Doc.	$\bar{q}_0$	$\bar{q}_1$
r_1	0.71	1.0
r_2	0.92	0.9
s_1	0.92	0.3
s_2	0.71	0.0

d) Correlations of query vecto $\bar{q}_0$ and $\bar{q}_1$ with doc. vectors

Geometrical Representation Of Relevance Feedback

Figure 3.2

query vector $\bar{q}_1$. The table compares the correlations of $\bar{q}_o$ and $\bar{q}_1$ with the document vectors.

The modifications to an initial query vector which are produced by the relevance feedback algorithm may receive the following interpretation: concepts, i.e. components of the initial query which are more significant in the document images of the relevant subset than in the nonrelevant subset will be emphasized (i.e. increased in weight and visa-versa). Thus the weighting of the original query terms, derived from frequency counting, will be adjusted on the basis of the statistical evidence derived from the sample output for which the user provides relevance feedback. In addition, concepts not included in the original query but which are also useful in differentiating the relevant from the nonrelevant documents will be added to the modified query image. Such concepts (components of the index space) can be expected to be useful in retrieving other relevant documents not explicitly identified by the original query, since all relevant documents (which can be successfully retrieved) must be sufficiently related to be localized in some region of the index space.

The basic relation for request modification using relevance feedback (equation (3.8)) can be modified in various ways by imposing additional constraints. For example, the weighting of the original query 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 non-zero components of the modified query on the basis of the degree of overlap of a component among the relevant documents identified by the user. There are a number of additional variations to this basic relation which might be investigated.

The modification process described above for generating $\bar{q}_1$ from $\bar{q}_0$ is amenable to iteration and therefore can be written in the general form:

$$\bar{q}_{i+1} = F(\bar{q}_i, R^i, S^i) \qquad (3.10)$$

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 is to 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 in so far as the index images of the relevant set are differentiable from those of the nonrelevant set.

In this context it is possible that the information needs of a user might be best satisfied by a multiple rather than a single search request. This would be the case, for example, if useful references happened to be mapped by the index transformation into several distinct regions of the index space. Since the user in general has no a priori means of determining whether he should use a single or a multiple search (other than his own intuition,) it is of interest to consider automatic means for generating multiple searches. Assume, for example, that the relevant set R identified by a user after an initial retrieval operation contains document images sufficiently seperated so as to be considered only slightly related. Figure 3.3 shows an example in two dimensions. Under the circumstances portrayed the relevance feedback adjustment algorithm is not useful since, in fact, there is no single vector close to both relevant document images. This suggests that useful information can be derived by measuring the degree of association among the elements of the relevant subset identified by the user. Such information is contained in the document-document correlation matrix which characterizes this subset.

Consider, for example, the situation described by the

document-document correlation matrix of Figure 3.4. It is assumed that the set $R = \{\bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{r}_4, \bar{r}_5\}$ has been identified as a relevant subset in response to some initial search request, $\bar{q}_0$. These document images are correlated against each other producing the correlation matrix shown. If this matrix is used as a basis for partitioning the set R by some clustering technique, two subsets $R^1 = \bar{r}_1, \bar{r}_2, \bar{r}_3$ and $R^2 = r_3, r_4$ will result. In this case then, the system can generate two new queries by using each of these subsets together with the non-relevant set S. Thus the following pair of new search requests can be formed:

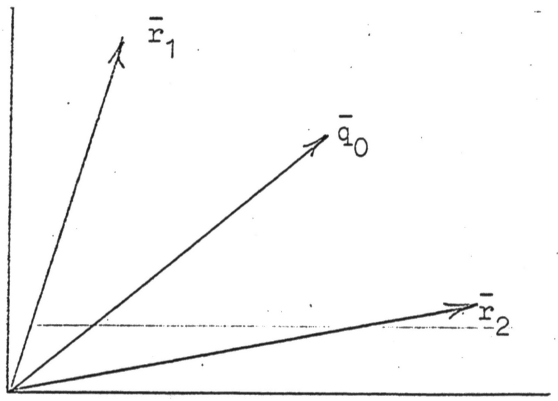
$$\bar{q}_1^1 = n_1' n_2 \bar{q}_0 + n_2 \sum_{\bar{r}_i \in R^1} \bar{r}_i - n_1' \sum_{\bar{s}_i \in S} \bar{s}_i \qquad (3.11)$$

and

$$\bar{q}_1^2 = n_1'' n_2 \bar{q}_0 + n_2 \sum_{\bar{r}_i \in R^2} \bar{r}_i - n_1'' \sum_{\bar{s}_i \in S} \bar{s}_i \quad , \qquad (3.12)$$

where $n_1' = n(R^1)$, $n_1'' = n(R^2)$, and $n_2 = n(S)$.

On the basis of a partition of the relevant subset identified by the user, two new search requests have been formed from a single original request. Roughly, this procedure amounts to allowing the user to identify particular documents in the collection and request additional references "like" those he has singled out. By examining the degree of association among the identified documents, it is possible to determine if this can be done efficiently with a single search request or whether multiple searching is required.



Relevant document images ($\bar{r}_1$ and $\bar{r}_2$) which have only slight association.

Figure 3.3

	r_1	r_2	r_3	r_4	r_5
r_1	1.0	.8	.7	.1	.2
r_2		1.0	.6	.2	.3
r_3			1.0	.2	.1
r_4				1.0	.8
r_5					1.0

a) Document-document correlations.

	r_1	r_2	r_3	r_4	r_5
r_1	1	1	1	0	0
r_2		1	1	0	0
r_3			1	0	0
r_4				1	1
r_5					1

b) Binary association matrix
(r_i associated with r_j if $\rho(r_i, r_j) > .5$)

Hypothetical document-document correlations among an assumed set of relevant documents.

Figure 3.4

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 document store.)

5. The Case of No Relevant Documents

The definition of an optimal search request assumed the existence of a nonempty set of documents relevant to each user's search request. The relevance feedback query optimization algorithm developed from the definition assumes that in response to the retrieval output generated by an initial query, feedback is received identifying both relevant and nonrelevant documents. Consider now the case in which either there are no relevant documents in the collection or none are identified by the user response to the initial retrieval operation. In this case the user is faced with a certain degree of uncertainty. If he is interested in ascertaining that there are in fact no useful documents in the collection, one possibility open to him is as follows: he may rephrase his search request and resubmit it. The relevance feedback query modification algorithm, when implemented with no relevant documents identified, will provide just the kind of adjustment to the original query which is useful in such a case. The modified query

would be produced by the equation

$$\bar{q}_1 = n_2\bar{q}_o - \sum_{i=1}^{n_2} \bar{s}_i \qquad (3.13)$$

where $S = \{\bar{s}_1, \bar{s}_2, \ldots\ldots \bar{s}_{n_2}\}$ constitutes those documents originally retrieved but judged nonrelevant.

In this case the modified query vector $\bar{q}_1$ represents a vector in the index space which is both related to the original query, and also moved from that region of the index space which was identified by the original retrieval operation. A sequence of such modifications allows controlled pertubations within the region of the index space of interest. Thus, even in the case where an initial search provides no relevant documents, the relevance feedback algorithm can be used to provide helpful query modifications to an iterative search process.

6. Experimental Results

A. Some Sample Search Requests

To test the effectiveness of the search request modification process based on relevance feedback as outlined above, some experiments were conducted using the SMART automatic document retrieval system. A set of sample search requests and a collection of 405 abstracts from the computer literature originally published in the

IRE Transactions on Electronic Computers (March - September, 1958) was available for this purpose.[3,4] Both the reference documents and each of the search requests which had been submitted at Harvard in the natural language were indexed using the SMART thesaurus. As the search requests had been used in a variety of previous retrieval experiments with this collection, relevance judgments for each query were also available, representing a full manual search through the complete reference collection.

A full retrieval ordering of the source documents with respect to each sample query was available, consisting of the correlation of each search request index image with every reference document image. From the initial portion of the retrieved list (ordered by descending correlations), two sets of documents were specified: one containing relevant documents and one containing nonrelevant documents. The vector index images of each search request, and the images of the documents in the two associated subsets were used as inputs to a Fortran program written to implement the query modification process. The output of this program was a new query vector suitable for input to the SMART system. The modified query images could then be correlated with the reference collection and the results compared with those of the original search requests.

Table 3.1 describes the program steps used to implement the relevance feedback query modification algorithm. Figure 3.5(a) shows the English text of a typical query. Figure 3.5(b) shows the explicit thesaurus mapping for the terms included in this query and part (c) shows the index image of the query in vector form (see Appendix A for

1. Read an initial query vector $\bar{q}_0$ (in integer format), convert it to a unit vector, and store it in the array Q(I), I = 1,N. (N is the dimension of the index language vector space.)

2. Read in the set of relevant document vectors $\bar{r}_j$, $j = 1,n_1$, convert them to unit vectors, and store them in the array R(I,J), I = 1,N; J = 1,N_1.

3. Read in the set of nonrelevant document vectors $\bar{s}_j$, $j = 1,n_2$, convert them to unit vectors, and store them in the array S(I,J), I = 1,N; J = 1,N_2. (Note that since the dimension of the index space was N = 511 for the thesaurus used, and since a document vector typically has about 35 nonzero components, the program actually handled the vectors in a condensed format.)

4. Form a new query vector represented by the array:

$$Q_1(I) = N_1 N_2 Q(I) + N_2 \sum_{J=1}^{N_1} R(I,J) - N_1 \sum_{J=1}^{N_2} S(I,J)$$

5. Normalize Q_1 to unit length:

$$Q_1(I) \leftarrow Q_1(I) \Big/ \left[\sum_{I=1}^{N} Q_1(I)^2 \right]^{\frac{1}{2}}$$

Program Steps for Producing Relevance Feedback Modified Queries

Table 3.1

6. Convert Q_1 to an integer format array:

$$IQ(I) \leftarrow \lfloor Q_1(I) \cdot 512 \rfloor ,$$

(where $\lfloor x \rfloor$ is the largest integer not exceeding x, and 512 is a scaling factor.)

7. Apply the screening algorithm to IQ(I) to produce the resultant modified query vector:

$$IQ(I) \leftarrow \begin{cases} IQ(I) & \text{if } X_1(I) \vee (X_2(I) \wedge X_3(I)) = 1 \\ 0 & \text{otherwise} \end{cases}$$

where $X_1(I)$, $X_2(I)$, and $X_3(I)$ are logical variables defined by:

$$X_1(I) = 1 \quad \text{if } Q(I) > 0$$

$$X_2(I) = 1 \quad \text{if } \sum_{J=1}^{N_1} R(I,J) \geq \lfloor N_1/2 \rfloor$$

$$X_3(I) = 1 \quad \text{if } \sum_{J=1}^{N_1} R(I,J) \geq \sum_{J=1}^{N_2} S(I,J)$$

Table 3.1 (continued)

an explanation of how the concept weights are derived.) The first part of the algorithm implements equation (3.9) directly and results in a new query image as shown in part (d) of Figure 3.5. The application of a screening process to this vector results in a final modified query image as shown in part (e). The screening process is designed to eliminate any negative components in the modified query image, as well as to reduce the positive nonzero components to those most likely to be useful. This latter feature is incorporated since the statistical evidence implicit in the user's relevance judgments may represent a relatively small sample.

Concepts which are retained after screening either a) occur in the original query, or b) occur in at least half of the relevant documents identified in addition to being more frequent in the relevant set than in the nonrelevant set. The screening algorithm thus serves to prevent the modified query from becoming too specialized to those relevant documents identified in the initial retrieval operation. In addition, reducing the number of nonzero components in the modified query image provides increased efficiency. With fewer components the modified search request requires less storage space and can be correlated with reference document images with fewer operations.

Negative components in a modified query image represent properties in the index space which are more significant among the nonrelevant documents retrieved by the user's original search request, than among the retrieved relevant documents. In principle then, there are no conceptual difficulties in allowing such negative weights. This is in contrast to generating property vector index representations of

Automatic Information Retrieval and Machine Indexing.

a) Text of Search Request "I-R Indexing".

Query Term	Concept Code(s)
automatic	119
information	53, 350
retrieval	26
machine	41, 119, 338
indexing	101

b) Thesaurus Mapping of Query Terms.

Concept Code	Weight
26	12
41	4
53	6
101	12
119	16
338	4
350	6

c) Compressed Representation of Query Vector.
(Only the components with nonzero weight are shown.)

A Typical Sample Search Request

Figure 3.5

12 (38)	13 (-157)	16 (24)	17 (17)	26 (202)
29 (-78)	34 (17)	36 (48)	39 (-76)	40 (12)
41 (18)	44 (12)	47 (24)	49 (24)	53 (47)
57 (13)	58 (37)	59 (24)	68 (-40)	73 (12)
74 (24)	77 (-53)	86 (25)	93 (13)	101 (273)
107 (-107)	108 (126)	110 (49)	113 (12)	114 (48)
119 (133)	121 (8)	130 (12)	132 (25)	135 (13)
136 (24)	142 (12)	143 (38)	146 (13)	147 (-53)
149 (49)	158 (24)	167 (37)	170 (72)	173 (12)
176 (12)	178 (13)	179 (12)	182 (8)	184 (12)
202 (25)	218 (24)	220 (12)	237 (24)	239 (24)
250 (12)	260 (51)	261 (25)	308 (24)	338 (14)
350 (101)	353 (24)	496 (25)	497 (86)	

d) Basic Modified Query Vector after Relevance Feedback.
(Shown in compressed form where n (w) implies that component n has weight w.)

26 (202)	41 (18)	53 (47)	58 (37)	101 (273)
108 (126)	110 (149)	119 (133)	149 (49)	167 (37)
338 (14)	350 (101)	497 (86)		

e) Final Modified Query Vector for Request I-R Indexing.
(Produced from the vector shown in part (d) by the screening process.)

Figure 3.5 (continued)

reference ~~documents~~ documents, where a negative weight would imply that one could measure the degree to which a certain attribute was lacking. For this reason, and because the recognition of negations or exceptions in search requests would require a high degree of syntactic sophistication, the simulation system (SMART) does not have any facilities for processing query or document vector images with negative components. Such components arising from the modification algorithm (see Figure 3.5 (d)) were eliminated then, to preserve compatibility with the simulation system. Allowing negative components in a modified version of a user's search request would amount to an effective increase in the quantization of the index space. It may be postulated then that this would lead to improved performance. An experimental investigation into this possibility was not feasible, because it would have required substantial changes in the simulation system.

Figure 3.6 (a) shows the initial portion of the retrieved output generated for the original query "I-R Indexing" which is described in Figure 3.5. The relevance feedback used in this example consisted in identifying the initial two relevant and two nonrelevant documents in the retrieved list. For reference, the titles of all the relevant documents for this search request are provided in Table 3.2. Figure 3.6 (b) compares the retrieval results of the original and modified queries with respect to this full set of six relevant documents which are in the reference collection. Note that the modification has substantially improved the performance with respect to three out of the four relevant documents <u>not</u> originally identified by the user. Figure 3.7 compares the correlation distributions of the original and relevance

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

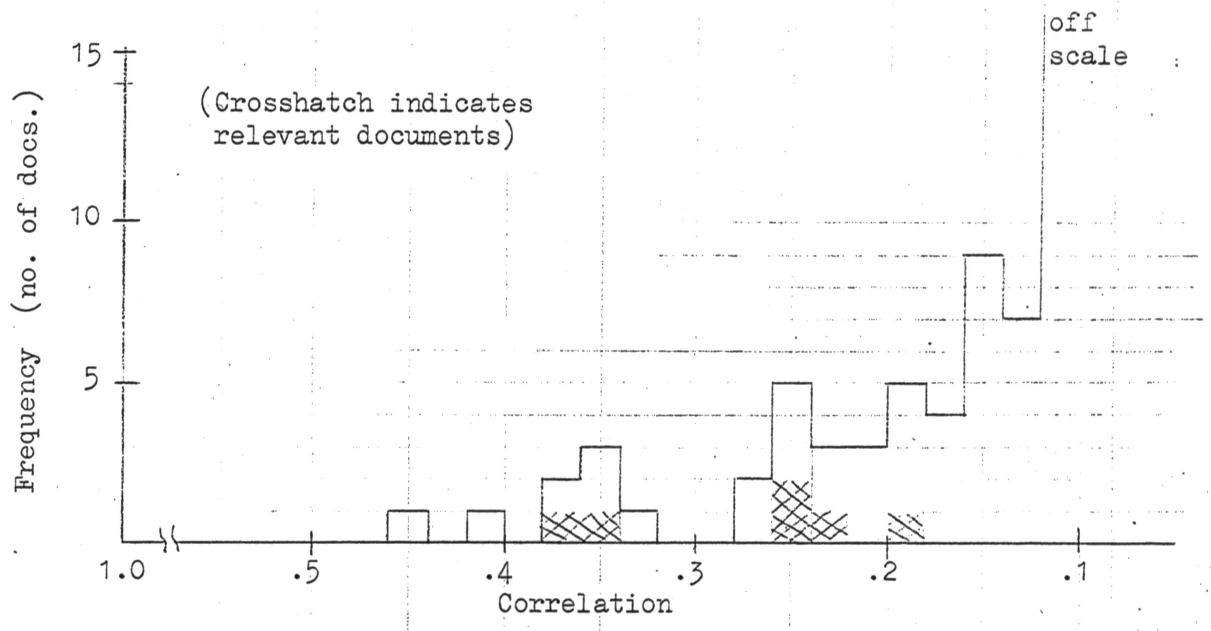
Query Processing Using Relevance Feedback for Search Request "I-R Indexing"

Figure 3.6

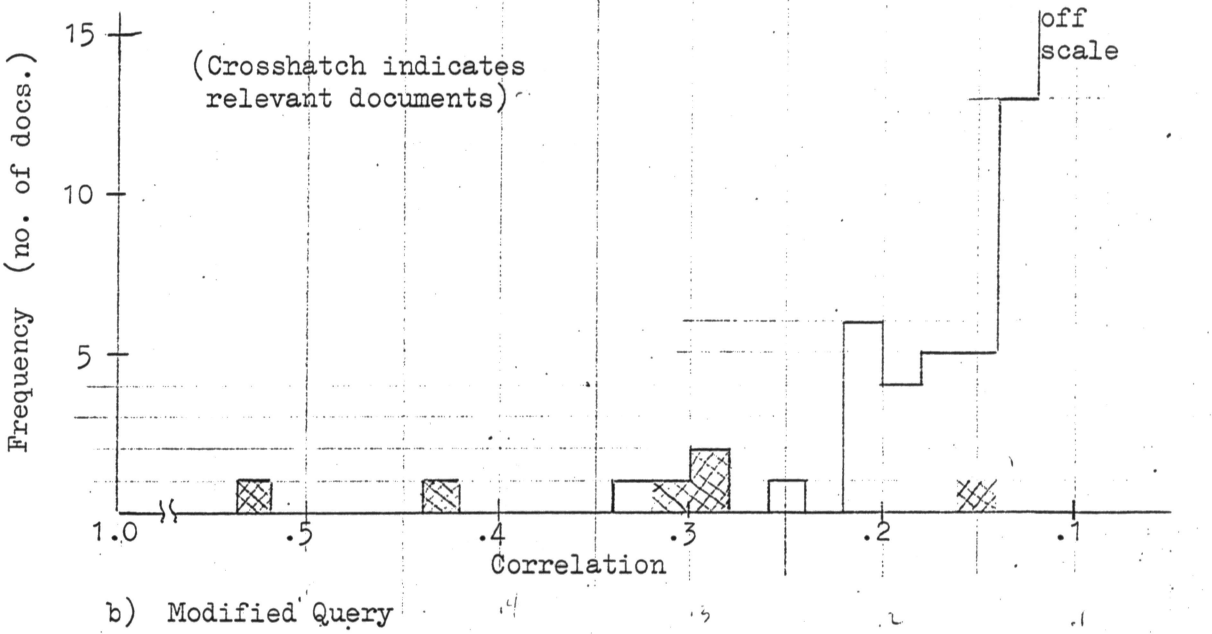
Doc. No.	Title
3	The Role of Large Memories in Scientific Communications
48	A Business Intelligence System
79	Machine-Made Index for Technical Literature- An Experiment
80	Rapid Retrieval of Information
126	How Much Science Can You Have At Your Fingertips
221	Chemical Structure Searching With Automatic Computers

Titles of Documents Relevant to Search Request "I-R Indexing"

Table 3.2



a) Original Query



b) Modified Query

Correlation Distributions for Search Request "I-R Indexing"

Figure 3.7

feedback modified queries.

Figures 3.8 and 3.9 compare the retrieval results of two additional original and relevance feedback queries. Note that for the query "Analog-Digital" shown in Figure 3.9, one of the relevant documents namely document 46, experiences a decrease in its retrieved rank from 21st to 46th while the ranks of the other relevant documents are substantially improved. This may be interpreted as implying that the index image of document 46 is less associated (in terms of angular distance in the index space) with the other relevant documents than it is with the original query. In general, this effect occurs, whenever the index images of the documents relevant to a search request form distinct clusters in the index space, and when the set of relevant documents identified by the relevance feedback consists substantially of members from only one of these clusters. In some cases it will be possible to identify such situations, and automatically to generate multiple queries for such search requests. For the case in point, however, the single document (no. 46) is assumed not to be identified by the original query. In this instance, then, there is no effective way to increase the probability of retrieving it. Such situations must then be interpreted (assuming that there are no grounds on which to question a user's relevance judgments), as arising from deficiencies in the indexing process or from the inherent information loss which necessarily accompanies it.

B. Average Results and Successive Iterations

The query modification procedure as illustrated in the

Document Rank	Document Number	Correlation	User Feedback
1	351	.65	Relevant
2	353	.42	Relevant
3	350	.41	Relevant
4	163	.36	Relevant
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 for Search Request "Pattern Recognition"

Figure 3.8

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 Feedbacks

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		

Query Processing Using Relevance Feedback for Search Request "Analog-Digital"

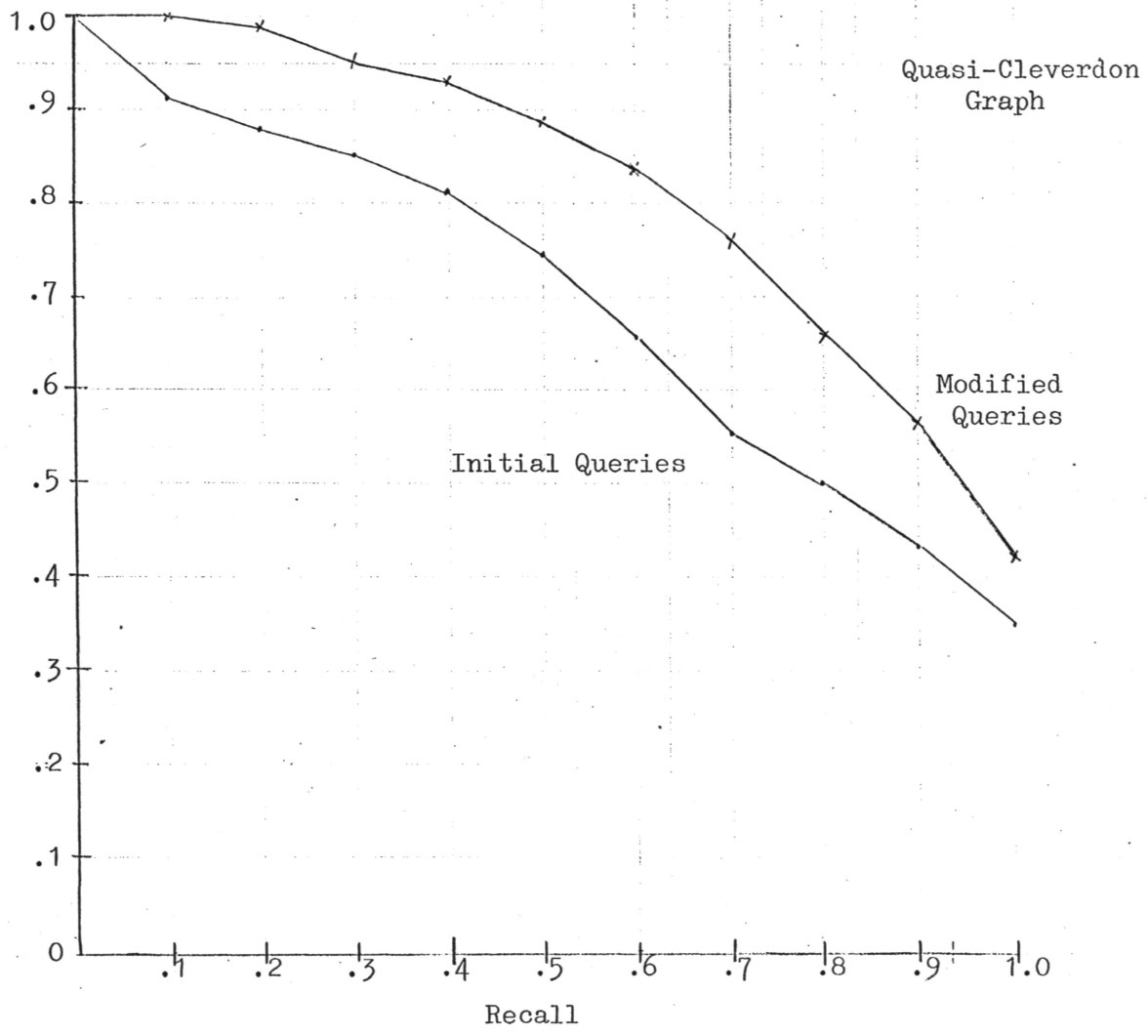
Figure 3.9

examples shown in Figures 3.6 through 3.9 was applied to the full set of search requests available. Average results comparing the performance of the original and relevance feedback modified queries are shown in the form of a precision vs. recall graph[x] in Figure 3.10. Since this means of exhibiting performance is based solely on the ranks of the relevant documents resulting from the query-document correlation process, it does not exhibit the true improvement which results from relevance feedback modification. This may be appreciated from the example shown in Figure 3.11, which illustrates another of the sample queries. In this case both the original and the modified query exhibit ideal performance (i.e. the relevant documents are all ranked higher than any nonrelevant documents). Thus the precision vs. recall graphs for both cases are identical. The correlation distribution, however, indicates that, in fact, the modified query provides greater discrimination of the relevant set from the nonrelevant set. In any case the average results as shown in Figure 3.10 indicate that the modification algorithm results in substantial improvement.

The request optimization procedure as illustrated by equation (3.10) can be used iteratively. The querist can, if he desires, provide evaluation information about the output generated by the first iteration and request that a second query modification take place. If R^1 and S^1 are the relevant and nonrelevant subsets

---

[x] The method of construction of such recall-precision plots has previously been described in detail.[4,5]



Precision vs Recall for Initial Queries and Queries Modified by Relevance Feedback (Averaged Over 24 Search Requests)

Figure 3.10

Document Rank	Document Number	Correlation	User Feedback
1	91	.57	Relevant
2	237	.56	Relevant
3	300	.55	Relevant
4	94	.50	Relevant
5	365	.33	---
6	219	.29	---
7	347	.28	---
8	68	.27	---

a) Retrieval Results Using Original Query for "Random Numbers" Including User Feedback

Document Rank	Document Number	Correlation
1	91	.74
2	94	.73
3	300	.70
4	237	.69
5	219	.37
6	254	.31
7	347	.26
8	195	.26

b) Retrieval Results for Modified Query

Comparison of Retrieval Results Using Original and Modified Query for Search Request "Random Numbers"

Figure 3.11

identified in response to the original query, and R^2 and S^2 are the corresponding subsets identified in response to the modified query, then, for the second iteration, the sets $R^T = R^1 \cup R^2$, and $S^T = S^1 \cup S^2$ are available for the optimization algorithm. Basically, two alternatives are possible: the optimal vector to differentiate documents in R^T from documents in S^T may be used as a perturbation of the user's original query; or this vector may be used as a perturbation of the query resulting from the first iteration. In practice this could be left for the user to decide, depending on his interpretation of the output from the first iteration.

In the general case the expression for the nth modified query in which all modifications are made to the original query can be written:

$$\bar{q}_n = \alpha_1 \bar{q}_o + \alpha_2 \left[\frac{1}{n_1} \sum_{\bar{r}_i \in R^T} \bar{r}_i - \frac{1}{n_2} \sum_{\bar{s}_i \in S^T} \bar{s}_i \right] \qquad (3.14)$$

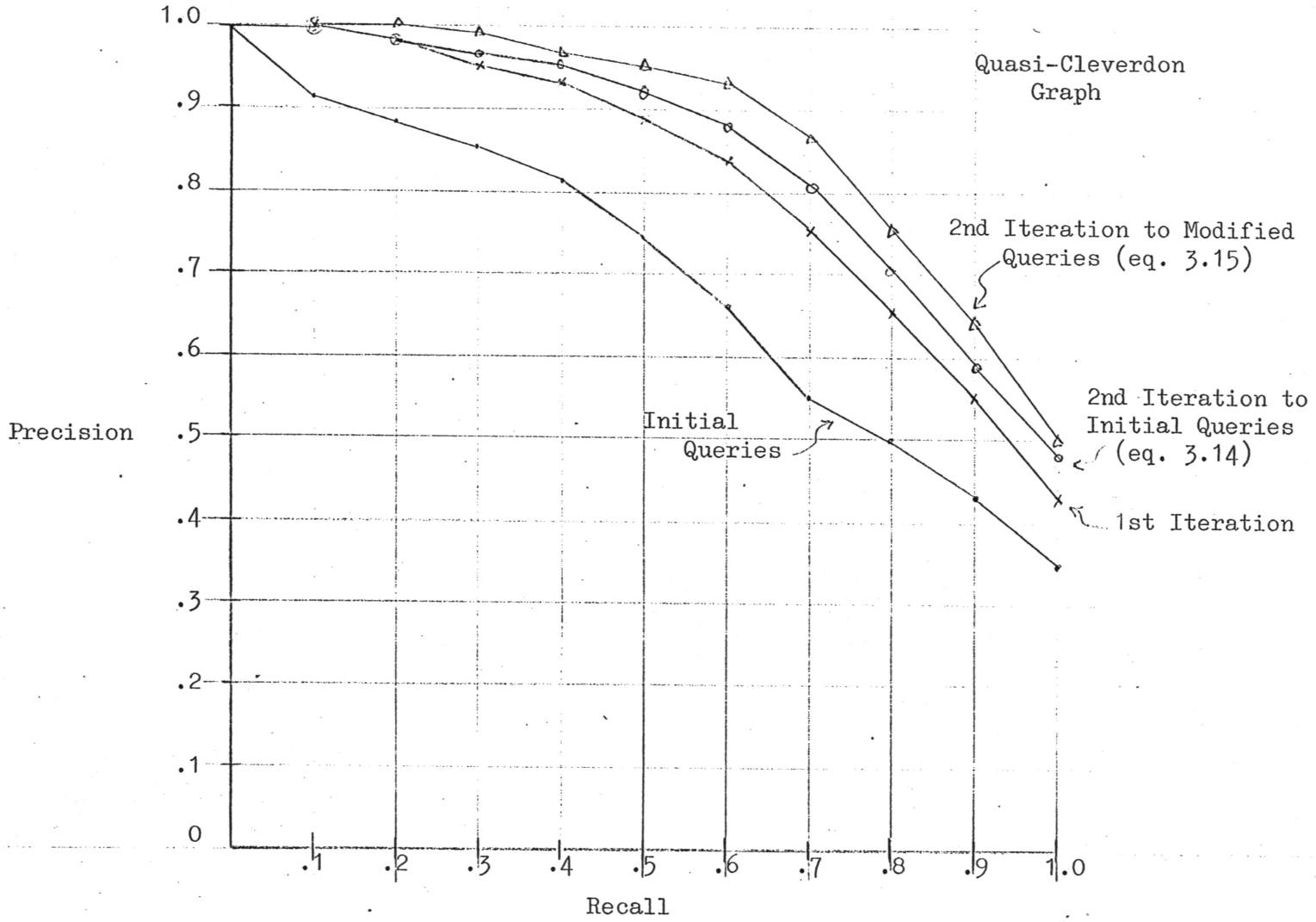
where $R^T = \bigcup_{i=1}^{n} R^i$, $S^T = \bigcup_{i=1}^{n} S^i$, $n = n(R^T)$, $n_2 = n(S^T)$, and α_1 and α_2 are weighting coefficients.

If each modification is made to the preceding resultant query image, the iteration formula becomes:

$$\bar{q}_n = \alpha_1 \bar{q}_{n-1} + \alpha_2 \left[\frac{1}{n_1} \sum_{\bar{r}_i \in R^T} \bar{r}_i - \frac{1}{n_2} \sum_{\bar{s}_i \in S^T} \bar{s}_i \right] \qquad (3.15)$$

Since operationally it is expected that only a few iterations would ever be used, the differences between these alternative formulations is not of major significance. If the user is satisfied with the relevant documents identified by the previous iterations and would, in effect, like to find others which are closely related to these documents, queries produced by equation (3.15) would be more suitable. If on the other hand, he is interested in maintaining a broader search, the iterations produced by equation (3.14) will not be as dependent on the relevant documents previously identified (members of R^T).

Average performance results for a second iteration of relevance feedback produced by each of these alternatives are shown in Figure 3.12. The results obtained with the original and first iteration queries are included for comparison. As can be seen by these graphs, the results obtained from using the iteration formula of equation (3.15) are somewhat better than when the second iteration starts from the original query. However, in comparing the behavior of these alternatives on individual queries, there are some cases in which the reverse is true. Figure 3.18 illustrates an example of this. In this case it is clear that documents 315 and 264 are not clustered in the index space with the other relevant documents; and therefore, these documents suffer more drastically from successive iterations (equation (3.15)) than from successive modifications to the original query (equation (3.14)).



Precision vs Recall for Initial Queries and Two Iterations of Relevance Feedback

Figure 3.12

Original Query		1st Iteration		2nd Iteration (eq. (3.15))		2nd Iteration (eq. (3.14))	
Rank	Doc. No.	Rank	Doc. No.	Rank	Doc. No.	Rank	Doc. No.
1	316	1	316 •	1	316 •	1	316 •
2	129	2	313 •	2	313 •	2	313 •
3	313	3	129 •	3	129 •	3	129 •
4	176	4	176 •	4	176 •	4	176 •
7	371	5	371	5	372 •	5	371 •
29	372	6	372	6	371 •	6	372 •
38	241	14	241	10	241	7	241
42	315	46	315	68	315	59	315
74	264	131	264	176	264	90	264

(The documents marked "•" were used in the modification algorithm.)

Retrieved Ranks of Relevant Documents for Query "Automata", Comparing Original Results with Those Obtained on Successive Iterations

Figure 3.13

C. Convergence

The performance improvement which results from a query optimization produced by relevance feedback modification is a function of the quality of the initial query, the degree of association of the index images of the relevant documents, and the amount of feedback. To investigate the influence of the latter parameters, some additional experiments were conducted. Figure 3.14 shows the retrieval results as a function of the amount of feedback for the query "IR-Indexing", and Figure 3.15 for the query "M9 Natlang". The document-document correlation matrix for those documents relevant to the query "IR-Indexing" are shown in Figure 3.16. Thus the rapid improvement obtained even with a small amount of feedback can be attributed to the fact that the members of the relevant set are all closely associated. In the case of the query "M9 Natlang", this is not true, and the document-document correlation of the first five relevant documents, retrieved by the original query indicates this. The relevance judgments for this query were made assuming a very general point of view. In this case it might be of use to produce multiple modified queries by seeking clusters in the relevant set. A possible partition based on the document-document correlations is shown in Figure 3.16, and this partition was used to generate two modified queries following equations (3.11) and (3.12). The retrieval results for these two modifications are shown in Figure 3.17. This figure illustrates how each of these queries is useful in retrieving some relevant documents. The fact that some of the relevant documents have low correlations with both of the

	Original Query	Modified Queries*			
		N=1	N=2	N=3	N=6
Ranks of Relevant Documents	4	1	1	1	1
	7	2	2	2	2
	13	3	4	3	3
	15	4	5	5	6
	17	5	6	6	7
	23	31	25	8	8
Average Corr. of All (6) Relevant Docs.	.288	.425	.354	.459	.461
Average Corr. of 1st 6 Non-relevant Docs.	.396	.321	.258	.312	.310
Difference	-.108	.104	.096	.147	.151

*(N relevant and N nonrelevant documents used for relevance feedback.)

Retrieval Performance as a Function of Amount of Feedback for Query "I-R Indexing"

Figure 3.14

	Original Query	Modified Queries*		
		N=1	N=3	N=5
Ranks of Relevant Documents	1	1	1	1
	6	3	2	2
	8	5	3	3
	9	7	5	4
	10	13	6	6
	13	15	12	9
	15	18	16	10
	18	22	18	15
	22	29	22	16
	24	32	23	17
	26	33	24	20
	27	34	27	21
	39	36	28	26
	52	53	32	44
	59	60	50	46
	69	69	66	51
	111	87	105	96
	251	98	122	137
	273	340	269	248
Average Corr. of All (19) Relevant Documents	.247	.222	.260	.263
Average Corr. of 1st 19 Nonrelevant Documents	.315	.255	.288	.267
Difference	-.068	-.033	-.028	-.004

*(N relevant and N nonrelevant documents used for relevance feedback)

Retrieval Performance as a Function of Amount of Feedback for Query "M9 Natlang"

Figure 3.15

Document No.	221	79	3	80	48	126
221	1.0	.24	.38	.33	.32	.37
79		1.0	.39	.29	.30	.36
3			1.0	.49	.42	.40
80				1.0	.56	.37
48					1.0	.41
126						1.0

a) Correlation Matrix for Documents Relevant to Query "I-R Indexing".

Document No.	221	223	314	80	112
221	1.0	.15	.32	.33	.14
223		1.0	.09	.19	.25
314			1.0	.13	.13
80				1.0	.09
112					1.0

b) Correlation Matrix for a Subset of the Documents Relevant to Query "M9 Natlang".

$$D_1 = \{221, 314, 80\} \qquad D_2 = \{223, 112\}$$

c) Document Clusters Derived from the Matrix of Part (b).

Document-Document Correlation Matricies

Figure 3.16

Retrieval Results With Relevance Feedback R = {112, 223}		Retrieval Results With Relevance Feedback R = {221, 80, 314}	
Ranks of Relevant Documents	Document Number	Ranks of Relevant Documents	Document Number
1	112	1	221
2	221	2	80
4	223	3	3
6	115	6	126
9	222	8	314
10	183	10	162
14	81	12	48
16	113	19	112
18	3	25	81
22	255	34	115
29	80	38	79
33	314	39	222
35	126	40	223
42	162	42	183
51	48	66	113
54	79	84	255
65	116	140	116
166	125	161	125
212	114	299	114

Retrieval Results for Two Queries Produced from Original Search Request "M9 Natlang"

Figure 3.17

modified queries is indicative of the broadness of the relevance judgments used in specifying D_R.

The use of multiple queries in general, must be justified in terms of the expected improvement in retrieval performance compared with the added cost of an additional search. Thus in respect to search cost, a two-way multiple search is equivalent to an additional iteration of a single query. Since the cost of query modification by relevance feedback is likely to be a small fraction of the cost of a search operation, the typical user maximizes his return by supplying a good relevance feedback sample (i.e. by carefully examining the initial retrieved output). Multiple searching may, however, be warranted in those cases where the user must obtain access to <u>every</u> relevant document. In such cases one can assume that the value of obtaining the reference justifies the additional cost.

REFERENCES

1. Rocchio, J.J. and Salton G., "Information Search Optimization and Iterative Retrieval Techniques", AFIPS Conference Proceedings, Vol. 27, Part 1, Spartan Books, Washington, D.C. 1965

2. Curtice, R.M. and Rosenberg, V., "Optimizing Retrieval Results with Man-Machine Interaction", Center for the Information Sciences, Lehigh University, Bethlehem, Pennsylvania (1965)

3. Salton, G., et al., "Information Storage and Retrieval", Reports ISR-7, ISR-8, National Science Foundation, Harvard Computation Lab., June and Dec. 1965

4. Salton G., "The Evaluation of Automatic Retrieval Procedures - Selected Test Results Using the SMART System", American Documentation, Vol. 16, No. 3, July 1965

5. Cleverdon, C.W., "The Testing of Index Language Devices", ASLIB Proceedings, Vol. 15, No. 4, April 1963