Interactive Search Strategies and Dynamic File Organization in Information Retrieval

E. Ide and G. Salton

Abstract

A great deal of effort has been devoted in recent years to the evaluation of automatic or semi-automatic information retrieval systems. Recent evaluation results indicate that the search effectiveness presently achieved, or likely to be achievable in the foreseeable future, is much smaller than expected by a majority of the potential user population. Furthermore, theoretical advances in language analysis and data organization promise only relatively modest future improvements.

The most significant advances in retrieval effectiveness are likely to be obtained by adaptive interaction techniques that extract information from the user during the search process to improve the organization of the data space, thereby providing more effective search and retrieval operations. The various user feedback techniques described either modify the user queries in such a way as to bring these queries closer to existing groups of relevant documents, or modify the document space to bring relevant documents closer to the corresponding search requests.

1. Retrieval System Performance

Over the past few years, the design of improved information storage and retrieval systems has become important to an increasing segment of the technically trained population. As a result, considerable attention has been paid to the development of automatic or semi-automatic hardware and software systems designed to store ever increasing amounts of information, and to make the stored data available to selected user classes. As the interest has grown in the development of automatic information systems, procedures for evaluating the performance of information systems have also become increasingly important, since the large investments necessarily required in a mechanization of information handling procedures would not be justified without some assurance that the resulting systems could render reasonably effective services.

Evaluation studies of information system performance are often carried out by choosing some subset of the information requests submitted to a given system, and identifying as "relevant" to each query a list of items that have been hand-selected by the user or by a subject expert. The effectiveness of the search and retrieval system is then measured by determining the extent to which the selected relevant items have been retrieved and other items have been rejected in answer to the query sample.

Two standard retrieval measures have been widely used to evaluate retrieval effectiveness: recall, defined as the proportion of relevant items actually retrieved; and precision, the proportion of retrieved items actually relevant. A perfect system, achieving both maximum recall and maximum precision, is not generally achieved in actual practice. In fact, recall is found to vary inversely with precision, that is, as the recall of a system goes up because more relevant items are retrieved, precision goes down because more irrelevant items are also retrieved. Therefore, the user must choose between obtaining either high recall and low precision, or high precision and low recall.

The average search results obtained in several recent retrieval evaluation studies vary between a recall of 0.1 at a high precision of 0.9, for specific and narrow search statements, and a recall of 0.9 at a low precision of 0.2, when the search statement is interpreted broadly. [1,2,3] Operational systems normally compromise by operating in the middle ranges where neither the recall nor the precision are very low. In fact, the Medlars system of the National Library of Medicine is said to operate at an average recall of 0.58 and an average precision of 0.50, thus producing the correct retrieval of about sixty percent of what is wanted, while keeping the amount of useless material also retrieved to about fifty percent. [3]

Two pragmatic approaches are being actively pursued in an attempt to improve the retrieval effectiveness of existing or proposed information systems. The first one consists in using more refined information analysis procedures designed to generate query and document identifiers more reflective of information content. For example, the experimental automatic SMART document retrieval system which provides fully automatic document and query analysis, includes procedures for automatic synonym recognition using stored dictionaries and thesauruses, for the assignment of phrase identifiers instead of simple terms, for the refinement or broadening of information identifiers using stored hierarchical subject arrangements, and for the use of statistical and syntactic language analysis methods. [4,5]

The second, more recently used method of improving retrieval effectiveness utilizes automatic information displays during an on-line search procedure in an attempt to prod the user into submitting more viable search statements. Excerpts of stored dictionaries or term lists can be displayed, as well as term frequency information, lists of related words, and titles or abstracts of stored documents. [6,7,8]

While both advanced language analysis methods and on-line interactive display techniques appear to improve retrieval effectiveness, the increment of improvement generated is relatively small, being generally from five to fifteen percent. [2,7] It thus appears that by methods which are well understood and seem economically reasonable, recall and precision figures of 0.60 to 0.65 are presently achievable at least in experimental environments.

Whether more dramatic improvements may be expected in the future — for example by the use of more refined grammatical models such as transformational language analysis — remains to be seen. Some evidence exists to suggest that presently obtainable results are only about twenty-five percent lower than those produced by an "ideal" search system, where human subject experts conduct exhaustive manual searches through the complete stored collection. [9] Therefore, recall and precision results of about 0.75 may constitute an upper bound to the performance of both automatic and manual retrieval systems. Whether any automatic system can achieve such results depends to some extent on the ability of the system to adapt to the expectations of the particular user population being serviced. Heuristic methods for this purpose are described in the remainder of this study.

2. Request Space Modifications

A) Relevance Feedback

A principal technique for improving the performance of automatic information retrieval consists in using information supplied by the customer in order to alter the request to correspond to the user's need. Specifically, the query representation — consisting in many retrieval systems of weighted sets of terms or concepts — can be changed by adding or stressing concepts which appropriately identify the user's information need and minimizing or even deleting concepts which are not representative of the user's need. The altered query should then be more similar to the stored representations of documents relevant to the user and less similar to the representations of nonrelevant documents.

One way in which this can be accomplished is by performing an initial search of the collection, using the original query, and retrieving for the user's attention a small amount of output, consisting of some of the highest scoring documents (those most similar to the query). These documents are examined by the user who identifies each retrieved item as either relevant (reflective of his information needs) or irrelevant. The stored representations of these judged documents are then used automatically to adjust the queries in such a way that terms present in the relevant documents are promoted, whereas terms occurring in documents designated as nonrelevant are demoted. In a somewhat simplified form, a typical query updating procedure is represented by the following equation:

$$\underline{q}_{i+1} = \underline{q}_i + \alpha \sum_{i=1}^{n_r} \underline{r}_i - \beta \sum_{i=1}^{n_s} \underline{s}_i \qquad (1)^*$$

where $\underline{q}_{i+1}$ represents the updated query vector, $\underline{q}_i$ the original query vector, $\underline{r}_i$ is one of n_r document representations identified as relevant, and $\underline{s}_i$ is one of n_s nonrelevant documents. [10,11]

Two major variants of the <u>relevance feedback</u> process described above are discussed in the following subsection. The simpler algorithm, called <u>positive feedback</u>, uses only the retrieved documents judged relevant to alter

---

*In the experimental system discussed here, terms having negative weights are deleted from the query (given zero weight).

the query (equation 1, $\beta = 0$). The second variant uses both the relevant and nonrelevant documents retrieved to modify the query (equation 1, $\beta > 0$). A study of the differences in performance between these two strategies reveals an important characteristic of the space of document representations, and leads to a proposal for several new techniques designed to improve retrieval in similar environments.

B) Positive and Negative Strategies

A typical positive query alteration process, where concepts may be added to the query but none are deleted is illustrated in the examples of Tables 1 and 2. An original query statement is given in Table 1, as well as the analyzed query "vector" in terms of a weighted term list. Following the addition of terms from document number 102, previously identified as relevant, the revised query vector retrieves two more relevant documents, numbers 80 and 81, with ranks 7 and 6, respectively (for retrieval purposes, documents are always ranked in decreasing order of similarity with the query). These two documents were originally assigned ranks 14 and 137 using the unaltered query vector.

Table 2 shows a typical retrieval output list, giving the ranks of retrieved documents in decreasing correlation order with the query. Relevant document numbers are identified by 'R'. The identified relevant document number 94 (originally retrieved with rank 14) is first used to update the query. This pulls up relevant documents 90 and 95 to ranks 7 and 10 respectively. When these two new documents are used in turn to update the query, additional relevant items are retrieved, until finally all five relevant documents are retrieved within the top twelve items following feedback run 3.

A typical recall-precision graph for positive feedback is shown in

Vector Type	Illustration
Initial Query Q 146	What informtion is available for dynamic response of airplanes to gusts or blasts in subsonic regime
Initial Query Vector	airplane 12, available 12, blast 12, dynamic 12 gust 12, information 12, regime 12, response 12 subsonic 12
Relevant Document 102 retrieved with rank 2 (partial vector)	gust 48, lift 48, oscillating 12, penetration 12 response 24, subsonic 12, sudden 12
Query Modified by Document 102	airplane 12, available 12, blast 12, dynamic 12 gust 60, information 12, lift 48, oscillating 12 penetration 12, regime 12, response 36, subsonic 24 sudden 12
Relevant Document 80 (improves from rank 14 to rank 7) (partial vector)	gust 24, lift 72, penetration 12, sudden 12
Relevant Document 81 (improves from rank 137 to rank 6) (partial vector)	lift 84, oscillating 12, sudden 12

Positive Feedback Illustration

Table 1

Query Q 147: Will forward or apex located controls be effective at low subsonic speeds.

Initial		Feedback Iterations		
		1	2	3
Rank	Doc	Doc	Doc	Doc
1	109	94R	94R	94R
2	60	81	95R	95R
3	121	195	90R	90R
4	192	123	195	195
5	193	80	81	91R
6	119	114	80	81
7	82	90R	114	80
8	24	193	193	111
9	86	122	123	114
10	123	95R	111	193
11	100	111	91R	93R
12	146	64	109	123
13	18	102	159	192
14	94R	109	103	109
15	167	82	192	159
16	125	103	82	155
17	163	78	93R	103
18	114	125	78	82
19	65	20	122	78
20	177	192	102	110
21	93R	124	64	122
22	90R	159	155	153
23	19	194	110	11
24	153	196	11	76
25	181	86	153	64
26	58	63	76	92
27	22	66	196	102
28	172	91R	20	152
29	200	10	194	161
30	64	93R	132	132
31	3	11	152	196
32	195	61	92	96
33	144	77	125	29
34	122	76	124	133
35	63	132	86	194
36	184	104	161	20
37	34	153	61	86
38	74	54	133	104
39	113	49	29	61
40	17	177	104	125
41	95R	144	63	176
42	75	67	96	124
43	67	29	83	121
44	140	60	77	83
76	91R	19	160	160

Positive Feedback Strategy for Query Q 147 Showing Improvements in Relevant Document Ranks

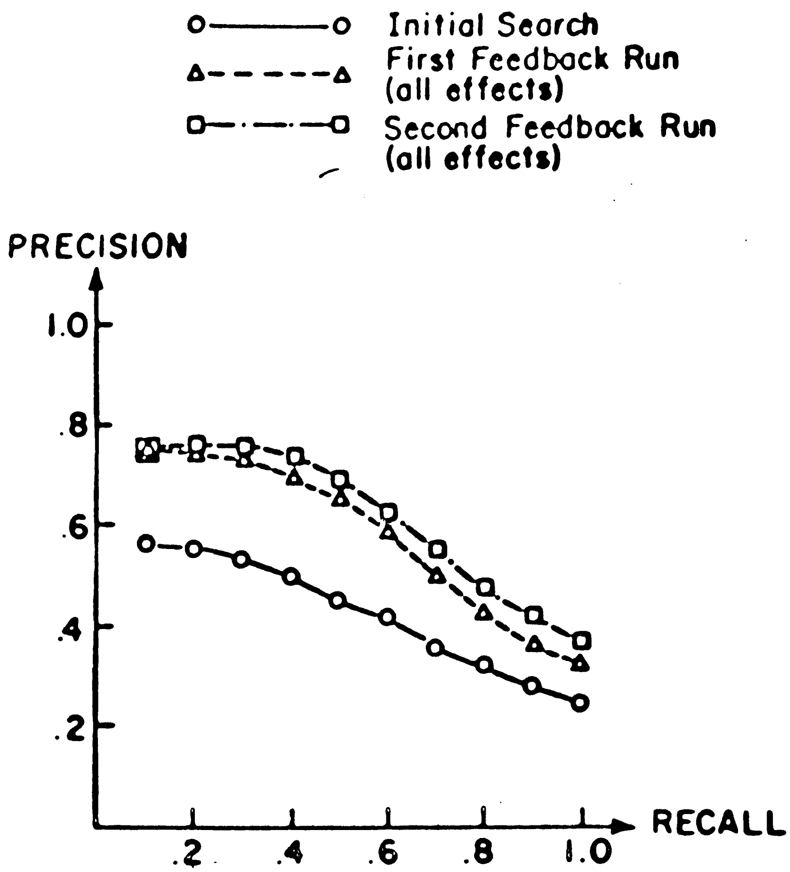
Fig. 1, giving averages over 42 queries for initial runs and two feedback iterations. The curves closest to the upper right-hand corner of the graph (where recall and precision are equal to 1) represent improved performance. It is seen that the updated queries produced by the feedback operations exhibit a precision average 10 to 20 percent better than the original queries for all recall points.

Although positive feedback is often successful, (for example for query 147 of Table 2), it fails to aid the retrieval performance of some queries. This occurs notably when no relevant items are retrieved, or when the retrieved relevant items are dissimilar. Performance may be improved even under these unfavorable conditions by a negative strategy that moves the queries away from those items specifically not wanted by the user.

An illustration of the potential usefulness of the negative feedback strategy is given in Table 3, showing positive and negative performance for query 3. Here the positive strategy produces no improvement on the first iteration, and then promotes relevant documents 57, 31, 4, 30 and 32, while demoting item 33 which goes down from rank 124 to 194. The negative strategy, on the other hand, retrieves documents 4, 57, 30, and 32 on the first iteration by moving away from the nonrelevant initially retrieved (documents 179, 42, 112, 39 and 117).

A thorough experimental comparison in a collection of 200 documents between positive and negative feedback strategies reveals the following differences in performance: [12]

a) the overall average differences in performance measured by the changes in rank of all documents strongly favor negative feedback, as is seen in Fig. 2;



Positive Feedback Performance

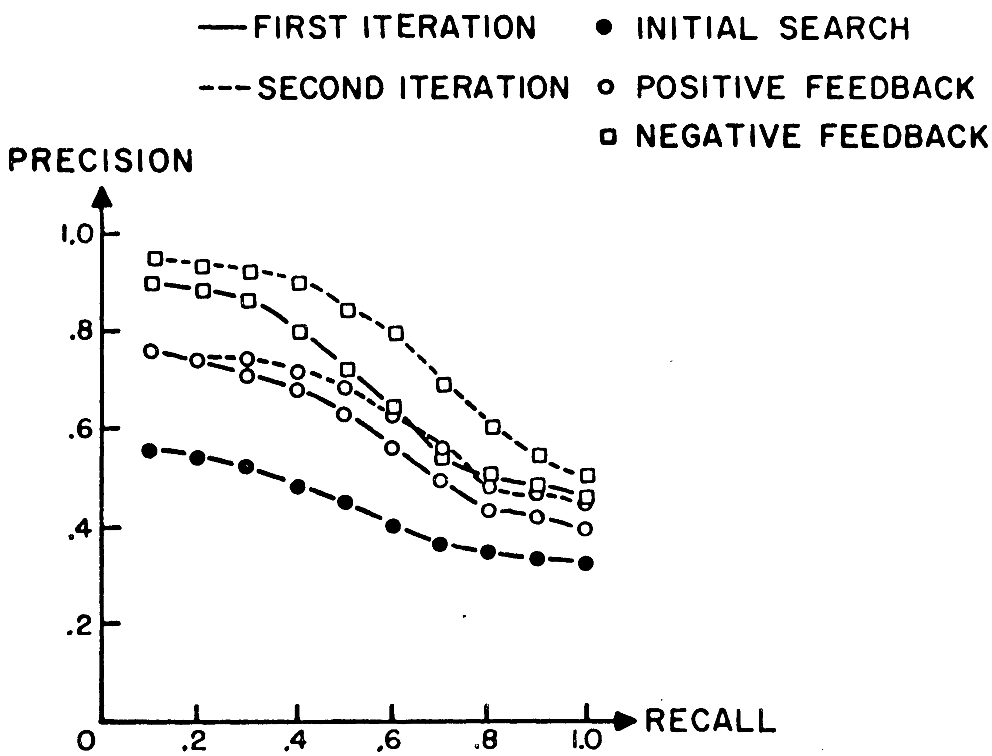
(200 documents — 42 queries)

Fig. 1

Rank	Positive Strategy Iteration 0	1	2	Rank	Negative Strategy Iteration 0	1	2
1	179	179	57R	1	179	4R	4R
2	42	42	31R	2	42	71	57R
3	112	112	179	3	112	57R	32R
4	39	39	4R	4	39	30R	30R
5	117	117	112	5	117	32R	31R
6	181	181	30R	6	181	182	200
7	57R	45	42	7	57R	152	189
8	45	57R	182	8	45	43	184
9	152	152	117	9	152	3	34
10	62	62	39	10	62	199	0
11	182	182	45	11	182	0	0
12	153	153	189	12	153	0	0
13	31R	31R	181	13	31R	0	0
14	43	43	0	14	43	0	0
15	116	116	0	15	116	0	0
16	0	0	32R	20	30R	0	0
20	30R	30R	0	22	0	31R	0
23	32R	32R	0	23	32R	0	0
25	4R	4R	0	25	4R	0	0
124	33R	33R	0	27	0	0	33R
194	0	0	33R	115	0	33R	0
				124	33R	0	0

Example of Improvements Obtainable with
Negative Feedback (Query 3)

Table 3



Comparison of Positive and Negative Feedback Performance

Fig. 2

b) the overall average differences measured by rank changes of unretrieved documents only are not statistically significant;

c) however, the variance in the performance is always greater for negative feedback than for positive, indicating that for some queries negative feedback is better and for other queries it is worse than positive feedback;

d) queries retrieving no relevant document in an initial search (which therefore cannot be updated on the first iteration by any positive strategy) are helped by the negative procedure;

e) on the average, the performance of queries that do retrieve relevant items in the initial search is not hindered by the negative strategy;

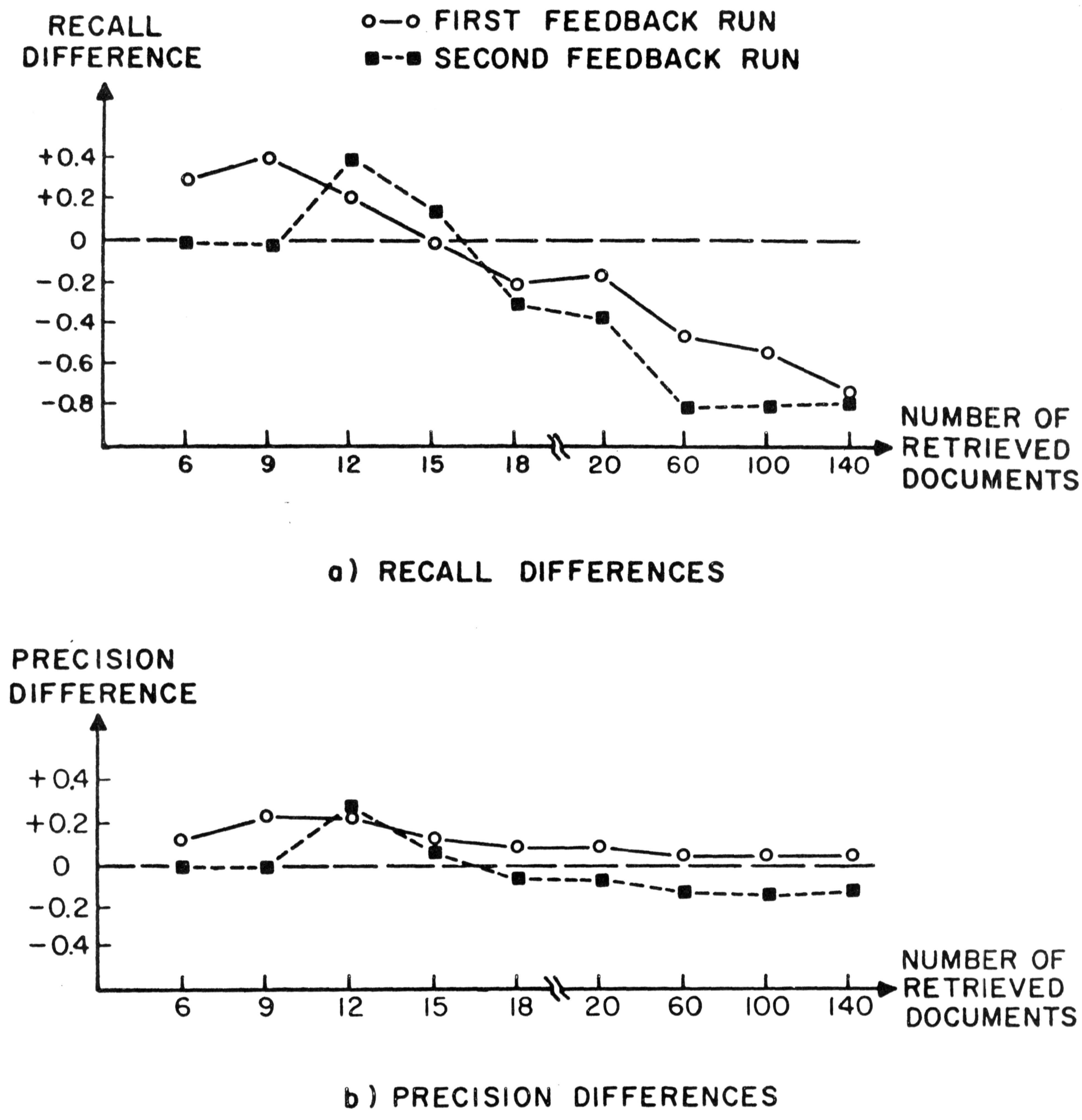
f) the negative strategy changes the query vector much more than the positive strategy (the average correlation between initial and updated queries is about 0.85 for the positive strategy, but only 0.50 for the negative strategy);

g) a plot of the average recall and precision differences between positive and negative feedback strategies is shown in Fig. 3; the following distinctions are apparent for the collection of 200 documents:

    i) if recall and precision are measured after the retrieval of about 15 documents, the negative strategy is better by about 5% in recall, and about 3% in precision;

    ii) after the retrieval of 20 to 30 items, the two strategies are about equal;

    iii) after 40 retrieved items, the positive strategy is better by about 10% in recall and 20% in precision.



Differences Between Negative and Positive Feedback
(averages 200 documents, 42 queries)

Fig. 3

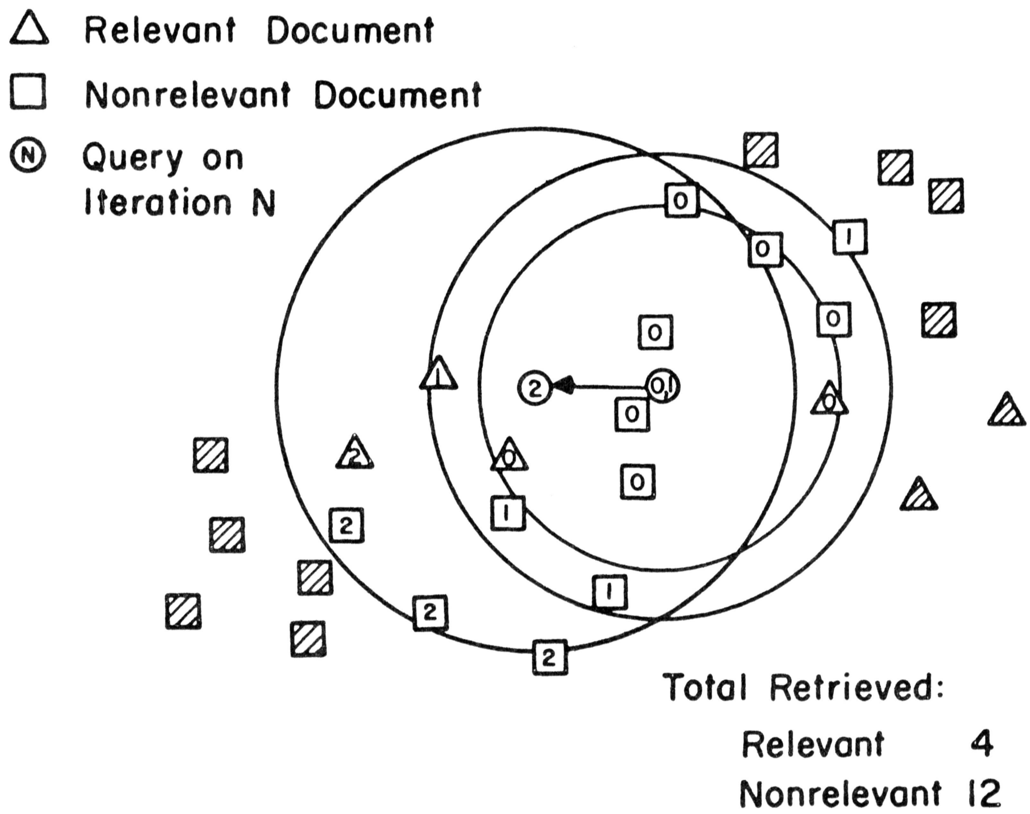
This indicates that negative feedback retrieves more relevant documents within the top 10% of the document collection than positive feedback, but that the relevant documents remaining in the lower 70% of the collection are assigned much lower ranks by the negative strategy than by the positive strategy. Thus, in general, the query produced by negative feedback is closer to some relevant documents and at the same time further from other relevant documents than the positive feedback query.

The evidence summarized above supports the following conclusion concerning the vector space of document representations:

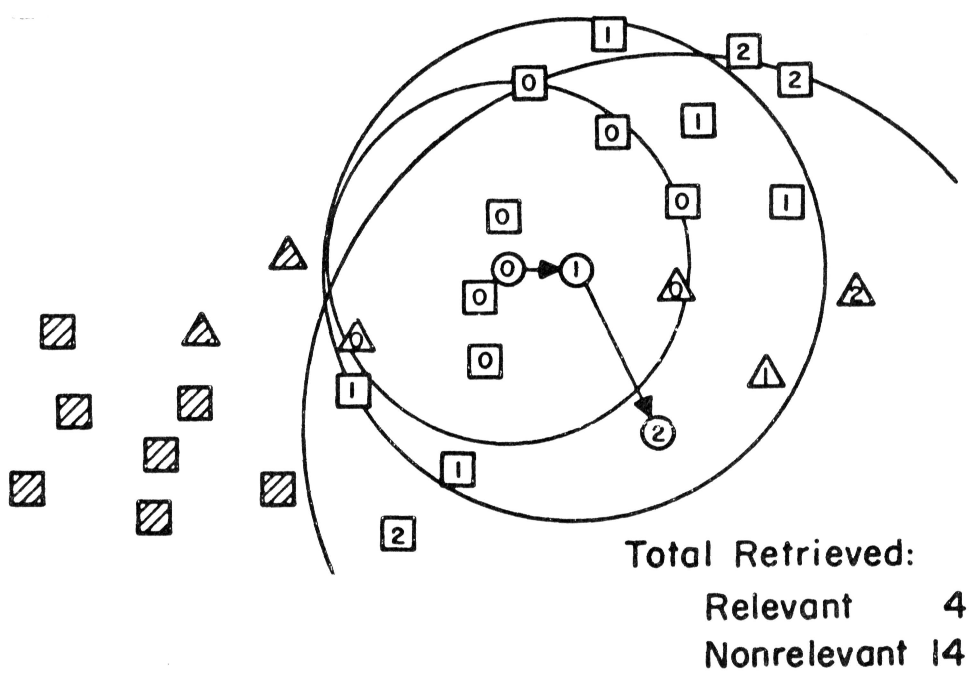
> the documents selected by the user as relevant to his query are often found in two or more distinct groups in the document vector space; and these groups are separated from one another by nonrelevant documents. For a significant number of queries, this separation of relevant document groups effectively prevents the retrieval of some relevant documents by conventional feedback strategies. [12]

Consider as an illustration the document space of Fig. 4. Here documents and queries are shown by points in the plane, and the distance between two points represents closeness of the corresponding subject matter.* Each query is assumed to retrieve all documents lying in a sphere around the query. The positive feedback illustration of Fig. 4(a) shows that the original query, identified by a circled zero, retrieves two relevant documents, one to the right of the query, and one to the left, as well as six nonrelevant documents.

*In actual fact each document or query must be represented by a t-dimensional vector, where t is the number of distinct allowable identifying terms; the two dimensional picture of Fig. 4 is thus a simplified analogy of the actual t-dimensional space.



a) Positive Feedback



b) Negative Feedback

Positive and Negative Feedback

Fig. 4

The expanded query, represented by a circled 1, retrieves additional documents, including a relevant one located to the left of the original circle. The new relevant item is used in a second updating operation to pull the query over to the left. The final updated query, represented by a circled 2, retrieves the three relevant items located on the left side of the picture; at the same time, two of the three relevant items on the right side are unfortunately lost.

The same document space and query are processed by a negative feedback strategy in Fig. 4(b). Here the three nonrelevant items located just left of center move the original query over to the right, away from the nonrelevant group. A new updating operation then moves the query further away from the original position in the general direction of the relevant document group on the right. The negative feedback strategy thus retrieves the relevant items on the right side of the picture, but "loses" two of the relevant ones on the left.

This type of retrieval behavior is illustrated for the example of query 9 in Table 4, where the positive strategy moves the query away from relevant document 82 when relevant document 116 is used for feedback. The negative strategy retrieves document 82 by using nonrelevant documents for feedback, but simultaneously, the query is moved away from document 116.

If the high retrieval performance sometimes achieved by human subject experts is to be duplicated in an automatic environment, new retrieval strategies must be specifically designed to select separated groups of relevant documents. Each of the techniques proposed in the following sections exhibit some advantages over conventional retrieval methods in the type of document vector space depicted in Fig. 4.

Rank	Positive Strategy Iteration			Rank	Negative Strategy Iteration		
	0	1	2		0	1	2
1	179	179	116R	1	179	25	25
2	112	112	179	2	112	71	71
3	39	39	62	3	39	41	41
4	42	42	102	4	42	64	3
5	181	181	181	5	181	3	98
6	45	45	39	6	45	85	178
7	62	62	42	7	62	88	82R
8	116R	116R	117	8	116R	23	160
9	97	97	3	9	97	101	64
10	188	188	45	10	188	17	101
11	31	31	115	11	31	82R	0
12	57	57	2	12	57	0	0
13	117	117	158	13	117	116R	0
14	2	2	0	14	2	0	0
15	25	25	0	15	25	0	116R
33	82R	82R	0	33	82R	0	0
54	0	0	82R				

Positive and Negative Feedback Strategies

Query 9 with Separate Relevant

Document Clusters

Table 4

C) Selective Negative Feedback

The discussion in the previous section indicates that the use of retrieved nonrelevant documents for feedback often further lowers the ranks assigned to low-ranking relevant documents. This suggests that a more selective process might be devised in applying the negative strategy in order to improve overall performance. Under the present procedure, <u>all</u> terms included in the identified set of nonrelevant documents are automatically deleted from the query or reduced in weight. This process may lead to the effective loss of important query terms, particularly terms which may have more than one meaning in the document collection. The illustration of Table 5, covering a query dealing with data sets, shows that the crucial term "data set" is eventually deleted from the query.*

Two selective negative procedures are proposed. The first one, illustrated in Table 6, consists in assigning negative weights to terms extracted from nonrelevant documents while leaving the original query terms unchanged. Thus, in the example, the term "data set" is still present in the final query, but the related terms derived from the nonrelevant document set which suggest "sets of data" are assigned negative weights.

The other selective procedure, illustrated in Table 7, makes use of a synonym dictionary, or thesaurus (or alternatively an associative indexing procedure) to provide for each term a set of related terms. These related dictionary terms are first added to the query statement, after which the terms obtained from the nonrelevant documents are subtracted out. In the

---

*"Data set" is an ambiguous term denoting both a communications device (the meaning assumed in the query), and a "set of data" (the meaning derived from the nonrelevant document set).

Type of Vector	Illustration
a) Original Query	Please give specification for all currently available data sets.
b) Initial Query Vector	available 12, current 12, data set 12, specification 12
c) Sum of Retrieved Nonrelevant Documents	access 48, data set 60, file 24, list 24, structure 84
d) Standard Negative Feedback Result	available 12, current 12, specification 12

Example of Inadequate Negative Feedback

Table 5

Type of Vector	Illustration
a) Initial Query Vector	available 12, current 12, data set 12, specification 12
b) Sum of Retrieved Nonrelevant Documents	access 48, data set 60, file 24, list 24, structure 84
c) Negative Context Vector (query concepts deleted)	access 48, file 24, list 24, structure 84
d) Selective Negative Feedback Result (b-c)	access -48, available 12, current 12, data set 12, file -24, list -24, specification 12, structure -84

Selective Negative Weighting

Table 6

Type of Vector	Illustration
a) Initial Query Vector	available 12, current 12, data set 12, specification 12
b) Concepts Related to "data set" with Correlation Strength	structure (79), access (77), interface (58), line (52), file (50), sort (50), retrieval (49), list (47), transmission (30), band-width (28)
c) Related Concept Vector (top 5 concepts with weight of 24)	access 24, file 24, interface 24, line 24, structure 24
d) Query Vector with Related Concepts	access 24, available 12, current 12, data set 12 file 24, interface 24, line 24, specification 12 structure 24
e) Sum of Retrieved Nonrelevant Documents	access 48, data set 60, file 24, list 24, structure 84
f) Feedback Result with Related Concepts	available 12, current 12, interface 24, line 24 specification 12
g) Related Concepts and Selective Negative Weighting	access -24, available 12, current 12, data set 12 interface 24, line 24, list -24, specification 12 structure -60

Negative Feedback with Related Concepts

Table 7

example of Table 7, the thesaurus provides contextual information for the term "data set" used both in the sense of a communications device and in the sense of "sets of data"; the latter context in then eliminated by the negative feedback operation.

Both of the suggested selective negative feedback strategies are intended to retain in the query the terms that might lead to the eventual retrieval of relevant documents, separated from the query by nonrelevant documents. Since the intervening nonrelevant documents are also retrieved, it remains to be seen whether these strategies improve performance for a significant number of queries.

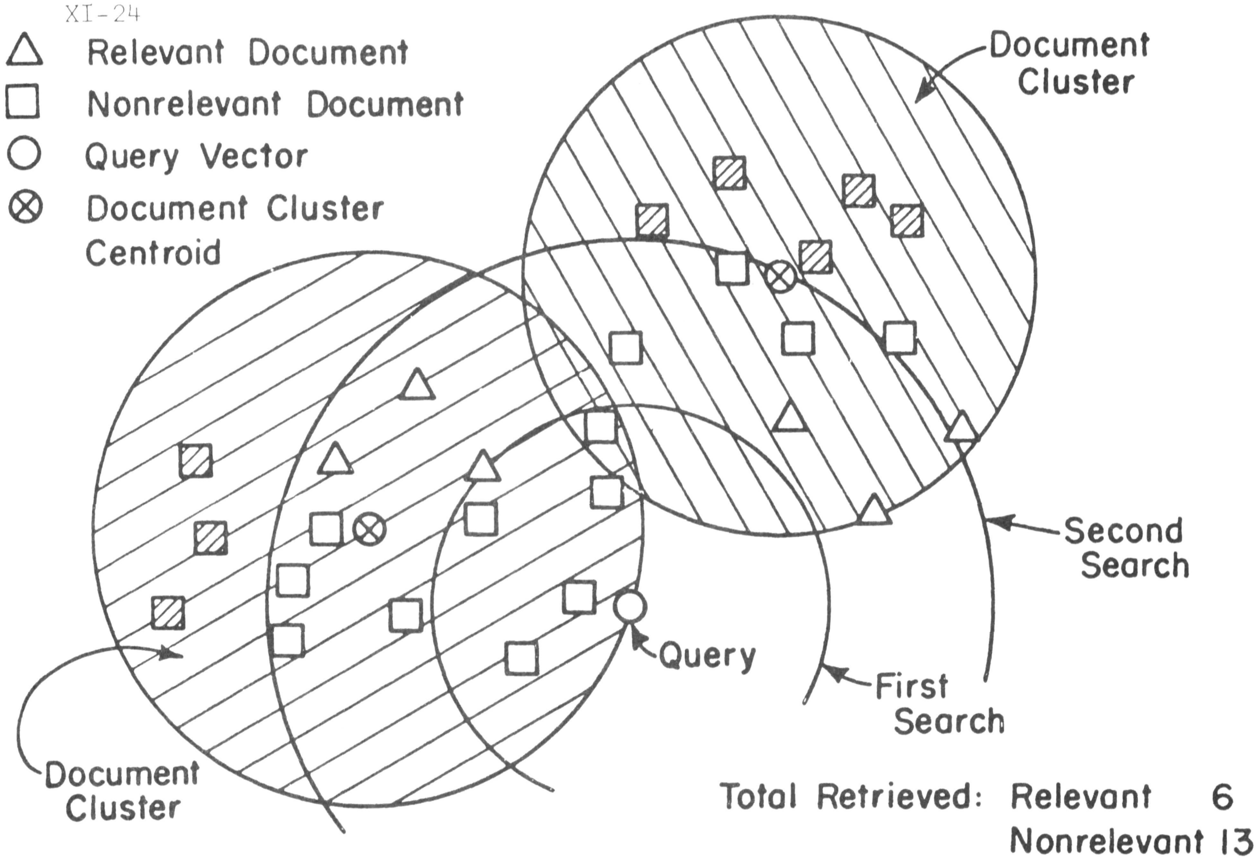
3. Document Clustering

When relevant documents are separated from each other by nonrelevant documents, no conceivable strategy which uses a single query to search the complete document space can identify the separate sets of relevant items, while properly rejecting the nonrelevant documents located between them. A multiple query set might then be used, instead of a single query, in such a way that each "subquery" searches a distinct part of the document space. This, in turn, suggests that the documents in a collection be grouped into "clusters" of similar documents, and that each document cluster be searched separately. It may then be easier to discriminate between relevant and nonrelevant items within a given document cluster than in the document collection as a whole.

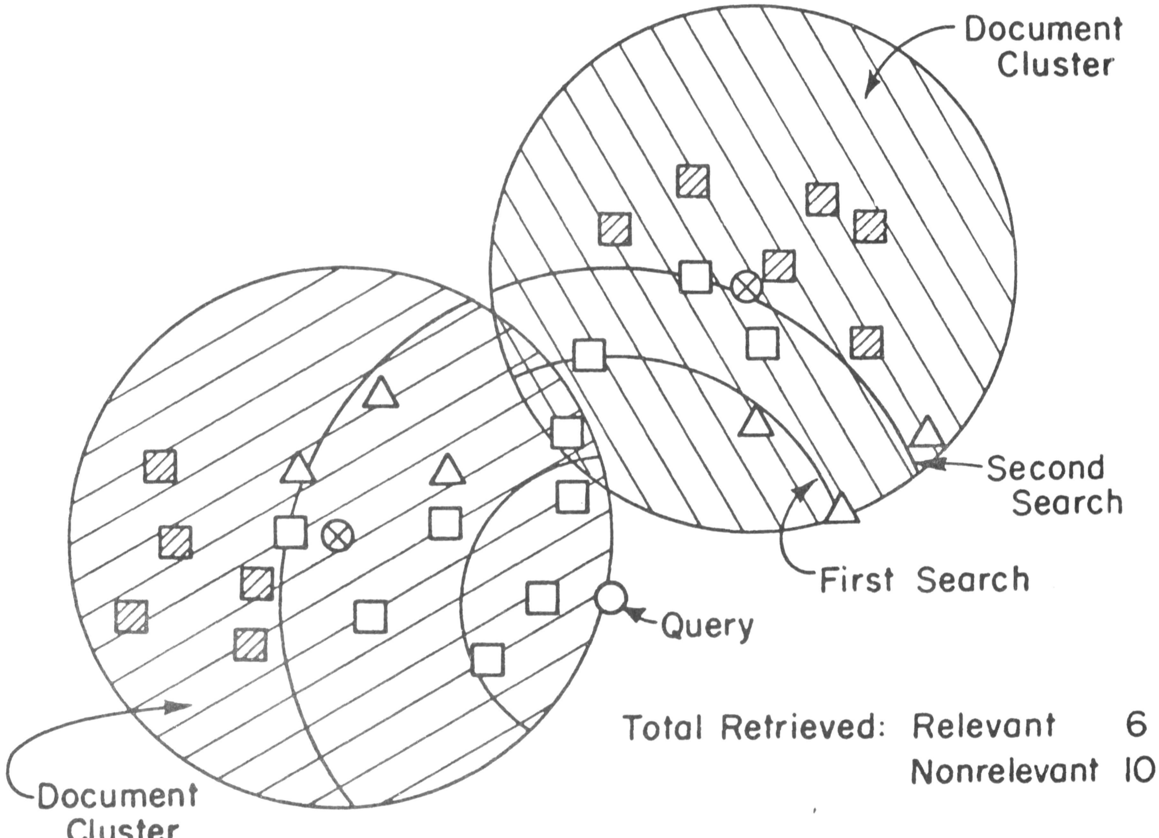
Several methods exist for automatically producing document clusters in such a way that items sufficiently similar to each other are placed in the same group. [13,14,15] Such clustered document colllections can conveniently be used in a retrieval environment to reduce the search to a small portion of the document space by comparing the query against only those documents located within a specified subset of clusters. [16,17]

Cluster searching can be performed in several distinct ways. The combined cluster search of Fig. 5(a) operates in such a way that all documents in the cluster set to be searched are ranked according to their distance from the query. Thus, the initial query of Fig. 5(a) first retrieves six documents all located in the left-hand cluster, including one relevant item; a second search operation is then used to retrieve 13 more items. Alternatively, a separate cluster search can be performed, as shown in the example of Fig. 5(b), where the documents are ranked separately within each cluster relative to other documents in the same cluster. The query then retrieves the highest ranking documents from each cluster searched. In the illustration the six relevant items are more efficiently retrieved in the separate cluster search than in the combined search, since the number of unwanted items obtained is only ten for the separate compared with thirteen for the combined strategy.

The cluster searches shown in Fig. 5 compare all documents in all selected clusters with the same initial query. In order to generate a distinct query for each cluster to be searched, it is possible to combine the notion of the cluster search with the query alteration methods used in relevance feedback. Specifically, a query alteration procedure can be utilized in which retrieved documents from separate clusters generate distinct queries, each of which operates within a distinct document cluster. The cluster feedback process illustrated in Fig. 6(a) is a partial search method of this type.



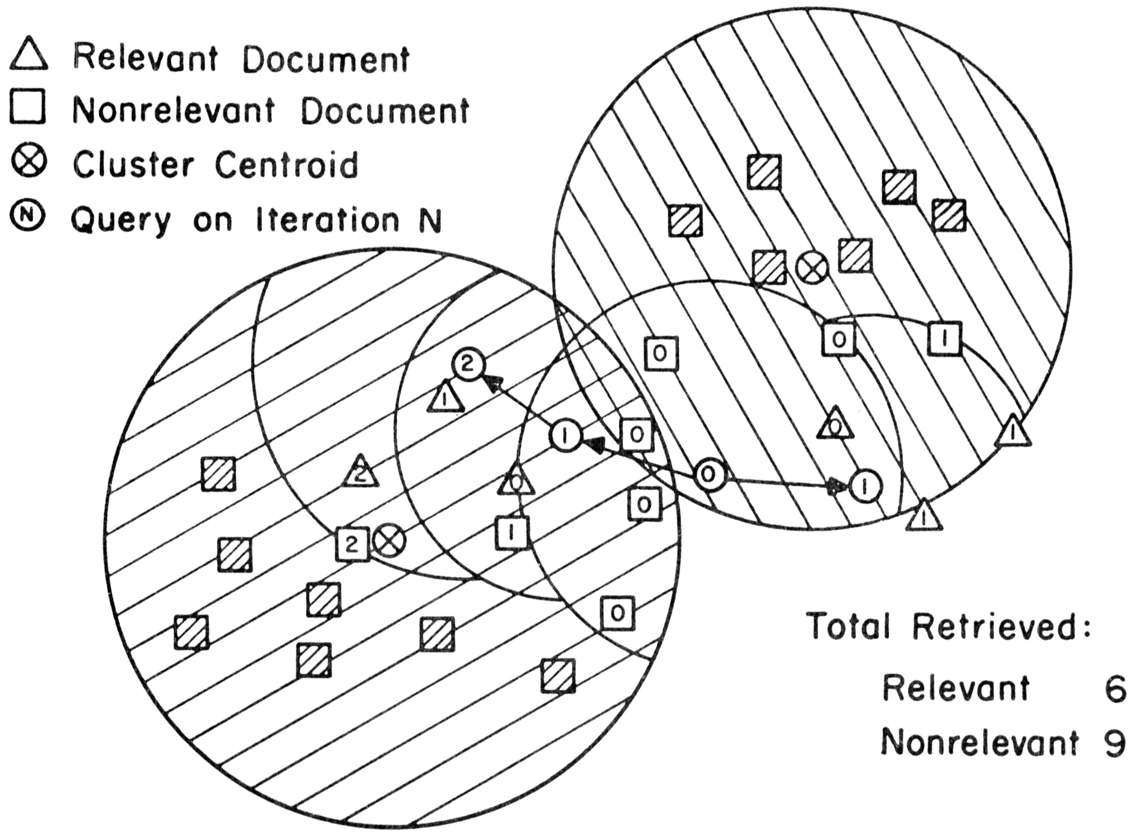
a) Combined Cluster Search



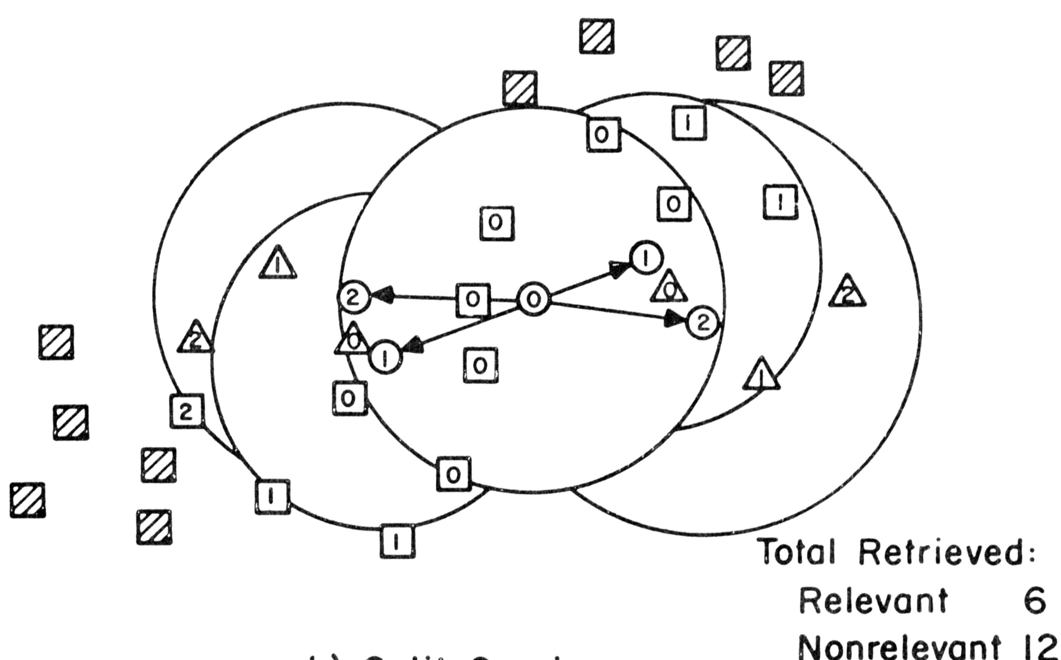
b) Separate Cluster Search

Single Query Cluster Searches

Fig. 5



a) Cluster Feedback



b) Split Queries

Multiple Query Searches

Fig. 6

The following principal steps are required:

a) the original query (designated in Fig. 6 by a circled zero) is first compared with the centers (centroids) of all document clusters;

b) the clusters whose centroids are closest to the original query are then selected, and the individual document vectors within the selected clusters are compared to the query;

c) relevance judgments are obtained for those documents found to be closest to the query;

d) a new query is constructed for each cluster, using the original query as well as relevant (or nonrelevant) documents from that particular cluster only — in the example of Fig. 6(a) the original query (circled 0) leads to two distinct new queries (circled 1) obtained by using the relevant documents from the right-hand and from the left-hand cluster, respectively;

e) each new query is now matched only against the documents in its own cluster, and only the documents retrieved by a given query are used to modify that query in further feedback iterations;

f)* all documents retrieved from all selected clusters may be used to generate from the initial query a new centroid search query to select additional clusters to be searched;

g)* since more than one query is generated, some means of discarding queries that seem unlikely to retrieve additional relevant items would be desirable. Several possible criteria for eliminating such queries are suggested elsewhere [12]

In the illustration of Fig. 6(a), only nine nonrelevant items are retrieved together with the six relevant.

*Steps f and g are not illustrated in Fig. 6(a).

The cluster feedback algorithm described above is equally feasible in combination with a technique called request clustering. This suggested alternative to document clustering assumes that documents formerly retrieved in answer to similar previous queries should be considered in processing a new query. In step a) the request cluster feedback algorithm would compare the new query to the centroids of clusters of previous queries submitted to the systems. The clusters of documents searched in steps b), c), d), and e) would then include documents judged relevant to the queries in the query clusters nearest the new query. Request clustering allows documents which are adjacent in the document space to be placed into different clusters and nonadjacent documents to be placed into the same cluster. This may turn out to be advantageous in an environment containing separated groups of relevant documents.

If the cluster search is to operate successfully, the retrieval problem (that is, the separation of relevant from nonrelevant) within each cluster must be simpler than the problem in the space as a whole; furthermore, the cluster selection method must pick few unproductive clusters to be searched. Should separated clusters of relevant documents still occur within one or more of the clusters, it may be necessary to construct multiple queries all of which search the same set of documents.

A "query splitting" process designed to do this has been investigated with some success on a small test collection. [18] A query is split into two subqueries whenever the correlation between two relevant documents previously retrieved is small compared to the average inter-document correlation between the first five retrieved documents. An alternative strategy might be to split the query whenever a retrieved nonrelevant document is located between two retrieved relevant ones; that is, relevant documents r and v are used to generate distinct (split) queries whenever, for some nonrelevant item n

$$\text{correlation } (\underline{n},\underline{v}) > \text{correlation } (\underline{r},\underline{v}),$$
$$\text{and correlation } (\underline{n},\underline{r}) > \text{correlation } (\underline{r},\underline{v}).$$

An illustration of the query splitting concept is shown in Fig. 6(b). The original query (circled 0) first retrieves two relevant items, one to the right and one to the left, whose interdocument correlation is small compared with the correlation of each relevant item to one of the nonrelevant in the middle. This leads to a split of the initial query into two pieces (circled 1), and to two additional queries (circled 2) after one more iteration. The subqueries on the right retrieve the right-most relevant, and the left subqueries handle the relevant on the left.

Both of the multiple query strategies illustrated in Fig. 6 remain to be tried out in a realistic document environment.

4. Document Space Modification

The feedback procedures described up to now all produce a modification of the query space in such a way that queries are moved close to certain identified relevant documents, or away from identified nonrelevant ones. The strategies suggested in this section attach the problem directly by permanently changing the document vector space. Specifically, the vector representations of documents judged relevant to a query are moved closer to the query vector. This strategy is more radical than query modification, since it implies that the queries are more fundamental as subject indicators than the original document identifying terms.

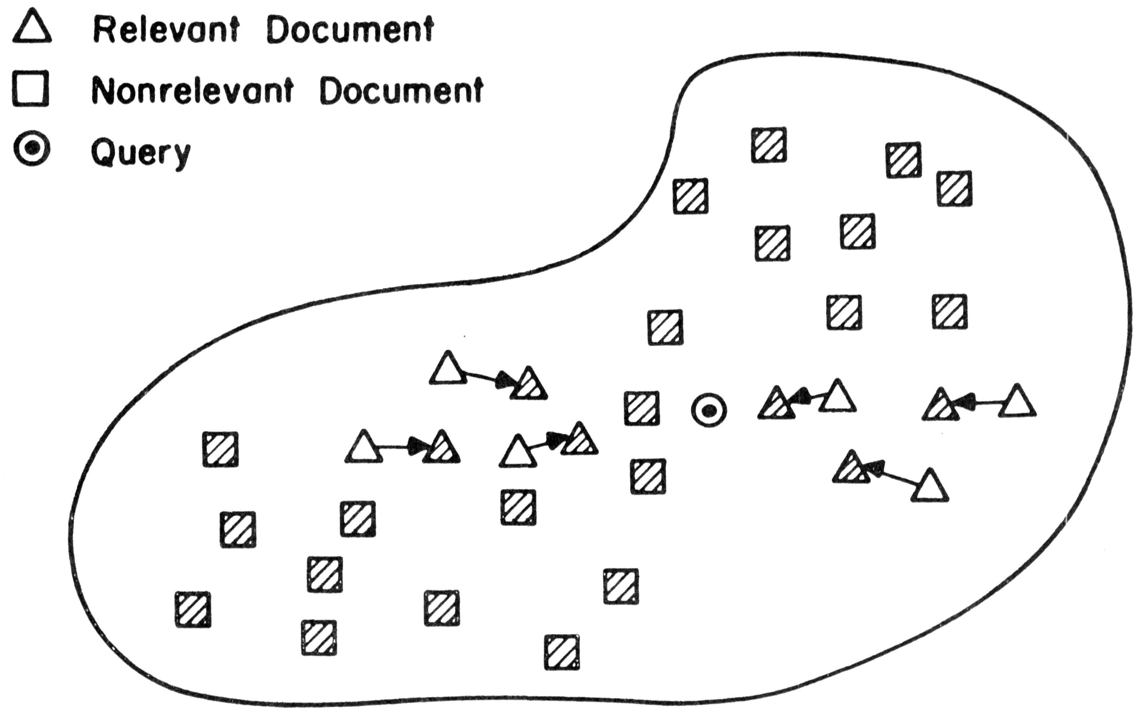
Two different document space modification methods are illustrated in Fig. 7. In the first one, (Fig. 7(a)), the previously identified relevant documents are modified by addition of query terms as follows:

$$\underline{d}_{i+1} = (1-\alpha)\, \underline{d}_i + \alpha \underline{q}_o \qquad (2)$$

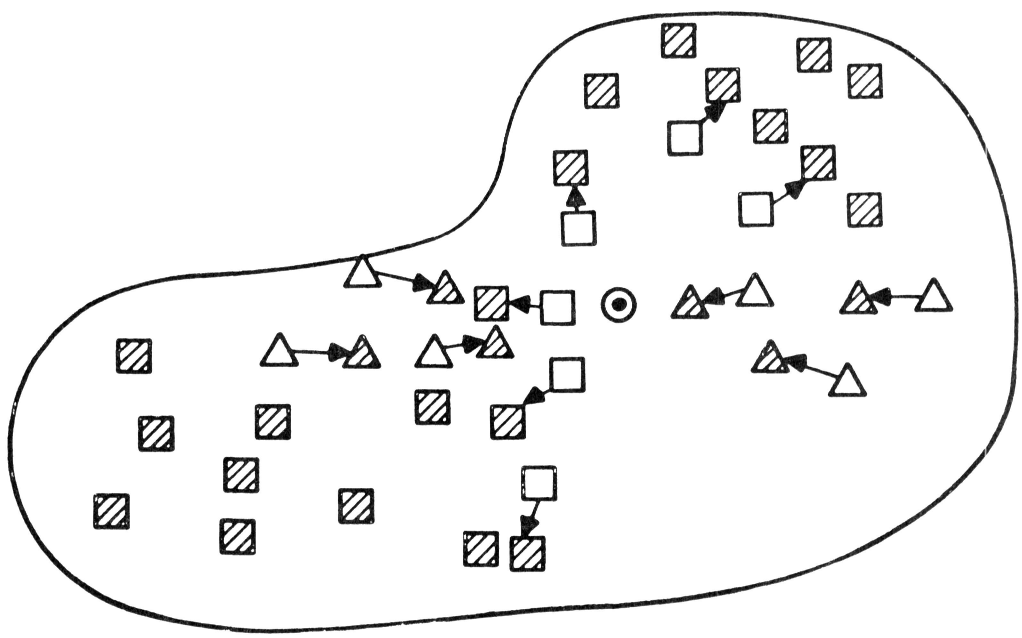
where $\underline{d}_{i+1}$ is the modified document, $\underline{d}_i$ the original document, and $\underline{q}_o$ the original query. A test was performed on this document modification process using a collection of 425 documents in aerodynamics, and a set of 125 queries to effect the space modification. A new set of 30 additional queries not previously used for space modification was then processed with the modified document space, and improvements in both recall and precision of 30 to 15 percent were detected, compared with the use of these same queries in conjunction with the original, unmodified document space. These relatively large improvements appear to indicate that new customers whose relevance criteria play no part in the space modification profit directly from the query-document associations derived from previous system users.

A second document space modification procedure illustrated in Fig. 7(b) is based on strategies previously tried in adaptive pattern recognition. [12] The basic idea is to pair each relevant document retrieved with a nonrelevant document not previously modified; if the nonrelevant item happens to be located closer to the query than the relevant one, an interchange procedure is used to move the relevant forward (closer to the query) and the nonrelevant backward (away from the query). More formally, the process is as follows:

a) if for <u>all</u> $\underline{d}_i$, such that $\underline{d}_i$ is relevant to query q_o,



a) Relevant Document Modification



b) Adaptive Document Modification

Document Space Modification

Fig. 7

and for all $\underline{d}_j$ such that $\underline{d}_j$ is nonrelevant

$$\text{correlation } (\underline{d}_i, \underline{q}_o) > \text{correlation } (\underline{d}_j, \underline{q}_o) + \theta$$

no adjustment is made;

b) otherwise, each vector $\underline{d}_i$ denoting relevant document i is processed in order with $\underline{q}_o$; if there is a document k, not yet adjusted by $\underline{q}_o$, and $\underline{d}_k$ is not relevant to $\underline{q}_o$, and

$$\text{correlation } (\underline{d}_k, \underline{q}_o) + \theta \geq \text{correlation } (\underline{d}_i, \underline{q}_o)$$

then

$$\underline{d}_i' = (1 - \alpha)\ \underline{d}_i + \alpha \underline{q}_o$$

$$\underline{d}_k' = (1 - \alpha)\ \underline{d}_k - \alpha \underline{q}_o$$

where $\underline{d}_k$ is that previously unmodified nonrelevant item having the highest correlation with $\underline{q}_o$, and $\underline{d}_i'$ and $\underline{d}_k'$ are the new adjusted document vectors;

c) if no nonrelevant document k exists which has not been previously adjusted, the modification of the relevant item $\underline{d}_i$ is still performed.

This procedure is intended to produce a document space which groups all the relevant items around the corresponding queries, while moving the nonrelevant items further away. The space alteration is moreover controlled in the sense that a different nonrelevant item is subtracted out each time.

The basic differences between the two suggested modification procedures is similar to the distinction between positive and negative feedback. The first technique adjusts only relevant documents, while the second alters both relevant and nonrelevant documents. A comparison of the two strategies in the 425 document collection shows the superiority of the second method when α (modifier in equation (2)) is relatively small (.05 to .10). The advantage in precision of 'negative modification' over 'positive modification' is greatest at relatively low recall levels, reaching 4% at 20% recall.

Both document space modification algorithms can easily be combined with the relevance feedback methods in an operating retrieval system to provide a continual adjustment of the document identifiers in accordance with the user's expectations. The simplest procedure consists in modifying only the retrieved documents. This modification could take place after the relevance judgments are rendered by the user. Only the vector representation of the user's initial query would be used to alter the document representations. The proposed combined query and document space modification has not yet been tested in a retrieval environment.

5. Conclusion

Several search and retrieval strategies are described in this study that use feedback information supplied by the user during the retrieval process to modify the query or document spaces. In each case, the space modification is intended to increase the correlation between queries and relevant documents, while decreasing the query correlation with nonrelevant items. Experimental evidence indicates that the improvements in retrieval effectiveness obtainable with these heuristic search strategies are much larger than the improvements immediately derivable from the more formal deterministic methods based on better document and query analyses and more sophisticated linguistic normalization tools.

References

[1] C. W. Cleverdon and E. M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 2 - Test Results, Aslib-Cranfield Research Project, Cranfield College of Aeronautics, 1966.

[2] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1968.

[3] F. W. Lancaster, Evaluating the Operating Efficiency of Medlars, Final Report, National Library of Medicine, January 1968.

[4] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Book Co., New York, 1968.

[5] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System — An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[6] J. L. Bennett, On-line Computer Aids for the Indexer, Proceedings of the 1968 meeting of the American Society for Information Science, Columbus, Ohio, October 1968.

[7] M. E. Lesk and G. Salton, Interactive Search and Retrieval Methods Using Automatic Information Displays, Scientific Report No. ISR-14 to the National Science Foundation, Section IX, Department of Computer Science, Cornell University, October 1968, to be presented at the 1969 AFIPS Spring Joint Computer Conference, May 1969.

[8] H. Borko, Utilization of On-line Interactive Displays, in Information Systems Science and Technology, D. Walker, editor, Thompson Book Co., Washington, D. C., 1967.

[9] M. E. Lesk and G. Salton, Relevance Assessments and Retrieval System Evaluation, Scientific Report No. ISR-14 to the National Science Foundation, Section III, Department of Computer Science, Cornell University, October 1968.

[10] J. J. Rocchio and G. Salton, Information Search Optimization and Interactive Retrieval Techniques. Proceedings of the AFIPS Fall Joint Computer Conference, Spartan Book Co., 1965.

[11] G. Salton, Search and Retrieval Experiments in Real-Time Information Retrieval, Proceedings IFIP Congress 68, Edinburgh, 1968.

[12] Eleanor Ide, Relevance Feedback in an Automatic Document Retrieval System, Master's Thesis, Cornell University, Report No. ISR-15 to the National Science Foundation, Department of Computer Science, Cornell University, January 1969.

References (contd)

[13] R. E. Bonner, On Some Clustering Techniques, IBM Journal of Research and Development, Vol. 8, No. 1, January 1964.

[14] H. Borko and M. D. Bernick, Automatic Document Classification, Journal of the ACM, Vol. 10, No. 2, April 1963.

[15] R. M. Needham and K. Sparck Jones, Keywords and Clumps, Journal of Documentation, Vol. 20, No. 1, March 1964.

[16] J. J. Rocchio, Jr., Document Retrieval Systems — Optimization and Evaluation, Harvard University Doctoral Thesis, Report No. ISR-10 to the National Science Foundation, Harvard Computation Laboratory, March 1966.

[17] G. Salton, Search Strategy and the Optimization of Retrieval Effectiveness, Proceedings of the FID/IFIP Conference on Mechanized Information Storage, Retrieval and Dissemination, North Holland Publishing Co., 1968.

[18] A. Borodin, L. Kerr and F. Lewis, Query Splitting in Relevance Feedback Systems, Scientific Report No. ISR-14 to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, October 1968.

[19] T. L. Brauen, R. C. Holt, and T. R. Wilcox, Document Indexing Based on Relevance Feedback, Scientific Report No. ISR-14 to the National Science Foundation, Section XI, Department of Computer Science, Cornell University, October 1968.