ED 048 913	LI 002 722
TITLE	User Feedback Procedures; Part III of Scientific Report No. ISR-18, Information Storage and Retrieval...
INSTITUTION	Cornell Univ., Ithaca, N.Y. Dept. of Computer Science.
SPONS AGENCY	National Library of Medicine (DHEW), Bethesda, Md.; National Science Foundation, Washington, D.C.
REPORT NO	ISR-18 [Part III]
PUB DATE	Oct 70
NOTE	111p.; Part of LI 002 719
EDRS PRICE	EDRS Price MF-$0.65 HC-$6.58
DESCRIPTORS	Automation, *Feedback, *Information Retrieval, *Relevance (Information Retrieval), *Search Strategies, *Use Studies
IDENTIFIERS	On Line Retrieval Systems, *Saltons Magical Automatic Retriever of Texts, SMART

ABSTRACT

Part Three of this five part report on Salton's Magical Automatic Retriever of Texts (SMART) project contains four papers. The first: "Variations on the Query Splitting Technique with Relevance Feedback" by T. P. Baker discusses some experiments in relevance feedback performed with variations on the technique of query splitting. The results indicate that these variations, as tested, offer no significant improvement over previously tried methods of query splitting. Pa er two: "Effectiveness of Feedback Strategies on Collections of Differing Generality" by B. Capps and M. Yin evaluates the comparative effectiveness of several feedback strategies on the Cranfield 200 and Cranfield 400 Collections. Results are assessed from both the user and the system viewpoint; some strategies appeared equally effective on both collections. The third paper: "Selective Negative Feedback Methods" by M. Kerchner deals with experiments performed with several new methods of using nonrelevant retrieval documents to modify queries which retrieved no relevant material in the first N documents retrieved. Paper four: "The Use of Past Relevance Decisions in Relevance Feedback" by L. Paavola investigates some possibilities for exploiting the potential similarity among documents judged relevant to the same query in relevance feedback. (See also LI 002 719-LI 002 721 and LI 002 723-LI 002 724.) (NH)

Department of Computer Science

Cornell University

Ithaca, New York 14850

User Feedback Procedures

Part III

Scientific Report No. ISR-18

INFORMATION STORAGE AND RETRIEVAL

to

The National Science Foundation

and to

The National Library of Medicine

Reports on Analysis, Dictionary Construction, User Feedback, Clustering, and On-Line Retrieval

Ithaca, New York

October 1970

Gerard Salton

Project Director

SMART Project Staff

Robert Crawford
Barbara Galaska
Eileen Gudat
Marcia Kerchner
Ellen Lundell
Robert Peck
Jacob Razon
Gerard Salton
Donna Williamson
Robert Williamson
Steven Worona
Joel Zumoff

TABLE OF CONTENTS

PART ONE

AUTOMATIC CONTENT ANALYSIS

Available as LI 002 720

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

Page

IV. continued

PART TWO

AUTOMATIC DICTIONARY CONSTRUCTION

Available as LI 002 721

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

Page

PART THREE

USER FEEDBACK PROCEDURES

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

Page

PART FOUR

CLUSTERING METHODS

Available as LI 002 723

TABLE OF CONTENTS (continued)

PART FIVE

ON-LINE RETRIEVAL SYSTEM DESIGN

Available as LI 002 724

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

Summary

The present report is the eighteenth in a series describing research in automatic information storage and retrieval conducted by the Department of Computer Science at Cornell University. The report covering work carried out by the SMART project for approximately one year (summer 1969 to summer 1970) is separated into five parts: automatic content analysis (Sections I to IV), automatic dictionary construction (Sections V to VII), user feedback procedures (Sections VIII to XI), document and query clustering methods (Sections XII and XIII), and SMART systems design for on-line operations (Sections XIV and XV).

Most recipients of SMART project reports will experience a gap in the series of scientific reports received to date. Report ISR-17, consisting of a master's thesis by Thomas Brauen entitled "Document Vector Modification in On-line Information Retrieval Systems" was prepared for limited distribution during the fall of 1969. Report ISR-17 is available from the National Technical Information Service in Springfield, Virginia 22151, under order number PB 186-135.

The SMART system continues to operate in a batch processing mode on the IBM 360 model 65 system at Cornell University. The standard processing mode is eventually to be replaced by an on-line system using time-shared console devices for input and output. The overall design for such an on-line version of SMART has been completed, and is described in Section XIV of the present report. While awaiting the time-sharing implementation of the system, new retrieval experiments have been performed using larger document collections within the existing system. Attempts to compare the performance of several collections of different sizes must take into account the collection "generality". A study of this problem is made in Section II of the present report. Of special interest may also be the new procedures for the automatic recognition of "common" words in English texts (Section VI), and the automatic construction of thesauruses and dictionaries for use in an automatic language analysis system (Section VII). Finally, a new inexpensive method of document classification and term grouping is described and evaluated in Section XII of the present report.

Sections I to IV cover experiments in automatic content analysis and automatic indexing. Section I by S. F. Weiss contains the results of experiments, using statistical and syntactic procedures for the automatic recognition of phrases in written texts. It is shown once again that because of the relative heterogeneity of most document collections, and the sparseness of the document space, phrases are not normally needed for content identification.

In Section II by G. Salton, the "generality" problem is examined which arises when two or more distinct collections are compared in a retrieval environment. It is shown that proportionately fewer nonrelevant items tend to be retrieved when larger collections (of low generality) are used, than when small, high generality collections serve for evaluation purposes. The systems viewpoint thus normally favors the larger, low generality output, whereas the user viewpoint prefers the performance of the smaller collection.

The effectiveness of bibliographic citations for content analysis purposes is examined in Section III by G. Salton. It is shown that in some situations when the citation space is reasonably dense, the use of citations attached to documents is even more effective than the use of standard keywords or descriptors. In any case, citations should be added to the normal descriptors whenever they happen to be available.

In the last section of Part 1, certain template analysis methods are applied to the automatic resolution of ambiguous constructions (Section IV by S. F. Weiss). It is shown that a set of contextual rules can be constructed by a semi-automatic learning process, which will eventually lead to an automatic recognition of over ninety percent of the existing textual ambiguities.

Part 2, consisting of Sections V, VI and VII covers procedures for the automatic construction of dictionaries and thesauruses useful in text analysis systems. In Section V by D. Bergmark it is shown that word stem methods using large common word lists are more effective in an information retrieval environment that some manually constructed thesauruses, even though the latter also include synonym recognition facilities.

A new model for the automatic determination of "common" words (which are not to be used for content identification) is proposed and evaluated in Section VI by K. Bonwit and J. Aste-Tonsmann. The resulting process can be incorporated into fully automatic dictionary construction systems. The complete thesaurus construction problem is reviewed in Section VII by G. Salton, and the effectiveness of a variety of automatic dictionaries is evaluated.

Part 3, consisting of Sections VIII through XI, deals with a number of refinements of the normal relevance feedback process which has been examined in a number of previous reports in this series. In Section VIII by T. P. Baker, a query splitting process is evaluated in which input queries are split into two or more parts during feedback whenever the relevant documents identified by the user are separated by one or more nonrelevant ones.

The effectiveness of relevance feedback techniques in an environment of variable generality is examined in Section IX by B. Capps and M. Yin. It is shown that some of the feedback techniques are equally applicable to collections of small and large generality. Techniques of negative feedback (when no relevant items are identified by the users, but only nonrelevant ones) are considered in Section X by M. Kerchner. It is shown that a number of selective negative techniques, in which only certain specific concepts are actually modified during the feedback process, bring good improvements in retrieval effectiveness over the standard nonselective methods.

Finally, a new feedback methodology in which a number of documents jointly identified as relevant to earlier queries are used as a set for relevance feedback purposes is proposed and evaluated in Section XI by L. Paavola.

Two new clustering techniques are examined in Part 3 of this report, consisting of Sections XII and XIII. A controlled, inexpensive, single-pass clustering algorithm is described and evaluated in Section XII by D. B. Johnson and J. M. Lafuente. In this clustering method, each document is examined only once, and the procedure is shown to be equivalent in certain circumstances to other more demanding clustering procedures.

The query clustering process, in which query groups are used to define the information search strategy is studied in Section XIII by S. Worona. A variety of parameter values is evaluated in a retrieval environment to be used for cluster generation, centroid definition, and final search strategy.

The last part, number five, consisting of Sections XIV and XV, covers the design of on-line information retrieval systems. A new SMART system design for on-line use is proposed in Section XIV by D. and R. Williamson, based on the concepts of pseudo-batching and the interaction of a cycling program with a console monitor. The user interface and conversational facilities are also described.

A template analysis technique is used in Section XV by S. F. Weiss for the implementation of conversational retrieval systems used in a time-sharing environment. The effectiveness of the method is discussed, as well as its implementation in a retrieval situation.

Additional automatic content analysis and search procedures used with the SMART system are described in several previous reports in this series, including notably reports ISR-11 to ISR-16 published between 1966 and 1969. These reports are all available from the National Technical Information Service in Springfield, Virginia.

G. Salton

VIII. Variations on the Query Splitting Technique with Relevance Feedback

T. P. Baker

Abstract

Some experiments in relevance feedback are performed with variations on the technique of query splitting. The results obtained indicate that these variations, as tested, offer no significant improvement over previously tried methods of query splitting.

1. Introduction to Query Splitting

In a document retrieval system with relevance feedback, query splitting refers to the creation of multiple queries from a single previous query, making use of user relevance judgments on documents retrieved by that query in a previous search. The intention in generating these multiple queries is to allow the search to be directed toward several individual clusters of relevant documents, a necessary assumption being that these clusters exist and do contain relevant documents which have not been previously retrieved.

There is little doubt that in a situation where several clusters of relevant documents are retrieved in the initial search it is desirable to generate multiple queries for succeeding iterations. The problem remaining is to distinguish this condition from those in which the relevant documents are unclustered or fall into a single cluster.

Borodin, Kerr, and Lewis [1] propose one method. Their algorithm makes use of the average interdocument correlation among the relevant documents available for feedback as a cutoff in determining whether a given pair of documents should be split. The results obtained with this algorithm are inconclusive, but indicate that it is not sufficiently selective.

Ide [2] suggests that a more sophisticated algorithm might look for separation of relevant documents by nonrelevant documents within the document space, splitting a pair of documents if and only if there exists a nonrelevant document more highly correlated with each of them than they are with each other. In certain respects, this separation criterion is more faithful to the conceptual basis of query splitting than the average correlation criterion. Unlike the average correlation criterion, the separation criterion takes into account the distribution of the nonrelevant documents. This may be significant, since what is desired is the detection of clusters of relevant documents. In contrast, what the average correlation criterion does is to cluster relevant documents. Since nonrelevant documents are not taken into account, this will not produce legitimate clusters, in terms of the whole docunent space, when relevant documents locally outnumber nonrelevant documents, or vice versa. For this reason it would seem that Ide's untested separation criterion deserves more attention.

The usual concept of query splitting, as discussed by Borodin, Kerr, and Lewis and by Ide, is limited in application to cases where more than one relevant document is retrieved by a previous search iteration. It seems that if query splitting is of any value, something similar could be done for the queries which do not retrieve enough relevant documents to consider splitting in the usual sense. After all, these are generally the queries most in need of modification. What is needed is a dual to the usual formulation of query splitting — a technique of clustering nonrelevant documents for the generation of multiple queries through negative feedback.

2. Algorithms for Query Splitting

Since the algorithm of Borodin, Kerr, and Lewis [1] using the average correlation criterion has been shown to be largely ineffective on the SMART document collections available, and because the separation criterion of Ide [2] remains untried, the primary algorithm tested in this study makes us of the separation criterion.

Since a pair splitting criterion does not by itself define a set of clusters, but rather an association matrix, a splitting algorithm may additionally choose between the use of multilevel associations and the use of direct associations for generating clusters. An examination of the document and query collections used here immediately discloses that multilevel association virtually eliminates cases of splitting in positive feedback. Therefore in order to facilitate experimentation, the splitting algorithm is weakened by permitting only directly connected pairs within clusters used for positive feedback.

Adding to this constraint the requirement that all clusters be maximal, the two conditions are sufficient to define for any pair splitting criterion a unique set of clusters (not necessarily disjoint).

The actual application of these clustering conditions for experimentation with the ADI Abstracts-Thesaurus and Cranfield 200-Thesaurus collections is performed manually using document-document correlations computed by the SMART system. To allow combining the results of the split queries in a consistent fashion, the number of clusters generated for each query (in cases where more would be generated) is limited to two by joining the pair of documents which most nearly fails to pass the separation criterion. The resulting pairs of clusters are fed to the SMART normalized relevance feedback facility in successive iterations.

The SMART relevance feedback formula used is:

$$Q' = MQ + \frac{M}{n} \sum_{i=1}^{n} \frac{R_i}{|R_i|} - \frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{|N_i|} \quad .$$

where Q' is the new query; Q is the original query; M is an integer constant; n is the number of relevant documents (R_i) fed back; m is the number of nonrelevant documents (N_i) fed back.

The top ranking seven documents according to the first "half" of the split query are frozen in place, while the succeeding ranks are determined by another search iteration with the other "half" of the split query. This is done with the two "halves" reversed, as well, so as to average out the effects of order.

The procedure described is applied to all queries retrieving more than one relevant document in the top five ranks according to the first search.

For those queries not retrieving sufficient relevant documents to be split for positive feedback, splitting in negative feedback is attempted.

Where one relevant document is known, the dual to the separation criterion is tried, splitting pairs of nonrelevant documents that are more similar to the one relevant than they are to each other. The resulting clusters of nonrelevant documents are treated like the clusters of relevant documents above, with the single relevant document additionally being fed back with each "half" of the split query.

Where no relevant documents are known, nonrelevant documents are separated by correlation less than the average correlation between documents

Sample separation criterion in its weak form applied to query Q250 of the CRN2TH collection, which retrieved three relevant documents out of five on the first search. The relevant documents retrieved are 3, 115, and 197. The two nonrelevant are 7 and 160.

The interdocument correlation matrix is (in part):

```
3:115 0.5744  3:197 0.4700  3:160 0.4828   3:7  0.2208
            115:197 0.7926 115:160 0.5797 115:7 0.3179
                           197:160 0.5506 197:7 0.3136
```

The pair of relevant documents which must be split for feedback purposes because they are separated by a nonrelevant document is 3:197, which is split by 160.

```
The remaining associations are 3........115
                                         .
                                         .
                                         .
                                         .
                                         .
                                         .
                                         .
                                        197  ,
```

and the two derived clusters are 3-115 and 115-197.

Separation Criteria for Query Q250

Example 1

Sample separation criterion applied to query B04 of the ADIABTH collection, which retrieves two relevant documents out of the top five ranks according to the first search. The two relevant documents retrieved are 33 and 20. The nonrelevant documents are 5, 46, and 62.

The interdocument correlations are:

33:20	0.1097	33:5	0.4843	33:46	0.2000	33:62	0.2026
		20:5	0.2292	20:46	0.1073	20:62	0.0593

Although it might be interesting to split the nonrelevant documents, there are relevant ones here to split, and the nonrelevant ones are therefore used only to split relevant pairs. We see that the pair 33:20 is split by 5, since 0.4843 and 0.2292 are both greater than 0.1097. Thus 33 and 20 are separated for feedback purposes, and since they are the only relevant documents available they are the two clusters which will be used.

Separation Criterion for Query B04

Example 2

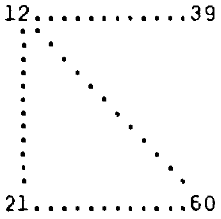
Application of the weak separation criterion to query A13 of the ADIABTH collection, which retrieves one relevant document in the five top-ranked documents according to the first search; the relevant document is 37 and the nonrelevant are 12, 21, 39, and 60.

The interdocument correlation matrix is:

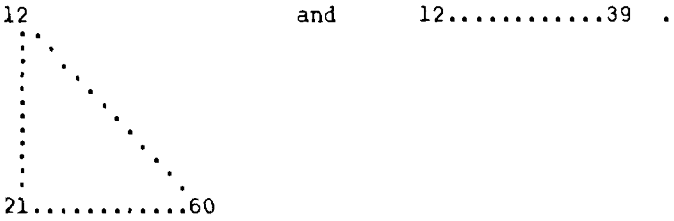
37:12 03411	37:21 0.3659	37:39 0.3225	37:60 0.4000
	12:21 0.3800	12:39 0.3769	12:60 0.3412
		21:39 0.1741	21:60 0.5066
			39:60 0.1061

The following pairs of documents are more highly correlated with 37 than they are with each other, and therefore are separated: 39:60; 21:39.

The remaining associations may be summarized:



Thus the resulting clusters of nonrelevant documents are:



Separation Criterion for Query A13

Example 3

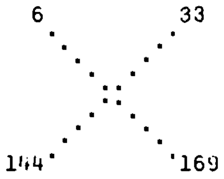
Application of the weak separation criterion to query Q189 of the CRN2TH collection, which retrieves one relevant out of five top-ranking documents according to the first search. The relevant document is 148 and the nonrelevant are 6, 33, 144, and 169.

The interdocument correlation matrix is:

148:6 0.1782	148:33 0.4881	148:144 0.6491	148:169 0.1816
	6:33 0.1630	6:144 0.1347-	6:169 0.2686
		33:144 0.5682	33:169 0.1218-
			144:169 0.0783

The following pairs of documents are more highly correlated with 148 than they are with each other, and therefore are separated for feedback purposes: 144:169; 33:169; 6:144; 6:33.

The remaining associations may be summarized:



The clusters of nonrelevant documents used for feedback are then:

6..........33 and 144..........169 .

Separation Criterion for Query Q189

Example 4

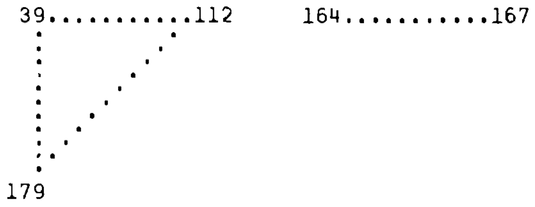
Application of the average correlation criterion to query Q182 of the CRN2TH collection, which retrieved no relevant documents in the top five ranks on the initial search:

Document-document correlations for the nonrelevant documents are:

```
39:112 .5367 39:164 .0100- 39:167  .0100- 39:179 .5696
            112:164 .1142 112:167  .1358 112:179 .6980
                          164:167  .7212 164:179 .2487
                                         167:179 .1663
```

The average correlation is 0.3190.

Thus the only associations permitted are:



which are the resulting clusters.

Correlation Criterion for QUERY Q182

Example 5

Application of the average correlation criterion to query Q266 o: the CRN2TH collection, which retrieved no relevant documents on the initial search.

The nonrelevant documents known are 58, 162, 163, 164, and 165.

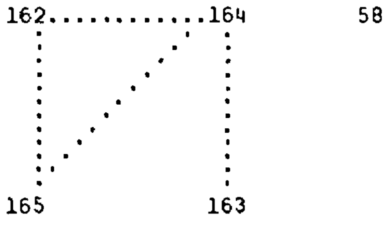
The interdocument correlations are:

58:162 .3932	162:163 .3240	163:164 .5194	164:165 .4585
58:163 .3368	162:164 .5679	163:165 .3744	
58:164 .3662	162:165 .5745		
58:165 .4113			

The average is 0.2819.

Thus the only permissable associations are 162:164, 164:165, 164:163, 162:165.

Thus the clusters are:



Correlation Criterion for Query Q266

Example 6

retrieved, and clusters are formed using multilevel associations (direct associations generally failing to produce any grouping at all). The clusters so formed are fed back in a manner similar to the clusters derived by the other two methods.

Results obtained with these three algorithms on the ADI Abstracts-Thesaurus and Cranfield 200-Thesaurus collections are summarized in the following section.

3. Results of Experimental Runs

The tables on the following pages summarize the results of runs made in the SMART system with splittable queries of the three categories mentioned in the preceding section for the ADI Abstracts-Thesaurus (82 documents and 35 queries — denoted by ADIABTH) and Cranfield 200 (200 documents and 42 queries — denoted by CRN2TH) collections.

The following conventions apply:

-	indicates that the results of the split query and control runs are indistinguishable in terms of the number of relevant documents retrieved.
*	indicates that all relevant documents are retrieved and no improvement is possible.
0	indicates that neither run retrieved any relevant documents.
•	indicates that this query would also have split according to the stronger version of the splitting requirement.
@	indicates a keypunching error detected too late to correct in one of the feedback document specifications for the trial run.

Queries from ADIABTH collection retrieving more than one relevant document on the first search, and splittable by the weak separation criterion:

Query	Improvement of split queries over ordinary normalized positive feedback in terms of relevant documents retrieved up to rank:			
	5	10	15	20
A03/a'	-	-	*	*
A03/b'	-	-	*	*
A15/a'	-1	-	-	-1
A15/b'	-1	-	1	-
B04/a'	-	1	-	-
B04/b'	-1	1	-	-
Average:	-0.5	0.33	0.17	-0.17

Query Splitting Results for ADIABTH Collection
(POSNEG)

Table 1

Queries from the CRN2TH collection retrieving more than one relevant document on the first search, and splittable by the weak separation criterion:

Improvement of split queries over ordinary normalized positive feedback in terms of relevant documents retrieved up to rank:

Query	5	10	15	20
Q122/a'	-	1	-	-
Q122/b'	-1	1	-	-
Q148/a'	-2	-	-	-
Q148/b'	-2	-	-	-
@ Q250/a	-	-	-	-
@ Q250/b	-	-	-	-
Q268/a	-	-	*	*
Q268/b	-	-	*	*
Q269/a'	-	-	*	*
Q269/b'	-	-	*	*
Average:	-4/10	2/10	-	-

Query Splitting Results for CRN2TH Collection
(SPLPOS)

Table 2

Queries from the CRN2TH collection retrieving more than one relevant document on the first search, and splittable by the weak separation criterion:

Improvement of split queries over ordinary normalized positive and negative feedback in terms of relevant documents retrieved up to rank:

Query	5	10	15	20
Q122/a'	-	1	-1	-
Q122/b'	-1	1	-1	-
Q148/a'	-1	-	-	-
Q148/b'	-2	-	-	-
@ Q250/a'	-1	-1	-1	-1
@ Q250/b'	-1	-1	-1	-1
Q268/a	-	-	*	*
Q268/b	-	-	*	*
Q269/a'	-1	-	*	*
Q269/b'	-1	-	*	*
Average:	-8/10	0	-4/10	-1/10

Note: This comparison is unfair to the split query run, since it made no use of negative feedback infoımation.

Query Splitting Results for CRN2TH Collection
(SPLPOS)

Table 3

Queries from the CRN2TH collection retrieving more than one relevant document out of five retrieved on the first search, and splittable by the weak separation criterion:

Query	Improvement of split queries over ordinary normalized positive and negative feedback in terms of relevant documents retrieved up to rank:			
	5	10	15	20
Q122/a'	-	-	-2	-1
Q122/b'	-	1	-1	-1
Q148/a'	-	-	-	-
Q148/b'	-	-	-	-
@ Q250/a'	-	-1	-1	-1
@ Q250/b'	-	-1	-1	-1
Q268/a	-	-	*	*
Q268/b'	-	-	*	*
Q269/a'	-	-	*	*
Q269/b	-	-	*	*
Average	-	-1/10	-5/10	-4/10

Note: Unlike the other tests, this run was done with the first five documents retrieved by the initial search frozen in their rank positions. It is also unfair to the split query run, since the control made use of negative feedback.

Query Splitting Results for CRN2TH Collection (SPLPOS)

Table 4

Queries from ADIABTH collection retrieving no relevant documents on the first search, and splittable by correlation less than average:

Improvement cf split queries over ordinary normalized negative feedback in relevant documents retrieved up to rank:

Query	5	10	15	20
A08/a	0	-	*	*
A08/b	-1	-	*	*
A09/a	0	-1	-	-
A09/b	1	-	-	-
B11/a	0	0	-	-
B11/b	0	0	-1	-
B13/a	0	-	1	-
B13/b	0	-	-	-
B15/a	0	0	-1	-3
B15/b	0	0	-1	-2
Average:	0	-1/10	-1/5	-1/2

Query Splitting Results for ADIABTH Collection
(ALLNEG)

Table 5

Queries from the CRN2TH collection retrieving no relevant documents in the top five ranks for the first search, and splittable by correlation below average.

Improvement of split queries over ordinary normalized negative feedback with only the top-ranked nonrelevant document used (as opposed to the previous run which used all five nonrelevant available) in terms of relevant documents retrieved up to rank:

Query	5	10	15	20
Q079/a	-1	-1	-1	-
Q079/b	-1	-1	-1	-
Q126/a	0	1	*	*
Q126/a	0	1	*	*
Q132/a	-	-	1	-
Q132/b	-1	-	-	-1
Q182/a	-	-	-	-
Q182/b	-1	-	-	--
Q266/a	0	1	3	2 2
Q266/b	0	2	4	3 3
Q323/a	-	-	-	- -
Q323/b	-	-	-	-
Average:	-4/12	3/12	6/12	4/12

Query Splitting Results for CRN2TH Collection
(NORELS)

Table 6

Queries from the CRN2TH collection retrieving no relevant documents in the first five ranks for the first search, and splittable by correlation below average.

Improvement of split queries over ordinary normalized negative feedback in terms of relevant documents retrieved up to rank:

Query	5	10	15	20
Q079/a	0	0	0	-
Q079/b	0	0	0	-
Q126/a	0	-	*	*
Q126/b	0	-	*	*
Q132/a	-	-1	-	-
Q132/b	-1	-1	-1	-1
Q182/a	-	-	-	-
Q182/b	-1	-	-	-
Q266/a	0	1	-	-
Q266/b	0	2	1	1
Q323/a	-	-	-	-
Q323/b	-	-	-	-
Average:	-2/12	1/12	0	0

Query Splitting Results for CRN2TH Collection
(NORELS)

Table 7

Queries from the ADIABTH collection retrieving one relevant document in the top-ranking five on the first search, and splittable by weak separation criterion for nonrelevant documents.

Improvement of split queries over ordinary normalized positive and negative feedback in terms of relevant documents retrieved up to rank:

Query	5	10	15	20
@ A01·	-1	-	-	-
	-	-	-	-
@ A02·	-	1	*	*
	-	1	*	*
A04·	-	-	-	-
	-	-	-	-
A06·	-	-	-	-
	-	-	-	-
A07	-	-	-	-2
	-	-	-	-
A10	-	-	-1	-
	-	-	-	1
A11	-	-	-	-
	-	-	-1	-
A12·	-	-	-	-1
	-	-	-	-2
A13	-	-	-	-
	-	-	-	-
A14·	-	-	-	-
	-	-	-	-
A17	*	*	*	*
	-1	*	*	*
B16·	-1	-	-	-
	-	-	-1	1
Average:	-3/24	2/24	-3/24	-2/24

Query Splitting Pesults for ADIABTH Collection
(SPLNEG)

Table 8

Queries from the CRN2TH collection retrieving only one relevant document in the top five ranks on the first sea. ·h, and splittable by the weak separation criterion for nonrelevant documents.

Improvement of split queries over ordinary normalized positive and negative feedback in terms of relevant documents retrieved up to rank:

Query	5	10	15	20
Q123/a	-	-	-	-
Q123/b	-	-	-	-
Q130/a	-	-1	*	*
Q130/b	-	-1	-1	-1
Q141/a	*	*	*	*
W141/b	*	*	*	*
Q170/a	-	-	-	-
Q170/b	-	-	-	-
Q189/a'	*	*	*	*
Q189/b'	*	*	*	*
Q272/a	-1	-	*	*
Q272/b	-	-	-1	*
Average:	-1/12	-2/12	-2/12	-1/12

Query Splitting Results for CRN2TH Collection
(ONEREL)

Table 9

Queries from the CRN2TH collection retrieving only one relevant document in the top five ranks on the first search, and splittable by the weak separation criterion for nonrelevant documents.

Improvement of split queries over ordinary normalized positive feedback in terms of relevant documents retrieved up to rank:

Query	5	10	15	20
Q123/a	-	-	-	-
Q123/b	-	-	-	-
Q130/a	-	-	1	-
Q130/b	-	-	-	-1
Q141/a	*	*	*	*
Q141/b	*	*	*	*
Q170/a	-	-	-	-
Q170/b	-	-	-	-
Q189/a'	*	*	*	*
Q189/b'	*	*	*	*
Q272/a	-	-1	*	*
Q272/b	1	-1	-1	*
Average:	1/12	-2/12	0	-1/12

Query Splitting Results for CRN2TH Collection
(ONEREL)

Table 10

Correlations			
0	1	2	3
0.4523	0.6176	0.8241	0.2811
0.3780	0.5598	0.4165	0.1603
0.3665	0.3343	0.4137	0.3070
0.3647	0.3325	0.3680	0.5045
0.3638	0.2731	0.3518	0.4064
0.3467	0.2363	0.3412	0.8449
0.3333	0.2347	0.3246	0.1727
0.3283	0.2334	0.2994	0.4017
0.3119	0.2206	0.2994	0.3635
0.3086	0.2141	0.2946	0.3530
0.3000	0.2130	0.2930	0.3529
0.3000	0.2092	0.2908	0.3390
0.2949	0.2033	0.2768	0.3360
0.2949	0.2001	0.2758	0.3356
0.2917	0.1763	0.2668	0.3350
	0.1606		
	0.1547		
0.2673			
	0.1418	0.2488	
		0.2482	
		0.2415	
0.2080			
0.1803			0.2109
0.1793			
			0.1705
			0.1607

Rank	Documents			
	0	1	2	3
1	33R	20R	20R	20R
2	5	33R	57R	57R
3	62	42	46	46
4	46	7	5	5
5	20R	9	56	56
6	70	67	33R	33R
7	10	56	7	7
8	11	27	37	10
9	1	5	60	21
10	80	71	21	76
11	37	28	9	37
12	60	18	45	3
13	47	23	12	62
14	56	68	10	60
15	3	24	79	53
17		36R		
19		16R		
20	57R			
22		57R	36R	
23			26R	
26			16R	
29	26R			
32	16R			36R
34	36R			
39				16R
43				26R

	Doc. Corr	Cent. Corr	Drop Doc	Corr. Rank	Old. Doc	Old Reldoc	New Doc
Run 0	82	0	17	65	0	0	0
Run 1	82	0	31	51	0	0	5
Run 2	82	0	12	70	5	2	0
Run 3	75	0	14	61	5	2	7

0 initial search
1 control run with positive and negative feedback
2 first "half" of split query
3 second "half" of split query

Sample Output for Query F04

Fig. 1

4. Evaluation

In general, the results for query splitting in positive feedback with the separation criterion are comparable to those achieved by Borodin, Kerr and Lewis in their experiments with the average correlation criterion. Although a slight improvement may be noted for split queries over ordinary feedback, it is not predictable enough to justify the use of query splitting in a working retrieval system.

Only if a more selective method can be devised for determining which queries will benefit from splitting will the technique become of practical value. Merely strengthening the splitting requirement by permitting multi-level associations in cluster formation appears to be of some value in eliminating nonproductive splitting in the queries tested. All queries split by the weaker method which show an improvement under splitting would be split in like manner by the stronger method. Strengthening the separation as well, by providing that pairs be separated only if they exceed the requirements by some margin, may also be of value in restricting the number of undesired splits.

For negative feedback, the situation is worse. The only run in which splitting exhibited any improvement over the usual negative feedback was on queries in the Cranfield collection retrieving no relevant documents in the first search. Even there, the improvement was erratic. This failure of splitting applied to negative feedback is not entirely surprising, since the hypothesis of separate clusters of relevant documents used to justify splitting in positive feedback does not apply. Here the best justification for splitting is that, since the locations of no relevant documents are known, multiple queries may offer more chance of success by means of a "shotgun" effect — scattering the search over a larger area of the document space.

Altogether, the results of the negative feedback runs indicate that the different "halves" of the split queries do not usually retrieve significantly different portions of the document space. Thus it would seem that this "shotgun" effect is not taking place. It may be that results can be improved by weighting the nonrelevant documents more heavily in feedback. In any case, negative query splitting as tested does not appear to benefit from this effect sufficiently to justify the effort of multiple query generation.*

Although the results of these experiments are largely negative, it is important in viewing them to consider that the queries tested were written by experts in their fields and are therefore generally consistent, thus making the probability of success in query splitting rather low. Also, being small, the document collections used are inimical to the existence of multiple clusters of relevant documents. Relevant documents in such small collections tend to fall into single clusters, or none. Although the success of query splitting in these adverse circumstances would be a strong argument in its favor, its failure in the same circumstances is less conclusive. It would appear that if truly significant results are to be achieved with query spli ting, they will be achieved in the environment of a larger more diverse document collection and with more realistically inconsistent queries.

* The only exception to this is query Q266 of CRN2TH, which showed remarkable improvement on splitting.

References

[1] A. Borodin, L. Kerr and F. Lewis, Query Splitting in Relevance Feedback Systems; Report ISR-14 to the National Science Foundation, Section XII, October 1968.

[2] E. Ide, Relevance Feedback In An Automatic Document Retrieval System; Report ISR-15 to the National Science Foundation, January 1968.

[3] D. Williamson, R. Williamson, M. Lesk, The Cornell Implementation of the SMART System; Report ISR-16 to the National Science Foundation, Section I, September 1969.

IX. Effectiveness of Feedback Strategies on Collections of Differing Generality

B. Capps and M. Yin

Abstract

This study evaluates the comparative effectiveness of several feedback strategies on collections which differ in generality, namely the Cranfield 200 and Cranfield 400 collections. A new query set which produces a constant number of relevant documents over the two collections is used to regulate the generality. The results are assessed from both the user and the system viewpoint; some strategies do appear equally effective on both collections.

1. Introduction

The ultimate goal of automatic information retrieval systems is to obtain a performance in "real life" situations equally as good as or better than in manual systems under operational conditions. Experiments done on automatic systems such as SMART are performed on controlled and limited collections. Therefore, in order to predict how the system will perform in a library situation, experiments on collections of different sizes are done and the results compared to see if there is a significant loss in performance as larger collections are used.

Generality is the proportion of relevant documents in a collection to total number of documents. In collections of varying sizes, generality is expected to differ, because the number of relevant documents does not increase proportionally to the number of nonrelevant documents. Therefore, results from test collections of different generality can be viewed as an indication of how the results from a test environment would be reflected in a real life situation.

This study is concerned with the relevance feedback aspect of information retrieval. Relevance feedback is one of the ways to utilize user opinion in improving search effectiveness [1]. A set of documents is given to the user who judges which documents are relevant to his request. This information is then used to modify his original query for another search through the collection. The rationale is that the original query might be badly worded, so that the incorporation of concepts from documents judged relevant might retrieve other related documents.

The method used in this study is to run several search strategies on collections of different generality and then to compare the retrieval performances. Several means are available to measure retrieval performance depending on the viewpoint taken. The recall-precision graph is used to represent the user viewpoint of how well the system is satisfying his needs. However, this is not adequate to measure system efficiency; consequently, fallout and adjusted precision have been developed. Fallout is the proportion of nonrelevant documents retrieved over total number of nonrelevant documents in the collection. When plotted against recall, this takes into account how much work the system has to do to retrieve equivalent numbers of relevant documents. When fallout is constant, precision can be adjusted to take generality into account so that the precision from collections of different generality can be compared on an equal basis [3].

2. Experimental Environment

The test collections should be similar in all respects except for generality. Ide [2] cites four factors which might account for the differences in results of the two collections she used — Cran 200 and ADI;

a) difference in subject matter

b) difference in collection scope

c) difference in variability within collection

d) difference in query construction and relevance judgment.

The CRN2NUL and CRN4NUL collections seem to eliminate these factors since they are subcollections of a homogeneous set — Cranfield 1400 — and are not mutually exclusive subcollections.

To vary the generality, the number of relevant items is held constant while the number of nonrelevant items varies. This can be done by creating a new query collection from the original CRN2NUL QUESTS and CRN4NUL QUESTS collections. The selected queries have the same relevance decisions in both the Cran 200 and Cran 400 collections. There are twenty-two such queries with a total of one-hundred and fifteen relevant documents. The formula for generality, averaged over all queries, is:

$$\text{Generality} = \frac{1000 \times \dfrac{\text{total relevant in collection}}{\text{number of queries}}}{\text{total number of documents in collection}} \qquad [4]$$

The generality for Cran 200 with respect to this new query set is 26.14 and for the Cran 400 is 12.30.

The query-update formula used for relevance feedback is:

$$Q_{i+1} = \pi Q_i + \omega Q_o + \alpha \sum_{1}^{\min(n_a, n'_r)} r_i + \mu \sum_{1}^{\min(n_b, n'_s)} s_i \qquad [2]$$

where

π, ω, α, μ	are multipliers
Q_{i+1}	is updated query
Q_i	is previous query
Q_o	is original query
n'_r	is number of relevant documents retrieved
r_i	is relevant document retrieved
n'_s	is number of nonrelevant documents retrieved
s_i	is nonrelevant document retrieved
n_a, n_b	specify the number of documents to be used.

Various strategies can be formulated using the above equation with the added parameters in the SEARCH routine of the SMART system, such as ALLOF, ATLEST and NOMOR. ALLOF is the number of documents to be retrieved. ATLEST is the minimum number of documents to be used in feedback and NOMOR is the maximum number of documents to be searched to provide documents for feedback. Only one iteration of feedback is used in this study because the most noticeable effect of feedback results from this iteration [5]. A frozen feedback iteration is used to eliminate the ranking effect for evaluation purposes.

Since the purpose of the experiment is to study the overall effect of feedback on these collections, a wide range of strategies are chosen:

Strategy 1 is positive feedback

Strategy is the "dec hi" strategy [2]

Strategy 4 is a modified "dec hi" strategy which uses a nonrelevant document for feedback only when no relevant documents are retrieved

Strategies 3 and 5 use varied multipliers.

The actual parameters are shown in Table 1.

This study attempts to determine whether feedback improves retrieval in one collection more than the other. That is, the initial full search results serve only as a base line and the improvement after using feedback is the result to be measured. Consequently, the following performance measures are stressed:

a) Precision improvement $-$ $P_1 - P_0$
This indicates whether a particular strategy is better for one collection than the other from a user viewpoint.

		Strategies				
		1	2	3	4	5
Original Multiplier	ω	1	1	1	1	2
Positive Rank Cut	n_a (ALLOF)	10	5	5	5	5
Positive Multiplier	α	1	1	4	1	1
Negative Rank Cut	n_b ($ALLOF)		1	5	1	5
Negative Multiplier	μ		-1	-1	-1	-1
Negative At Least	$ATLST		1	0	1	0
Negative No More	$NOMOR		10	5	5	5

Parameters for the Feedback Strategies

Table 1

P_0 is precision of initial search
P_1 is precision of feedback iteration
These are taken at fixed recall points.

b) Percentage of precision improvement — $\frac{P_1 - P_0}{P_0} \times 100\%$

This takes into account the fact that precision is better for a collection with a higher generality number [4] by taking the difference with respect to the original precision.

c) Fallout improvement — $F_0 - F_1$

A performance improvement implies that fallout for the feedback iteration is less than fallout for the initial search. This equation is equivalent to $(-1) \times (F_1 - F_0)$ and multiplying by -1 serves to transform the difference onto the positive scale.

F_0 is fallout of initial search
F_1 is fallout of feedback iteration
These are taken at fixed recall points.

d) Percentage of fallout improvement — $\frac{F_0 - F_1}{F_0} \times 100\%$

This takes into account the fact that the fallout is not the same for the initial searches on both collections. Therefore, the difference is computed as a percentage of the original.

e) Adjusted precision — $P_A = \frac{R_1 \times G_2}{(R_1 \times G_2) + F_1(1000 - G_2)}$ [4]

Precision of the Cran 200 is adjusted to that of Cran 400 and not vice versa, because the emphasis of this study is on performance of larger collections.

R_i is fixed recall points
F_1 is fallout of Cran 200 at R_1 recall
G_2 is generality of Cran 400

ln this manner, the results from two collections of different generality can be compared on an equal basis. This comparison is from a system viewpoint.

f) Adjusted precision improvement — $P_{A_1} - P_{A_0}$

Similar to a).

g) Percentage of adjusted precision improvement —

$$\frac{P_{A_1} - P_{A_0}}{P_{A_0}} \times 100\%$$

Similar to b).

3. Experimental Results

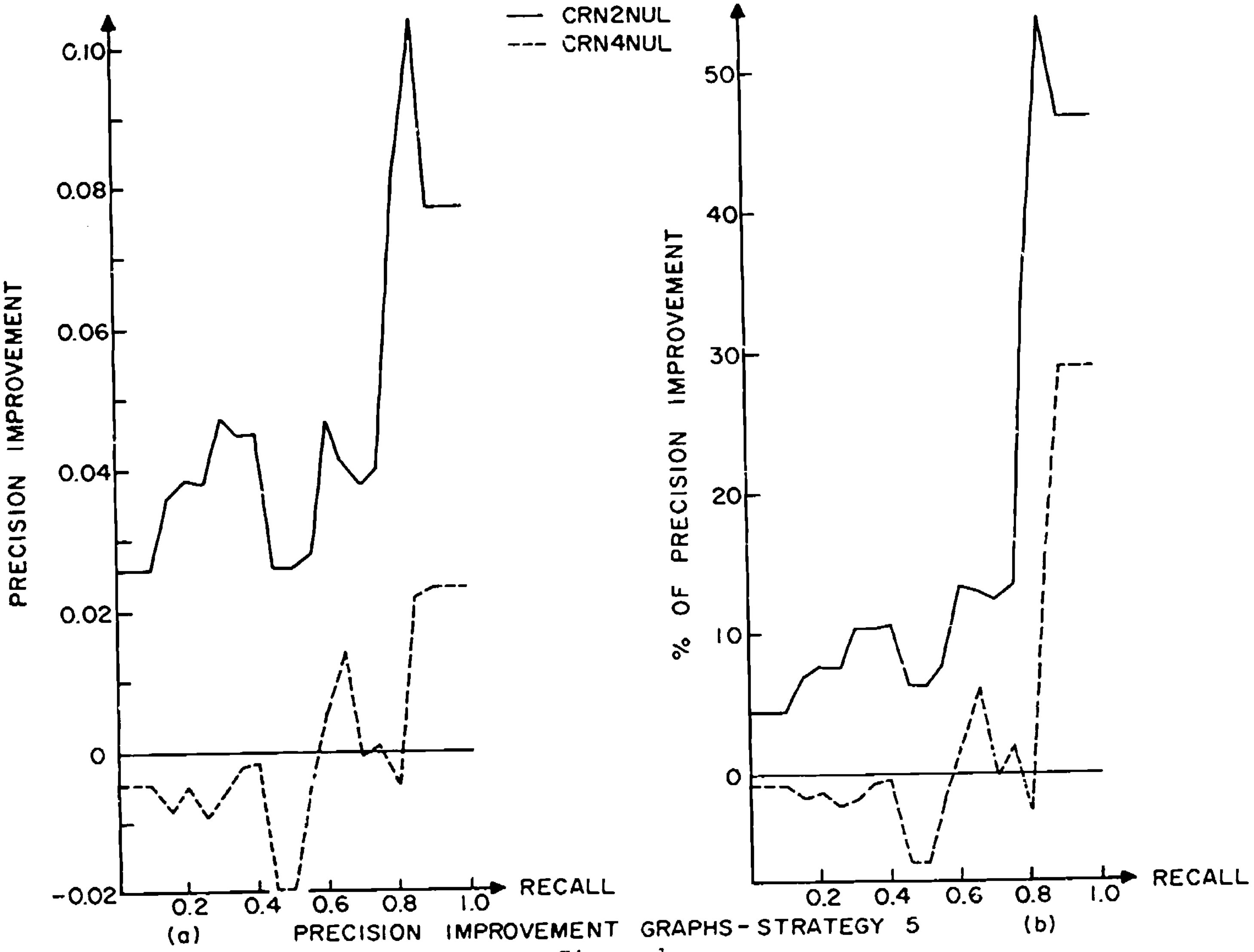
The results seem to fall into two categories with strategies 2, 3 and 5 in one group (group A) and strategies 1 and 4 in the other group (group B). The former group consistently shows a good performance for Cran 200, but there is little improvement for Cran 400, whereas the latter group shows an equivalent improvement. The average improvements for one strategy from each group are shown in Table 2.

In group A from both a system and a user viewpoint, the Cran 200 performs better as can be seen in all the improvement graphs. In fact, for strategy 3, Cran 400 performs worse using feedback than for the full search as shown by the negative values in the precision improvement curve (Fig. 1a) and the percentage fallout improvement curve (Fig. 2b). This result seems to indicate that this class of feedback strategies will not perform well in a library situation. For strategies 3 and 5, the result is probably due to the large number of nonrelevant documents

	Strategy 1		Strategy 3	
	Cran 200	Cran 400	Cran 200	Cran 400
Precision improvement	.0293	.0307	.0526	.0218
% of Precision improvement	12.50	16.99	19.38	12.73
Fallout improvement	.0104	.0770	.0130	.0066
% of Fallout improvement	13.38	15.94	21.73	12.16
Adj Precision improvement	.0194		.0379	
% Adj Precision improvement	24.21		23.79	

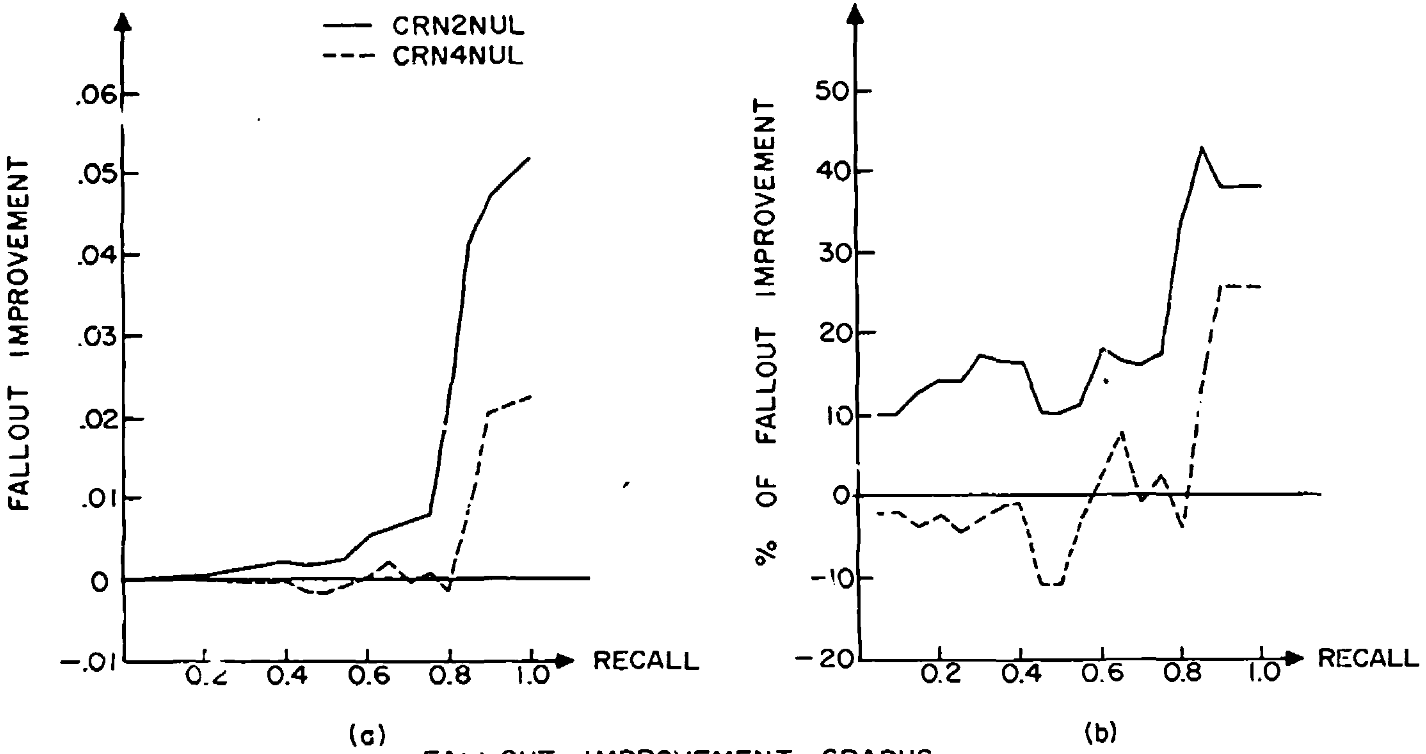
Average Improvement Results for Strategies 1 & 3

Table 2



PRECISION IMPROVEMENT GRAPHS - STRATEGY 5

Figure 1



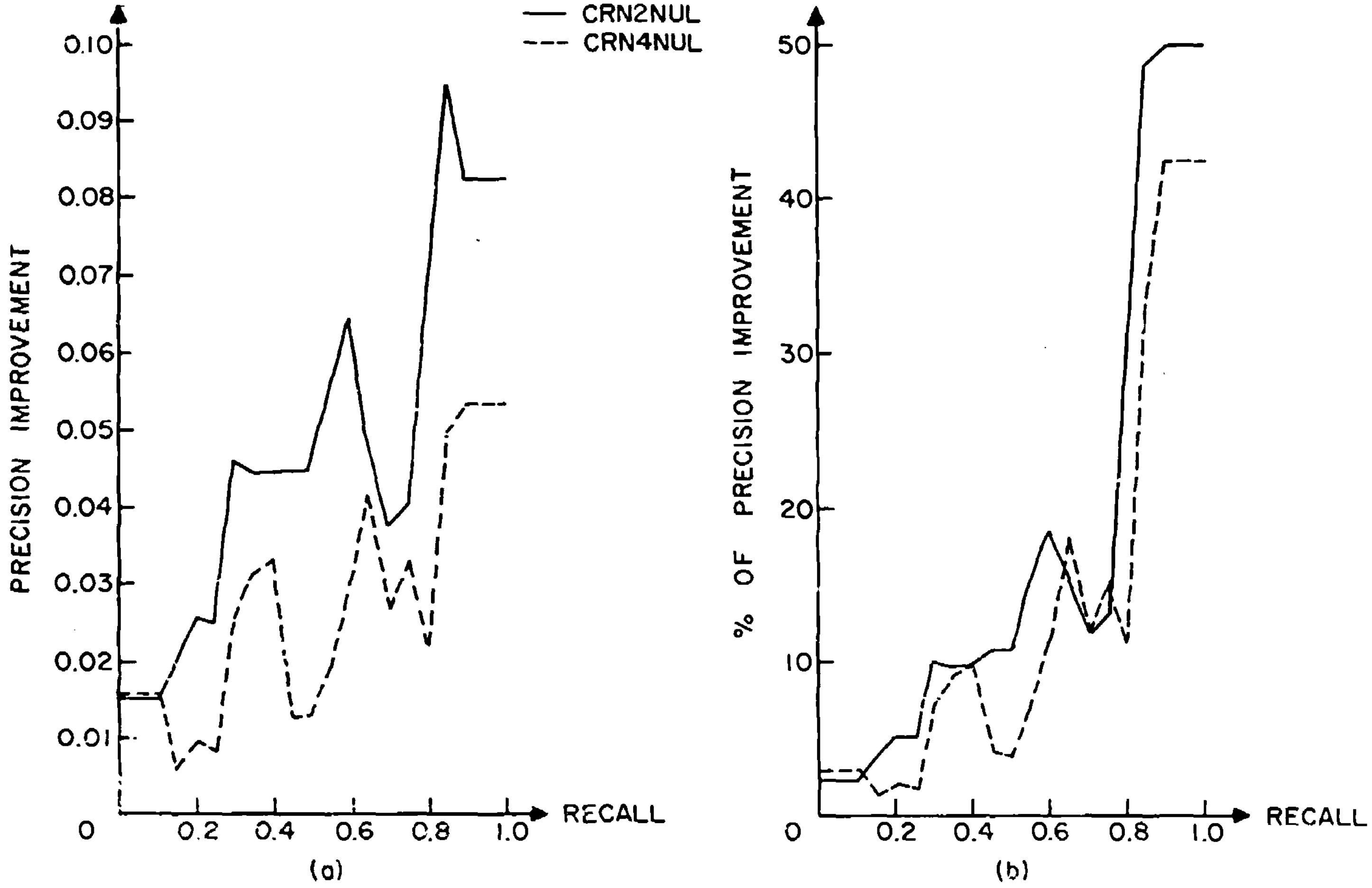
FALLOUT IMPROVEMENT GRAPHS.
STRATEGY 5

Figure 2

used for feedback which tend to eliminate the query. There is a median of four nonrelevant documents used for feedback on the Cran 400 and of three documents on the Cran 200. In strategy 5 on the Cran 400, out of the twenty-two queries, seven queries have two or fewer concepts left after feedback whereas on the Cran 200 there are only four such queries. The larger multiplier for the original query in strategy 3 partially offsets this effect of erasing the query.

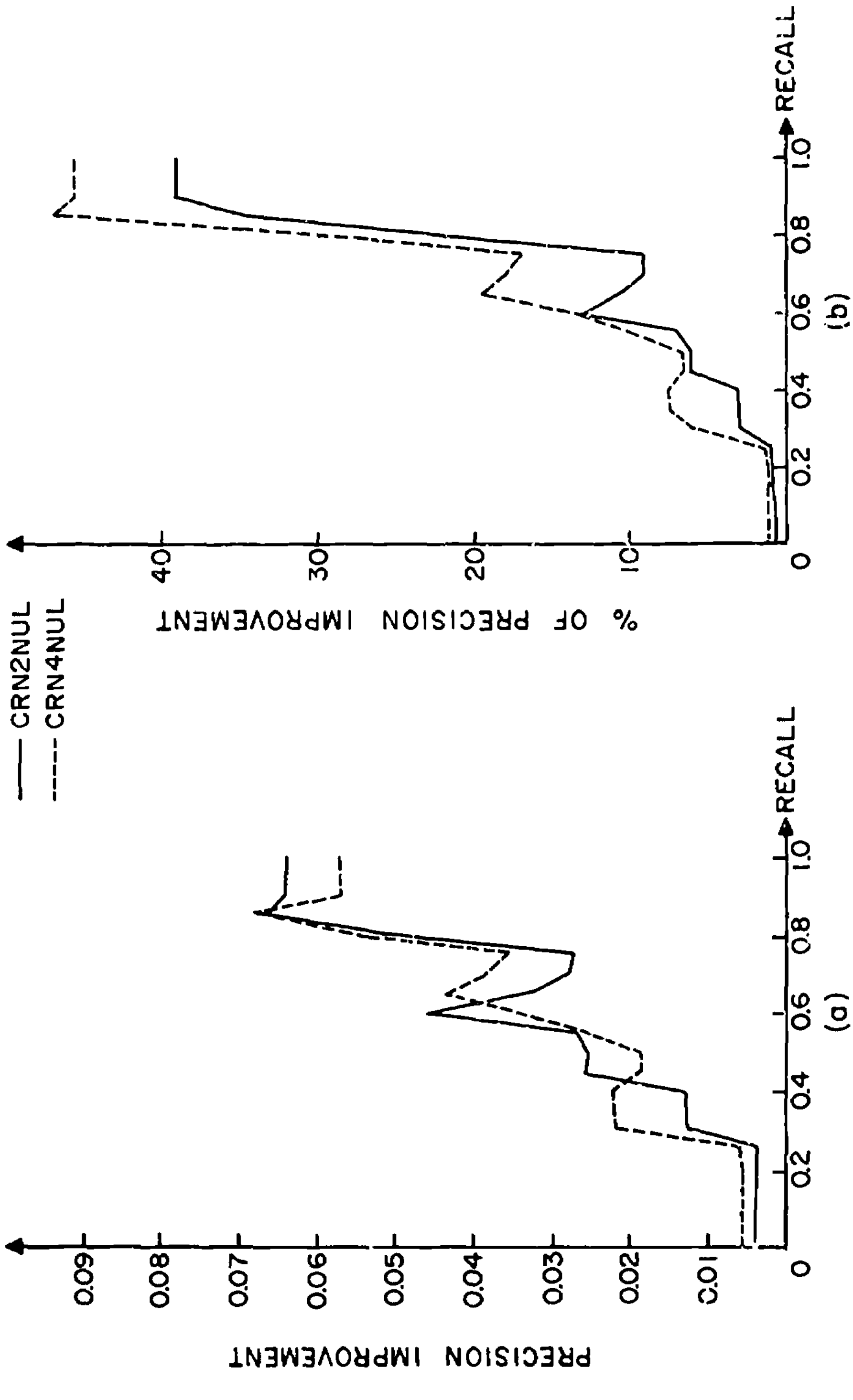
As for strategy 2 which always uses one nonrelevant document for feedback, the Cran 200 precision improves while the Cran 400 precision does not (Fig. 3a). This is due to the fact that on the Cran 200 there would be more relevant documents retrieved (median of 2); therefore, one nonrelevant document does not erase the query. On the Cran 400, however, fewer relevant documents would be retrieved (median of 1); herefore, one nonrelevant document might remove more concepts than are added by the relevant documents in the feedback.

Looking at the precision improvement graphs for group B, Cran 200 and Cran 400 curves using strategy 1 (Fig. 4a) are interspersed whereas for strategy 4, the Cran 200 curve is usually higher (Fig. 5a). But looking at the percentage precision graphs, for strategy 1 (Fig. 4b), the Cran 400 is better at all recall points. This is rot unexpected, since the original precision of the Cran 400 is lower than that of the Cran 200. Therefore even with a similar increase in precision, from a system viewpoint, the feedback is more helpful in im[illegible]ing retrieval for the Cran 400 since this larger collection



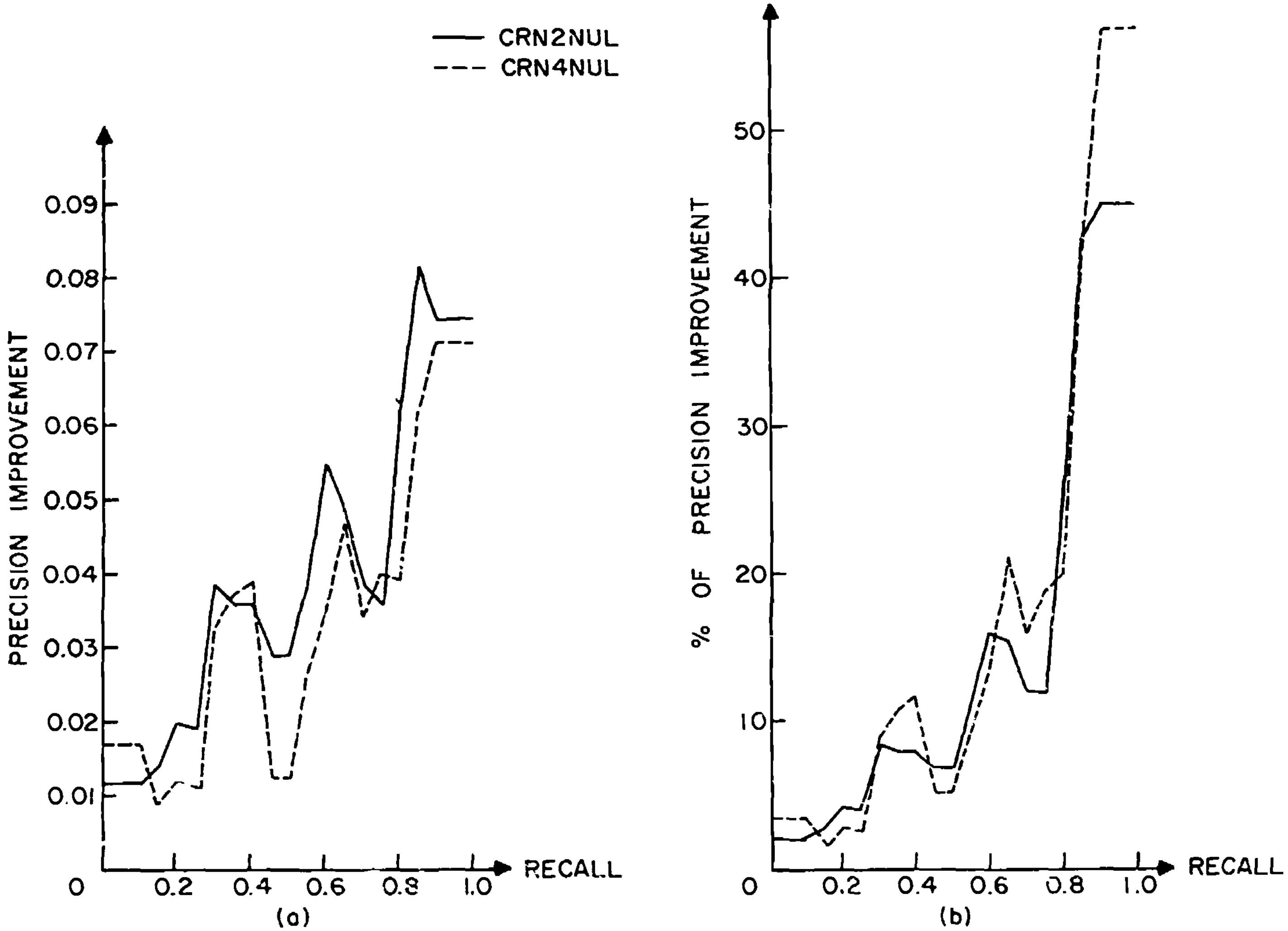
PRECISION IMPROVEMENT GRAPHS - STRATEGY 2

Figure 3



PRECISION IMPROVEMENT GRAPHS - STRATEGY I

Figure 4



PRECISION IMPROVEMENT GRAPHS – STRATEGY 4

Figure 5

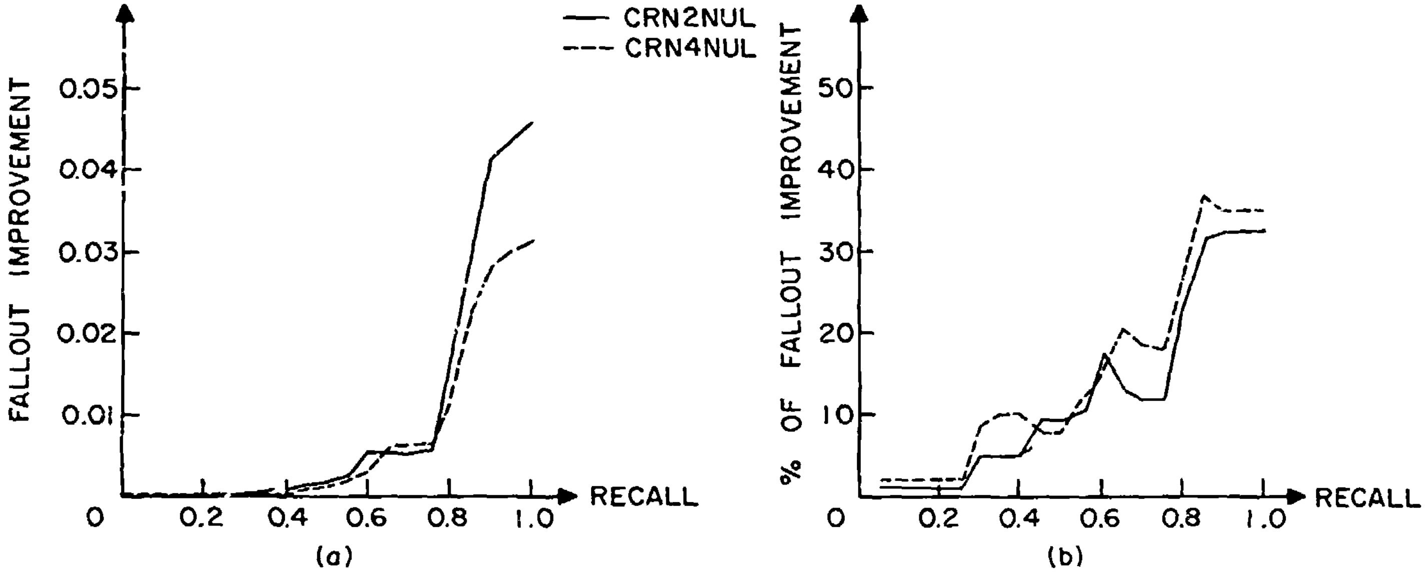
is not as favorable to retrieval as a smaller collection in the first place (lower original precision). For strategy 4 (Fig. 5b), the two curves are interspersed instead of the Cran 400 being lower because once the original precision is taken into account, the percentage increase becomes similar.

Theoretically, the fallout curves (see Appendix) for the two collections should be the same. However, there is probably a subset in the Cran 400 collection of nonrelevant documents which have a very low probability of being retrieved [4]. This explains why fallout for the Cran 400 seems better, a fact to be remembered when comparing fallout values.

For strategy 1, in the fallout improvement graph (Fig. 6a), Cran 200 is for the most part better. On the corresponding percentage fallout improvement graph (Fig. 6b) the Cran 400 is slightly better. For stragety 4, on the other hand, the difference in fallout improvement is more pronounced and the percentage fallout improvement is more similar (Fig. 7a, 7b).

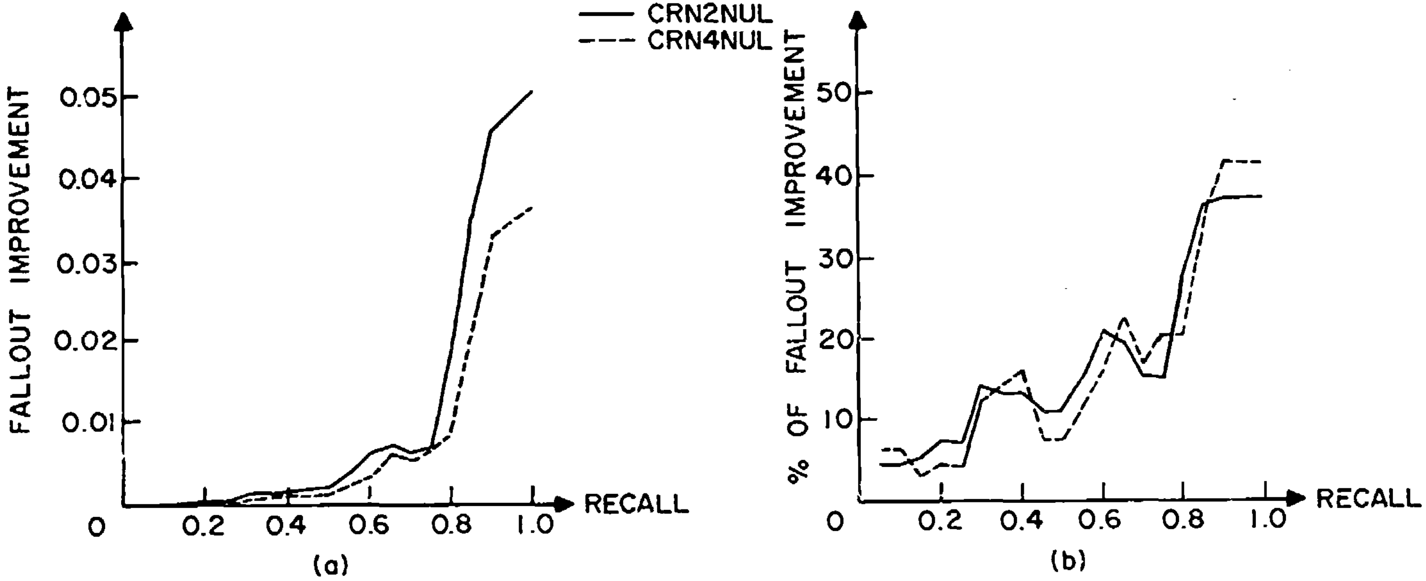
These fallout results are quite logical. Since on the feedback run, the number of relevant documents retrieved on the Cran 200 tends to be larger than for the Cran 400 (usually one more relevant document for Cran 200), the number of nonrelevant documents would be smaller. Therefore, the fallout improvement for the Cran 200 is larger. However, when the original fallout values are considered, the two collections become similar.

Once precision for the Cran 200 is adjusted to that of the Cran 400, the recall-precision curve for the Cran 400 is



FALLOUT IMPROVEMENT GRAPHS
STRATEGY I

Figure 6



FALLOUT IMPROVEMENT GRAPHS
STRATEGY 4

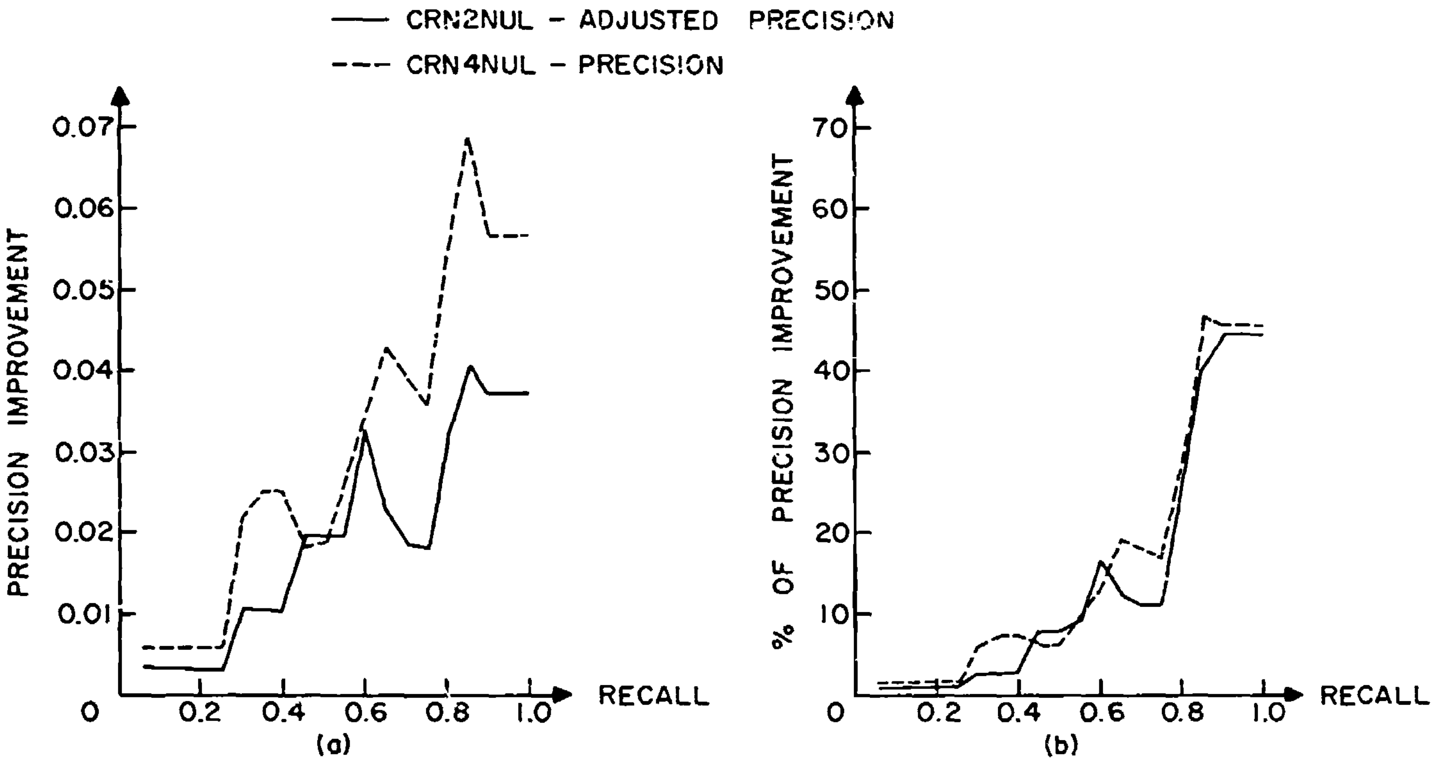
Figure 7

is lower than that for the Cran 400 (see Appendix). Therefore, according to these graphs, from a system viewpoint, Cran 400 definitely shows better performance. From the adjusted precision improvement graphs (Fig. 8a, 9a), the improvement of Cran 400 is at least equal if not more than that of Cran 200. This result is also supported by the percentage adjusted precision improvement graphs (Fig. 8b, 9b). From both a user and a system viewpoint, it would appear that use of these feedback strategies is at least as effective for a larger collection (lower generality number).

An interesting comparison can be made between strategies 2 and 4 since they are similar in that both use negative feedback of one nonrelevant document. However, the fact that strategy 4 uses negative feedback only when no positive feedback can be performed, as opposed to strategy 2 which uses it for all queries, causes strategy 4 to be effective and strategy 2 to fail on the Cran 400. For strategy 4, the few relevant documents used in feedback are not offset by any negative feedback as they would be for strategy 3 (see discussion of strategy 2 above).

4. Conclusion

Results of this study are encouraging in that they seem to indicate that some feedback strategies can indeed be used in a realistic environment. Those commonly used strategies such as pure positive feedback and the strategy which uses the top ranking nonrelevant document only when no relevant documents are

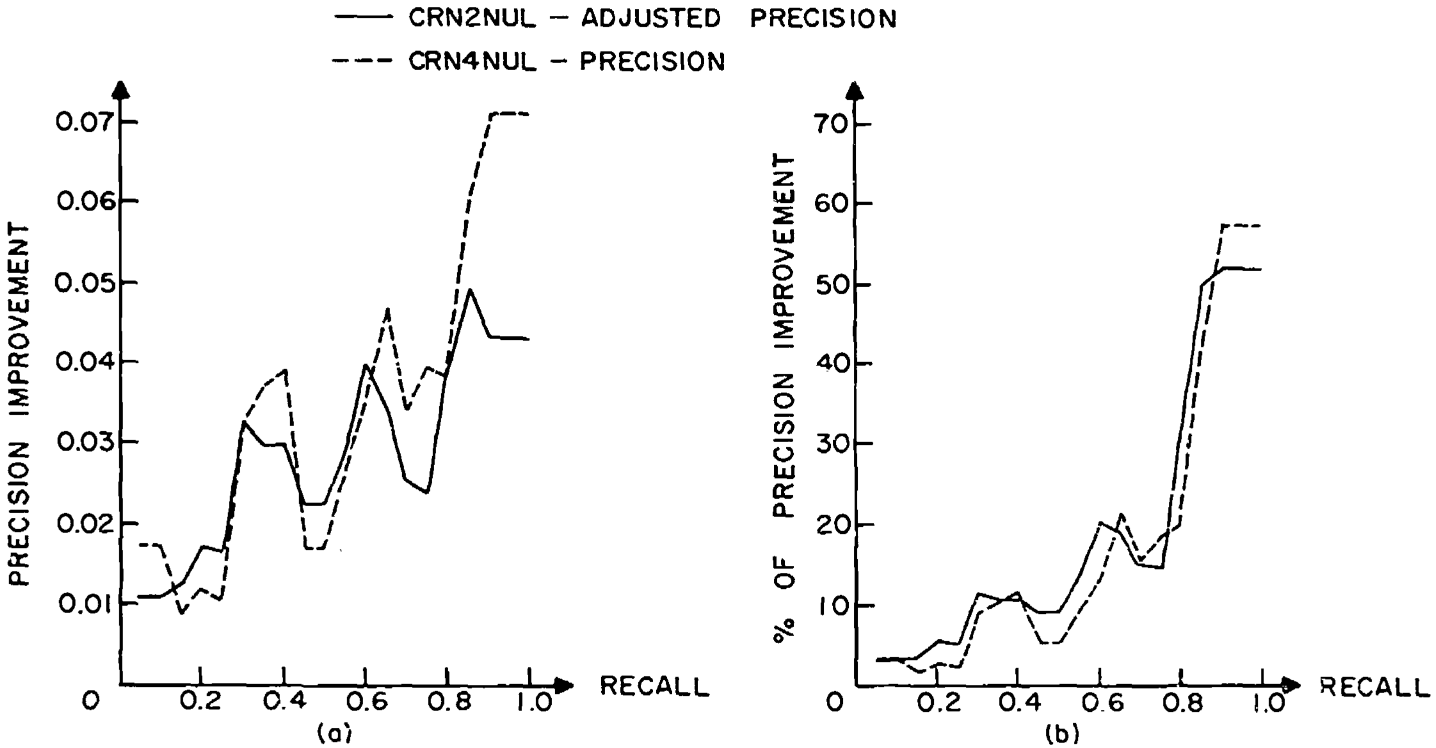


PRECISION IMPROVEMENT GRAPHS

WITH PRECISION OF CRAN 200 ADJUSTED

STRATEGY I

Figure 8



PRECISION IMPROVEMENT GRAPHS

WITH PRECISION OF CRAN 200 ADJUSTED

STRATEGY 4

Figure 9

retrieved, are equally effective on the Cran 200 and the Cran 400.

It is generally believed that feedback on a collection of lower generality will not be as effective and that feedback on a collection as large as a library is not promising. However, the results of this study to seem to point out that relevance feedback would be operative on a library collection, contrary to common belief. Of course this is highly dependent on which feedback method is used, since some strategies (such as those using a large number of nonrelevant documents) perform poorly on collections of lower generality. Furthermore, as the fall-out curves indicate, the Cran 400 collection might have a disjoint subset of documents never retrieved. Thus generality should be recomputed by removing such documents. In addition, the test collections used here are limited in that they pertain to only one subject area.

A suggestion for future experiments is that queries should be examined individually to isolate irregular behavior. Also a larger query collection and document collection on more than one subject area would be advisable to substantiate the results. Based on the findings of this study, variations of the two feedback strategies in group B — e.g. requiring a constant number of relevant documents to be fed back or using different rank cut values — should be explored.

References

[1] G. Salton, Automatic Information Organization and Retrieval, McGraw-Hill Book Company, New York, 1968, pp. 266-267.

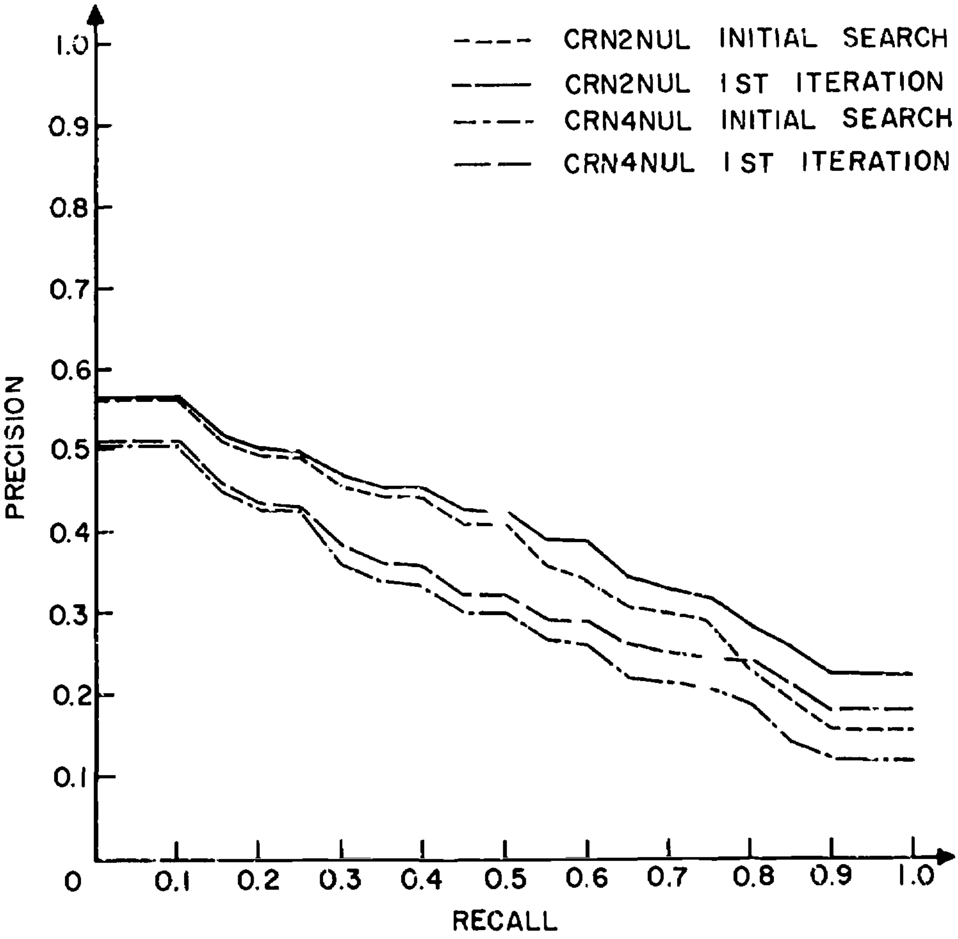
[2] E. Ide, User Interaction with an Automated Information Retrieval System, Report No. ISR-12 to the National Science Foundation, Section VIII, Deparment of Conputer Science, Cornell University, June 1967.

[3] E.M. Keen, Evaluation Parameters, Report No. ISR-13 to the National Science Foundation, Section II, Department of Computer Science, Cornell University, January 1968.

[4] C. Cleverdon and M. Keen, Factors Determining the Performance of Indexing Systems, Vol. 2 Test Results of the ASLIB Cranfield Research Project, C. Cleverdon, Wharley End, Bedford, 1966.

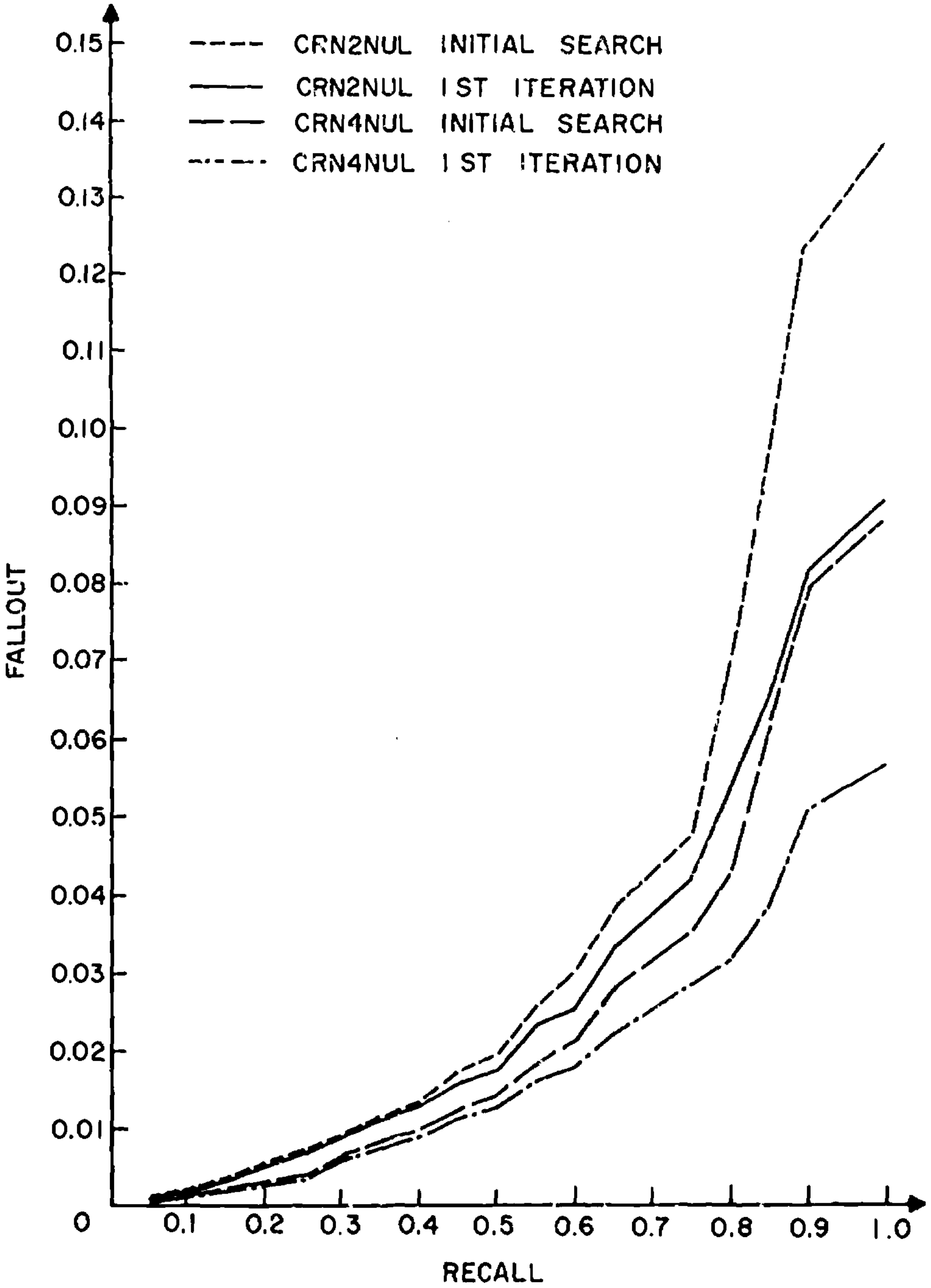
[5] G. Salton, Search Strategy and the Optimization of Retrieval Effectiveness, Report No. ISR-12 to the National Science Foundation, Section V, Department of Computer Science, Cornell University, June 1967.

Appendix



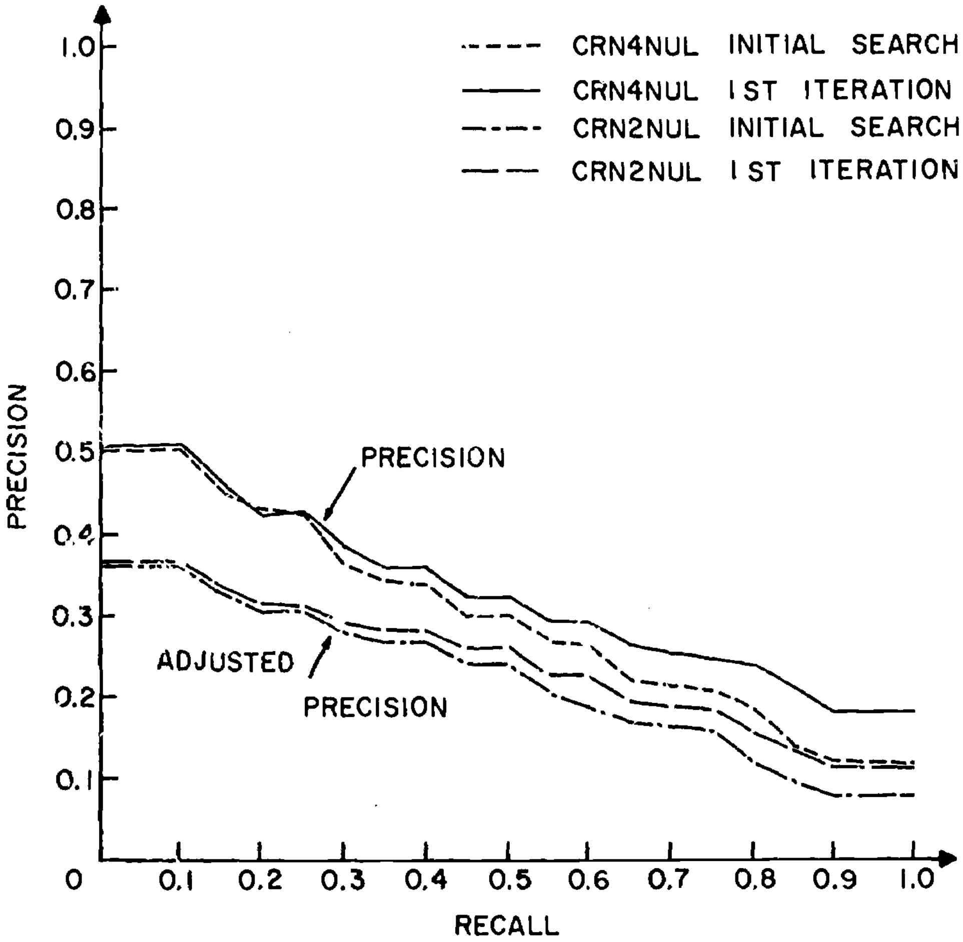
STRATEGY I

RECALL-PRECISION GRAPH



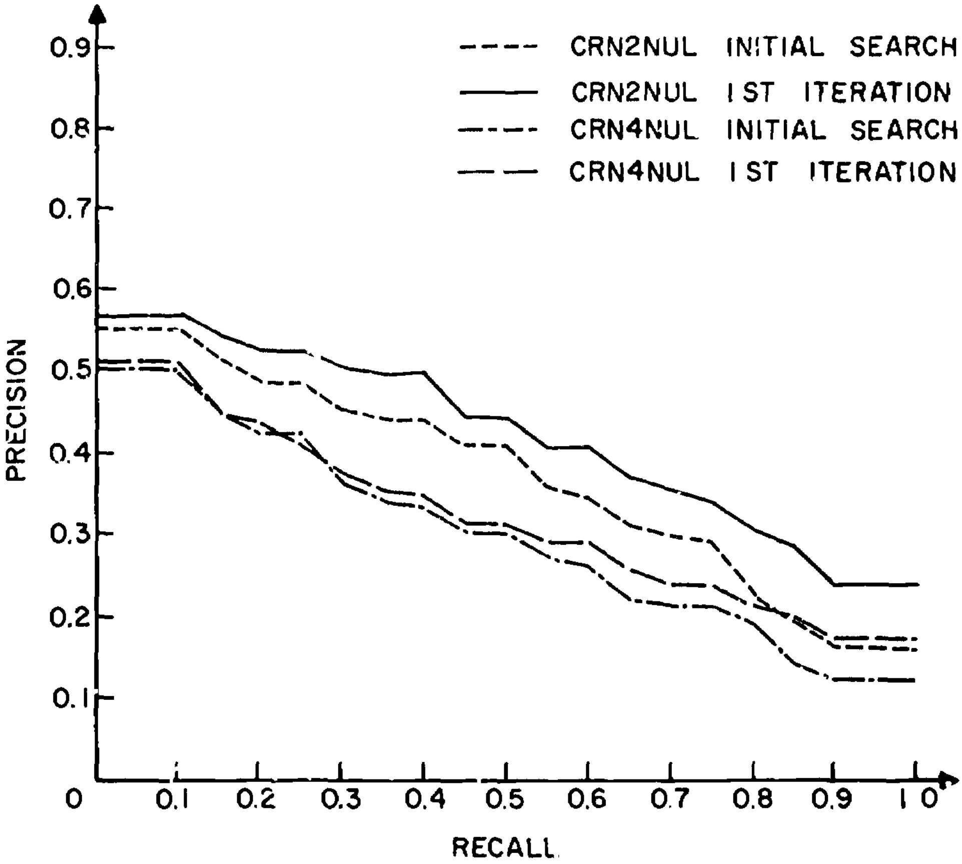
STRATEGY I

RECALL-FALLOUT GRAPH



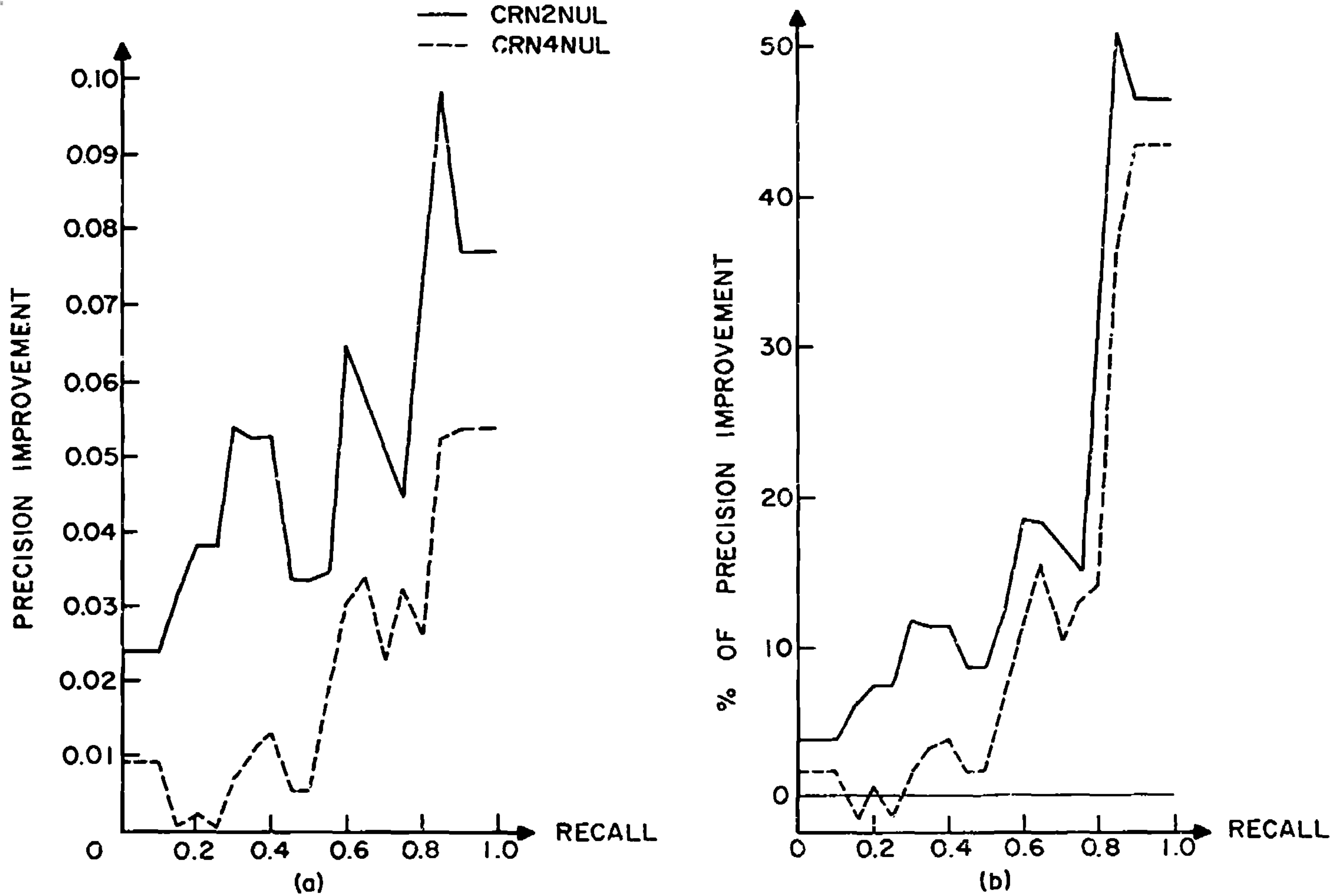
STRATEGY I

RECALL-PRECISION GRAPH WITH PRECISION OF CRAN 200 ADJUSTED



STRATEGY 3

RECALL-PRECISION GRAPH

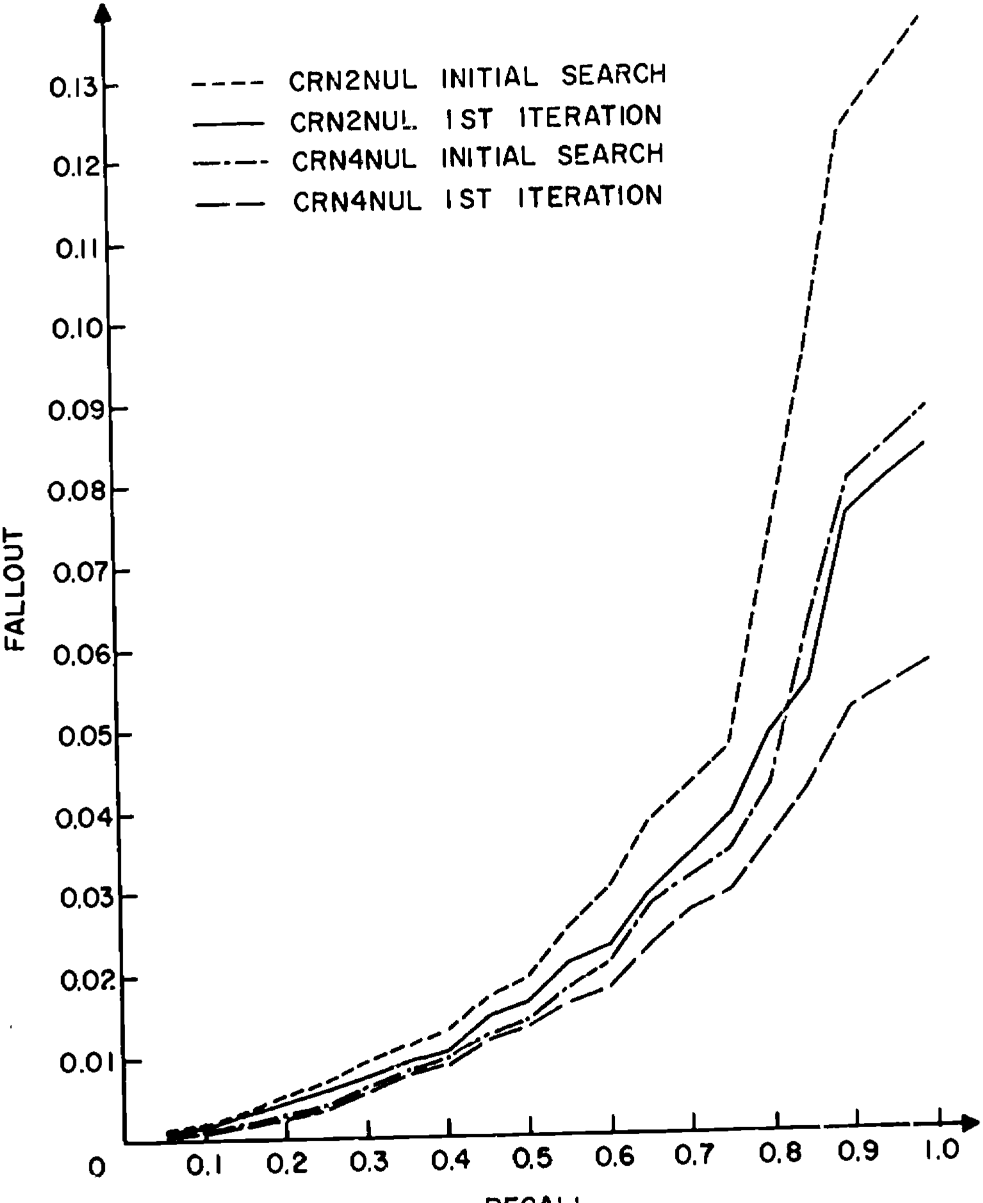


(a)

(b)

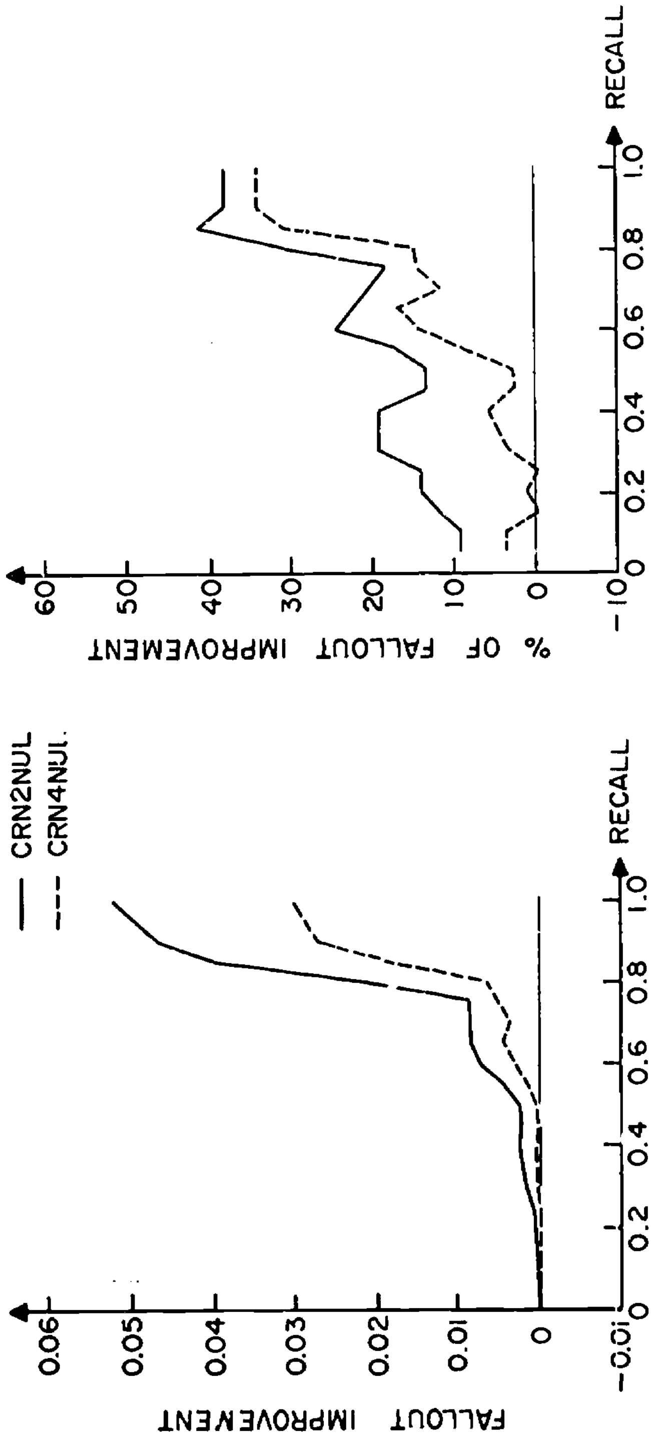
PRECISION IMPROVEMENT GRAPHS

STRATEGY 3



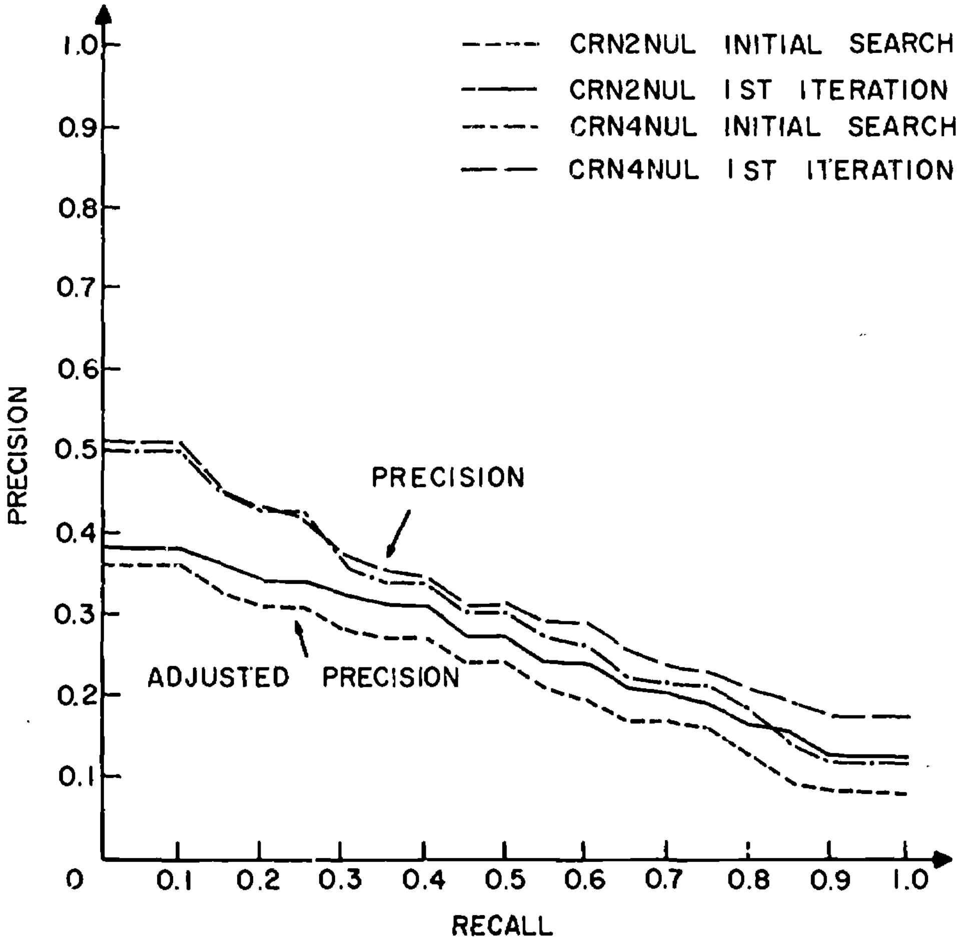
STRATEGY 3

RECALL – FALLOUT GRAPH



FALLOUT IMPROVEMENT GPAPHS

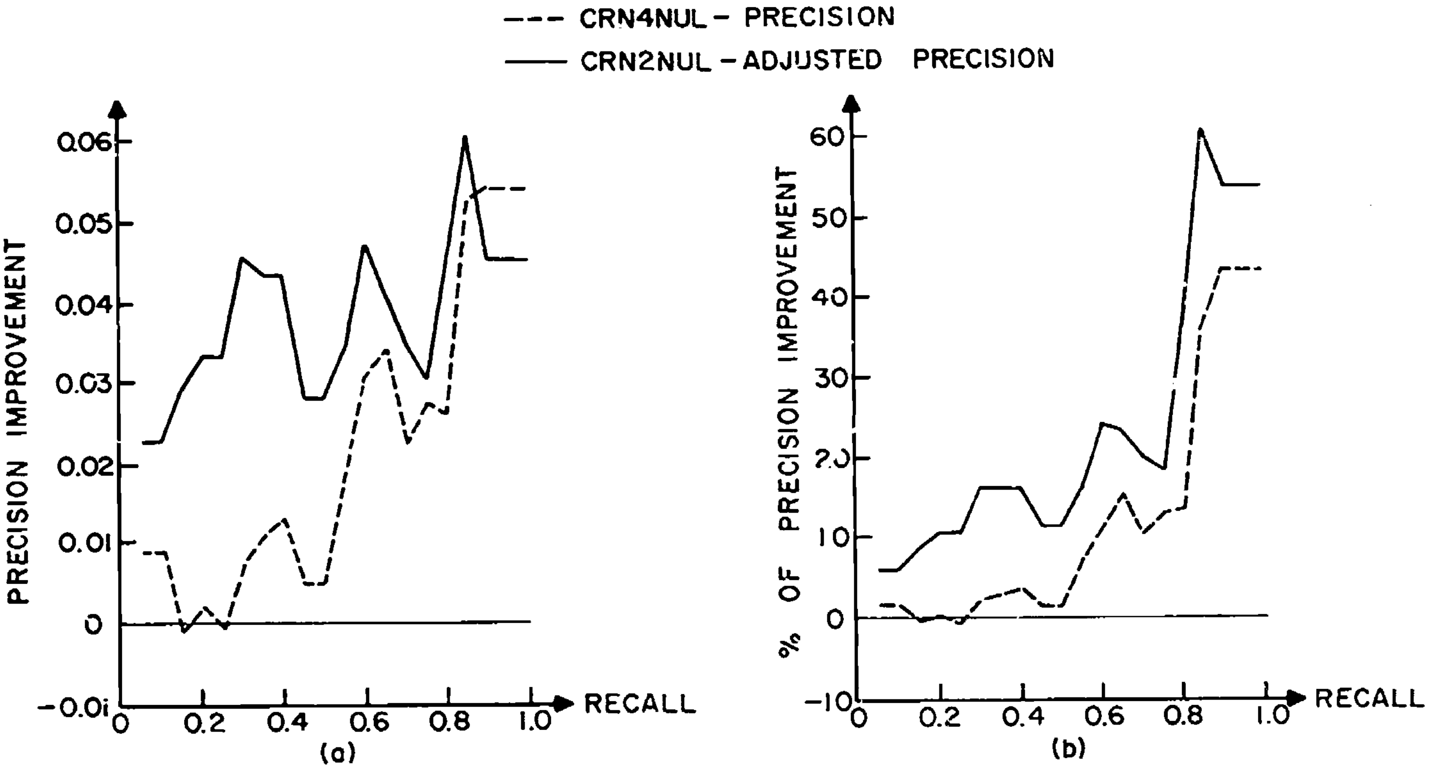
STRATEGY 3



STRATEGY 3

RECALL-PRECISION GRAPH WITH

PRECISION OF CRAN 200 ADJUSTED



PRECISION IMPROVEMENT GRAPH

WITH PRECISION OF CRAN 200 ADJUSTED

STRATEGY 3

X. Selective Negative Feedback Methods

M. Kerchner

Abstract

A great deal of work has already been done in automatic information retrieval in an effort to improve performance and to satisfy user needs. In particular various techniques have been described which modify the initial query submitted by the user, including the use of nonrelevant and relevant retrieved documents. The present study deals with experiments performed with several new methods of using nonrelevant retrieved documents to modify queries which retrieve no relevant in the first N documents retrieved. The results of the experiments are evaluated and suggestions are made for possible further investigations.

1. Introduction

Relevance feedback is a technique for improving the performance of an information retrieval system to better satisfy the needs of its users. [1] A search of the document collection is made with an initial query and a set of retrieved documents, ranked in order of correlation with the query, is presented to the user. After examining the set of retrieved documents, the user indicates whether each is relevant or not relevant to his query. [3] The relevance judgments are used by the system to modify the original search query in such a way that the modified query will retrieve additional relevant documents.

Experiments have been made with several methods of positive relevance feedback in which highly ranked relevant documents are used to modify the query. [2,4] In the case where no relevant documents are retrieved in the first N documents considered, negative relevance feedback — the use of nonrelevant documents for query modification — has been the basis for experimentation. [1,2,5] However, some problems arise with the use of nonrelevant documents for query modification. Riddle et. al. [4] and Ide [5] confirm that in some cases the use of nonrelevant documents perturbs the query vector so grossly that no additional relevant documents are retrieved in subsequent searches with the modified query. [6]

In the present study, the SMART document retrieval system is used as the basis for experiments on methods which propose to deal with the above and related problems.

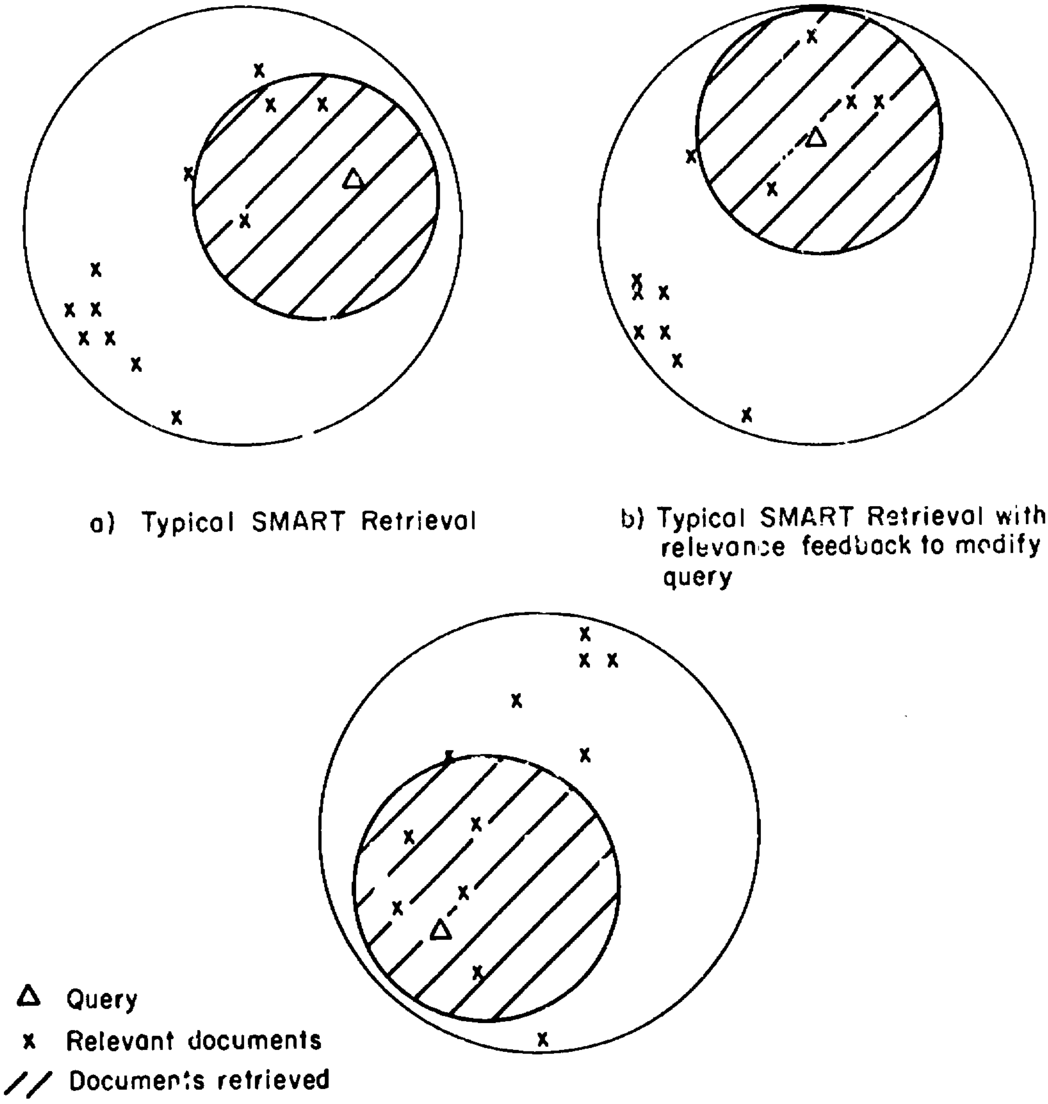
2. Methodology

It has been shown by previous work that methods using positive relevance feedback are reasonably successful for queries retrieving at least one relevant document in the first N retrieved. Therefore, the experiments in this study are only concerned with those queries which retrieve no relevant in the first N (N=5) documents retrieved.

To deal with the problem of overdistortion of the query which occurs with standard negative feedback schemes in which highly ranked nonrelevant documents are subtracted from the query, Johnson and Krablin [6] propose that more selective methods be used in order to "insure the integrity of the original relevant concepts in the query" and to move the query out of an area of nonrelevant concepts in the document space by using a series of selected terms for negative feedback. The approach suggested by Johnson and Krablin is to select those terms which appear in several of the highly correlated nonrelevant documents, but not in the original query and to add these terms, with negative weights, to the query.

In connection with this approach, it is important to note that a large portion of normal queries covers more than one subject area. [7] In addition, concepts which appear in highly correlated nonrelevant may also be significant in retrieved relevant documents. As a result, since the basic selective negative feedback strategy of Johnson and Krablin leaves untouched those concepts in the query which may have been found in several of the highly correlated <u>nonrelevant</u> documents (and, as noted, several of the <u>relevant</u> retrieved as well), the query appears to remain in approximately the same areaof the document space, as seen in Fig. 1. The highly correlated nonrelevant documents in the area may no longer be retrieved but the query also does not approach the documents relating to any secondary relevant subject area. The retrieval results confirm that most of the improvement obtained is caused by raising the ranks of the relevant documents in the primary subject area, and, in some cases, retrieving several other relevant in the same part of the document space.

In contrast, by removing those concepts in the query which are shown to be significant in the highly ranked <u>nonrelevant</u> documents, the query is moved from that part of the document space in which those documents appear, i.e. from an area of the space which is, in a sense, "more" nonrelevant than relevant to the query. It is hypothesized, as shown in Fig. 1, that the query is moved nearer to the set of documents related to its second subject area since presumably, the concepts which remain in the query relate to this area and, by removing the other concepts (or decreasing their weights), the remaining (or more weighty) concepts now assume primary importance in the query. In fact, a situation analagous to query splitting



a) Typical SMART Retrieval

b) Typical SMART Retrieval with relevance feedback to modify query

c) Typical retrieval with query modified by selective negative feedback (Methods 1,4)

Selective Negative Feedback Illustration

Fig. 1

is achieved, although relevant documents in the original areaof the document space may now be overlooked. However, while missing these relevant documents, experiments show that the query is moved significantly nearer the second subject area and more new documents in this area are retrieved than would be the case if additional documents in the first subject area were retrieved by not modifying those selected concepts which appear in the query.

3. Selective Negative Relevance Feedback Strategies

The following procedure is used in testing the various selective negative feedback methods to be described.

1. A full search is made with the original queries (Note: As mentioned above, only those queries which retrieve no relevant in the first 5 documents retrieved are used in this study.)

2. Modify the query in one of the following ways (as summarized in Table 1):

Method 1: Any concept which appears in at least 3 of the first 5 nonrelevant documents is deleted it it appears in the query. No new concepts are added to the query.

Method 2: Any concept which appears in at at least 3 of the first 5 nonrelevant documents is assigned a weight equal to the average of its weights in these documents multiplied by -1. If the concept appears in the query, its weight is replaced by the new calculated weight. If the concept does not appear in the query, it is added to the query.

Method 3: This method is similar to Method 2 but if the selected concept appears in the query, the new (negative) weight of the concept is added to its present weight in the query.

Method 4: This method is similar to Method 1 but a concept must appear in all 5 nonrelevant documents in order to be selected.

Method 5: This technique is similar to Method 2 but a concept must appear in all 5 nonrelevant documents to be selected.

3. Search the document collection with the modified query, and repeat procedure of part 2.

This process is halted when a satisfactory proportion of relevant documents are retrieved.

For comparison, searches are made with the test queries using a standard method of negative relevance feedback in which the nonrelevant document retrieved with rank 1 in the original search is subtracted from the query and a subsequent search is made with the modified query. Two feedback iterations are performed.

In Methods 1 and 4, the danger exists of reducing the query to the zero vector. It has been found that such reduction occurs after the second iteration of Method 1 with only 2 queries. However, the experiments performed indicate that two iterations are the maximum number desirable, as further iterations cause too much distortion in the query.

4. *The Experimental Environment*

The strategies outlined above have been tested on the Cranfield collection of 424 document vector abstracts produced using a word form the-

Method 1:	Any concept which appears in at least 3 of the 5 nonrelevant documents is deleted if it appears in the query. No new concepts are added to the query.
Method 2:	Any concept which appears in at least 3 of the 5 nonrelevant documents is assigned a weight equal to the average of its weight in the 5 documents multiplied by -1. If the concept appears in the query, its weight is replaced by the calculated weight. If the concept does not appear in the query, it is added to the query.
Method 3:	This method is similar to Method 2 but if the selected concept appears in the query, the calculated (negative) weight of the concept is added to its present weight in the query.
Method 4:	This method is similar to Method 1 but a concept must appear in all of the 5 nonrelevant documents in order to be selected.
Method 5:	This method is similar to Method 3 but a concept must appear in all 5 of the nonrelevant documents to be selected.

Five Proposed Selective Negative Feedback Schemes

Table 1

saurus and 155 queries, 35 of which retrieve no relevant in the first five documents retrieved. These queries are used as the experimental base. In the experiment, 15 documents are shown to the user but only the first five are used for relevance feedback.

5. Experimental Results

Since it is hypothesized that modification of the query by the proposed methods moves it to a part of the document space which represents the second subject area, it is important to consider the number of new relevant documents which are retrieved in the first 15 documents, i. e. those which have not previously been shown to the user. [7,8] As seen in Table 2, Method 1 is the most successful in retrieving new relevant documents. In one iteration 24 relevant documents appear in the first 15 documents retrieved or 15.5% of the remaining relevant documents, with an average of 3.0 concepts deleted from each query. In two iterations, a total of 30 new relevant documents are shown to the user or 19.4% of the remaining relevant in the collection for this particular set of queries. Method 4, which requires that a concept appear in all 5 nonrelevant documents in order to be deleted, retrieves 16 new documents or 10.3% of the remaining relevant, with an average of 1.6 concepts deleted from each query. Tne techniques which add concepts with negative weights to the query show inferior results. Method 2 retrieves only 9 new documents or 5.8% of the remaining relevant while Method 3 retrieves 8 new relevant documents. Thus it appears that assigning a weight of zero to a concept, i. e., deleting it from the query, results in less distortion of the query than assigning it a negative weight. In addition, Methods 1 and 4, which both neglect to add new concepts with nega-

	Method 1 (1 iter.)	Method 1 (2 iters.)	Method 2	Method 3	Method 4	Method 5
Number of queries modified	34	34	34	34	22	22
Number of relevant in first 5 retrieved	13	14	8	9	12	10
Number of relevant in first 15 retrieved	38	28	10	10	33	21
Number of new relevant in first 5 retrieved	13	16	8	9	11	9
Number of new relevant in first 15 retrieved	24	30	9	9	16	13
% of remaining relevant retrieved in first 15	15.5	19.4	5.8	5.8	10.3	8.4
Number of queries which retrieve at least 1 new relevant in the first 15	1[illegible]	24	9	9	13	11

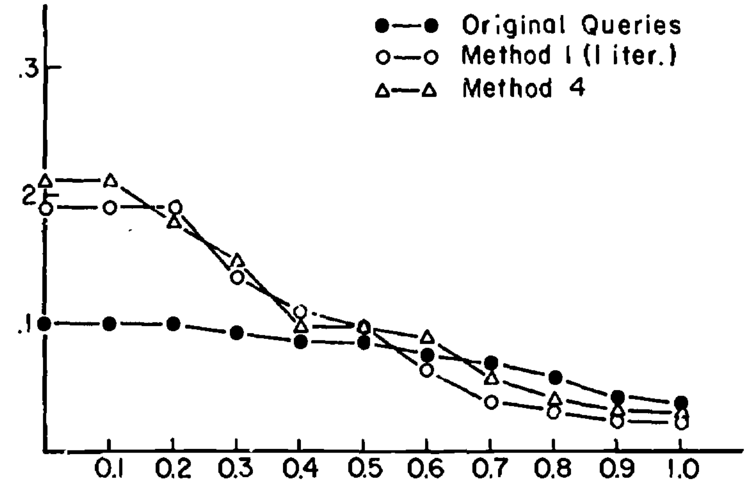
Comparison of Methods 1-5

Table 2

tive weights to the query, are significantly more successful in retrieving new relevant documents than Methods 2, 3, and 5, which do add new concepts with negative weights.

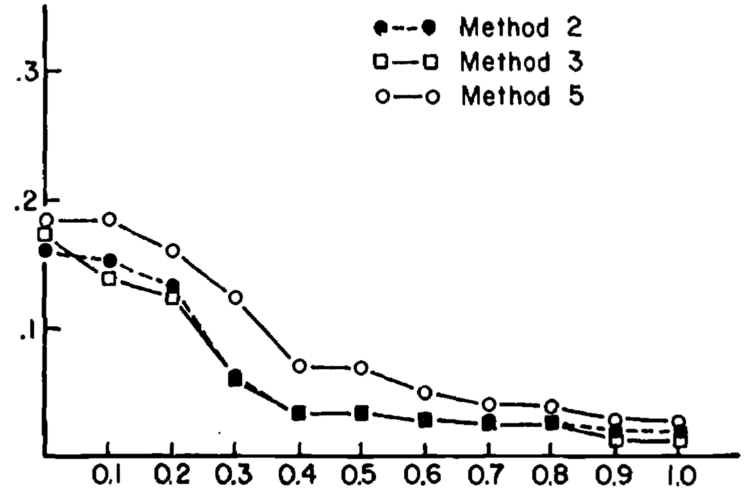
As seen in Fig. 2, the more selective modification technique of Method 4 results in higher precision figures at recall levels up to 0.5 than those achieved by Method 1, although precision figures for Method 1 are higher at the higher recall levels. It is also seen by examination of retrieval results that in some cases for Method 1, the ranks of relevant documents which are retrieved among the top 15 documents in the original search decrease significantly since, as hypothesized, the query is moving in a direction away from these highly correlated documents. As shown in Table 2, for Method 1, 24 of the 38 relevant documents retrieved, or 63%, are new relevant documents. Since Method 4 leaves 13 queries unchanged, the high ranks of these relevant documents remain the same and thus help in achieving high precision figures for Method 4 at low recall levels. In the same way, Method 1 tends to push low ranking relevant documents lower if these documents are in the area of the document space from which the query is being moved, as they tend to be. In fact, using Method 1, 47 relevant documents which have a nonzero correlation with the queries are reduced to having a zero correlation with the modified queries after one iteration. It is to be noted that some of these relevant documents have been seen by the user, as they appear in the top 15 retrieved documents, but, nonetheless, such factors affect the precision and recall calculations.

As seen in Table 3, the standard feedback technique of subtracting the nonrelevant document with rank 1 from the query only retrieves 13 new vant documents after 2 iterations, or 8.4% of the remaining relevant.



Recall–Precision Curve for Original Queries, Methods 1 and 4

Fig. 2



Recall–Precision Curve for Methods 2,3 and 5

Fig. 3

	Iteration 1	Iteration 2	Combined
Number of relevant in first 5 retrieved	3	3	4
Number of relevant in first 15 retrieved	9	12	17
Number of new relevant in first 5 retrieved	3	1	4
Number of new relevant in first 15 retrieved	8	5	13
% of remaining relevant retrieved in first 15	5.2	3.2	8.4
Number of queries which retrieve at least 1 new relevant in the first 15	5	4	9
Average number of concepts subtracted from the query	56.2	35.9	92.1

Results for Nonselective Negative Feedback Scheme

Table 3

The high average number of concepts subtracted from the query after two iterations, 92.1, may explain the poor performance as the query is probably overperturbed.

6. Evaluation of Experimental Results

As the criteria cited above (number of new relevant retrieved, etc.) as well as the statistical T- and Wilcoxon Signed Rank tests favor Methods 1 and 4 significantly over Methods 2, 3, and 5, only the former are compared with the standard nonselective negative feedback scheme and with each other.

According to the T-test, the differences in performance between Method 1 and Method 4 are statistically significant. Using measures of rank recall, log precision, normalized recall, normalized precision, and recall level averages, Method 4 is concluded to be "better" than Method 1. The Wilcoxon Signed Rank test confirms this conclusion.

The Sign test favors the nonselective negative feedback strategy over Method 1 while the same test favors Method 4 over nonselective negative feedback. However, as noted above and by others, [7,8] several other factors must be considered in evaluating the various strategies.

Methods 1 and 4 both perform better than the nonselective negative feedback scheme as reflected by the number of new relevant retrieved. This is also reflected in the standard precision-recall curves (see Tables 2 and 3, Figs. 1 and 3). As noted previously, the improved precision-recall curves for these methods do not result from simply raising the ranks of already retrieved relevant for, as shown in Table 2, 63% of the relevant documents retrieved by Method 1 are new documents not seen before by the user. For Method 4, 48% of the relevant documents retrieved are new.

To determine which of Methods 1 or 4 is to be favored, it must be considered that although the precision-recall curve of Method 4 is higher than that of Method 1 at recall levels up to 0.5, the curve for Method 1 shows higher precision at recall levels greater than 0.5, since more relevant are retrieved using Method 1 than if Method 4 is used. At low recall levels, precision may be improved by raising the ranks of relevant documents already shown to the user. As noted by Hall et al [7] and Cirillo et al [8], assuming that 15 documents are shown to the user, whether a relevant document is ranked 8 or 13 is not important to the user since he is shown both documents; it is in the higher ranks of relevant documents retrieved that Method 4 seems to show better performance figures than Method 1.

It is, in addition, important to note that Method 4, due to its more selective modification procedure which requires that a concept appear in all 5 nonrelevant documents in order to be deleted from the query, fails to alter 13 of the 35 queries while Method 1 modifies 34 of the 35 queries. For those queries which are modified, their performances as far as the number of new relevant documents retrieved are similar. Method 1 retrieves an average of .71 new documents per query and Method 4 retrieves an average of .73 new documents per query.

Since negative feedback schemes are conceived for the purpose of dealing with problem queries, i.e. those which retrieve no relevant in the first 5 documents retrieved, and thus cannot be modified by positive feedback schemes employing relevant documents, a strategy which leaves 37% of the queries unmodified must be considered unsatisfactory for the purpose for which it is designed.

Therefore, it is recommended that Method 4, which deletes from the query those concepts which appear in at least 3 of the 5 nonrelevant documents, be used as a negative feedback scheme for those queries which retrieve no relevant documents in the first 5 retrieved. However, as it is hypothesized in the present study that the large number of new relevant documents retrieved by queries modified by this strategy are obtained by moving the query to a new section of the document space, which represents its second subject area, it is necessary to perform further experiments to determine how to retrieve the relevant which remain unretrieved in that part of the document space which relates to its first subject area. A combination of such techniques would presumably result in significantly better retrieval results for the problem queries dealt with in this study.

	Initial		Nonselective		Method 1		Method 4	
Rank	Doc	Corr	Doc	Corr	Doc	Corr	Doc	Corr
1	106	.4471	372	.0261	226R	.3993	226R	.3458
2	87	.4429	321	.0148	340	.3592	340	.3111
3	74	.4248	373	.0097	227R	.3118	225R	.2725
4	27	.4025	103	.0	238	.3093	321	.2722
5	128	.3643	197	.0	244	.3020	227R	.2700
6	91	.3593	241	.0	267	.2993	238	.2679
7	72	.3542	264	.0	167	.2867	244	.2616
8	83	.3539	267	.0	372	.2774	267	.2592
9	387	.3501	273	.0	225R	.2697	167	.2483
10	107	.3476	320	.0	339	.2649	372	.2402
11	167	.3441	106	-.9917	321	.2357	339	.2294
12	234	.3274	107	-.5190	270	.2200	270	.2223
13	225R	.3237	91	-.4715	374	.2025	228R	.1816
14	62	.3227	36	-.4627	243	.1992	374	.1754
15	65	.3227	415	-.4446	242	.1911	243	.1725

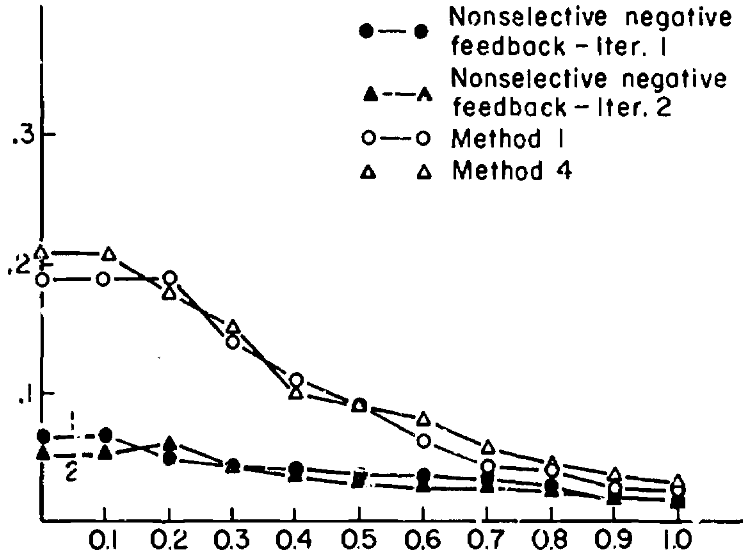
a) Three Negative Feedback Strategies for Query 34

Rank	Doc	Corr	Doc	Corr	Doc	Corr	Doc	Corr
1	73	.3230	73	-.9971	73	.3322	163R	.2011
2	406	.2926	174	-.4847	406	.3084	202	.1964
3	40	.2363	133	-.4199	40	.2491	413	.1474
4	367	.2349	134	-.4158	174	.2430	385	.1367
5	398	.2333	398	-.4071	74	.2185	203	.1297
6	174	.2305	406	-.4054	7	.2148	384R	.1235
7	381	.2173	419R	-.4032	367	.2063	61	.1223
8	74	.2073	234	-.3997	90R	.2010	90R	.1066
9	7	.2038	381	-.3737	234	.1991	73	.1057
10	163R	.1962	28R	-.3733	398	.1933	122	.1003
11	90R	.1907	136	-.3674	394	.1907	26	.0898
12	234	.1889	40	-.3593	202	.1852	70	.0898
13	394	.1809	7	-.3513	65	.1811	22	.0884
14	202	.1757	74	-.3486	64	.1794	39	.0881
15	65	.1718	376	-.3485	163R	.1724	7	.0876

b) Three Negative Feedback Strategies for Query 137

Retrieval Results for Three Negative Feedback Strategies

Table 4



Recall-Precision Curve for Nonselective Negative Feedback Technique, Methods 1 and 4.

Fig. 4

X-18

Query	Orig. Search	Method 1 1st iter	New Rel	Method 2 2nd iter	New Rel	Total New	Method 2	New Rel	Method 3	New Rel
3	3	3	0	3	0	0	1	0	1	0
6	1	2	1	2	0	1	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	1	1	1
12	1	4	3	5	1	4	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0
29	1	1	0	0	0	0	0	0	0	0
32	0	4	4	0	0	4	1	1	1	1
33	1	1	1	2	0	1	0	0	0	0
34	1	3	2	0	0	2	1	1	1	1
37	1	1	0	0	0	0	0	0	0	0
38	1	0	0	0	0	0	0	0	0	0
45	1	2	1	0	0	1	0	0	0	0
46	2	4	3	3	1	4	1	1	1	1
48	0	1	1	1	0	1	1	1	1	1
54	0	0	0	1	1	1	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0
58	1	0	0	0	0	0	0	0	0	0
66	1	0	0	1	1	1	0	0	0	0
74	2	1	1	1	0	1	0	0	0	0
76	1	2	1	1	0	1	1	1	1	1
78	0	0	0	0	0	0	0	0	0	0
79	1	1	1	0	0	1	0	0	0	0
83	1	0	0	0	0	0	0	0	0	0
91	2	1	1	0	0	1	0	0	0	0
95	0	0	0	1	1	1	2	2	2	2
96	0	1	1	2	1	2	0	0	0	0
102	0	1	1	0	0	1	0	0	0	0
103	0	0	0	0	0	0	0	0	0	0
113	2	1	0	0	0	0	0	0	0	0
118	0	2	2	2	0	2	0	0	0	0
134	1	0	0	1	0	0	0	0	0	0
137	2	2	0	2	0	0	1	1	1	1
140	1	0	0	0	0	0	0	0	0	0

mber of Relevant in First 15 Retrieved for Various Feedback Methods

Table 5

Query	Method 4	New Rel	Method 5	New Rel	Nonsel 1st iter	New Rel	Nonsel 2nd iter	New Rel	Total New
3	3	0	0	0	0	0	0	0	0
6	2	1	2	1	0	0	1	1	1
8	0	0	0	0	0	0	0	0	0
10	0	0	0	0	1	1	0	0	1
12	1	0	2	1	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0
29	1	0	1	0	1	1	2	1	2
32	0	0	0	0	2	2	1	0	2
33	1	0	1	0	0	0	0	0	0
34	4	3	2	2	0	0	0	0	0
37	1	0	1	0	0	0	0	0	0
38	1	0	1	0	0	0	0	0	0
45	0	0	0	0	0	0	2	2	2
46	4	3	2	2	0	0	0	0	0
48	1	1	1	1	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0
55	1	1	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0
66	1	0	1	0	3	2	0	0	2
74	1	1	1	1	0	0	0	0	0
76	2	1	1	1	0	0	0	0	0
78	0	0	0	0	0	0	0	0	0
79	1	0	1	0	0	0	0	0	0
83	0	0	0	0	0	0	0	0	0
91	1	1	0	0	0	0	0	0	0
95	0	0	0	0	0	0	0	0	0
96	0	0	0	0	0	0	0	0	0
102	0	0	0	0	0	0	1	1	1
103	0	0	1	1	0	0	1	1	1
113	1	0	0	0	0	0	1	0	0
118	1	1	0	0	0	0	0	0	0
134	1	0	1	0	0	0	0	0	0
137	3	2	1	1	2	2	2	0	2
140	1	1	1	1	0	0	1	1	1

er of Relevant in First 15 Retrieved for Various Feedback Methods

Table 5 (contd.)

References

[1] J. Kelly, Negative Response Relevance Feedback, Information Storage and Retrieval, Scientific Report ISR-12 to the National Science Foundation, Section IX, Department of Computer Science, Cornell University, June 1967.

[2] E. Ide and G. Salton, Interactive Search Strategies and Dynamic File Organization in Information Retrieval, Information Storage and Retrieval, Scientific Report ISR-16 to the National Science Foundation, Section XI, Department of Computer Science, Cornell University, September 1969.

[3] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill, Inc., New York 1968.

[4] W. Riddle, T. Horwitz, and R. Dietz, Relevance Feedback in an Information Retrieval System, Information Storage and Retrieval, Scientific Report ISR-11 to the National Science Foundation, Section VI, Department of Computer Science, Cornell University, June 1966.

[5] E. Ide, New Experiments in Relevance Feedback, Information Storage and Retrieval, Scientific Report ISR-14 to the National Science Foundation, Section VIII, Department of Computer Science, Cornell University, September 1968.

[6] P. Johnson and L. Krablin, Proposal for Modified Nonrelevant Feedback. Computer Science 635 Term Project Report, June 1969.

[7] H. Hall and N. Weiderman, The Evaluation Problem in Relevance Feedback Systems, Scientific Report ISR-12 to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, June 1967.

[8] C. Cirillo, Y. K. Chang, and J. Razon, Evaluation of Feedback Retrieval using Modified Freezing, Residual Collection and Test and Control Groups, Scientific Report ISR-16 to the National Science Foundation, Section X, Department of Computer Science, Cornell University, September 1969.

XI. The Use of Past Relevance Decisions in Relevance Feedback

L. Paavola

Abstract

A high degree of similarity may be expected to exist among documents judged to be relevant to the same query. This paper investigates some possibilities for exploiting this potential similarity in relevance feedback. Runs are made on the ADI and Cranfield 424 collections of the SMART retrieval system. In these runs all "jointly relevant" documents are incorporated into feedback as if they were a single relevant document. Standard recall-precision evaluation measures are used, and the performance of some individual queries is illustrated. Some directions for further research are suggested.

1. Introduction

In the SMART system, statistical and syntactic analyses of search queries and documents are used for text analysis, and automatic comparisons of analyzed queries to documents or to sets of centroids of document clusters are used for the selection of documents to be displayed to query authors. [1] However, the utility of these methods alone is severely limited, and attempts have been made to introduce subjective judgments into the retrieval process. The usual method, known as relevance feedback, uses a query author's decisions about the relevance to his query of specified documents in order to modify the vector representation of his query. [1, section 7-4] Occasionally such judgments are used to modify document vectors. [2] Methods which do not alter query or document vectors include query splitting [3] and query clustering. [4,5]

2. Assumptions and Hypotheses

This paper details another method of using the history of a system to improve its performance. The assumption is made that if a given document is known to be relevant to a query, another document is more likely to be relevant to the query if both have been judged relevant to some past query. It is further assumed that the number of such past occurrences of joint relevance may be a useful index to inter-document similarity.

The following problems may be anticipated in such a system: the system may be handicapped in dealing with queries of a type which it has not encountered frequently earlier; user ideas of relevance and nonrelevance may differ widely; unless special measures are taken, documents which may be relevant to a given query but never initially retrieved (e.g. situations in which query splitting would be in order) may become increasingly less likely ever to be retrieved.

The proposed method is expected to have the following advantages: general queries with a high number of relevant documents may establish a loose connection between documents of the same general subject area, while specific queries may set up stronger connections between more closely similar documents; the system may function well for the "average" user, if queries do not vary too widely; groups of documents of which all are relevant to each of several queries may be used to better performance.

3. Experimental Method

The procedure is first tried on the ADI collection of the SMART retrieval system, consisting of 82 documents and 35 queries, then on the Cranfield 424 collection, which has 424 documents and 155 queries. In each case, the query collection is divided into two equal groups by random methods. The documents relevant to each query are known. From the relevance decisions for the queries in the first group a list is made for each document of the other documents with which it has been included in such decisions and the number of times for each, as shown in Fig. 1.

The other half of the query collection is used to make three searches of the entire document collection. The first search is a full search using unaltered query vectors. The second search incorporates in positive feedback those documents among the first five shown the user which are judged relevant by him. The third search alters the query vectors in the way described below.

In general, the altered query is constructed according to the following formula:

$$q = a_0 q_0 + a_1 \left(\sum_{i=1}^{N_R} D_{R_i} \right) + a_2 \left(\sum_{i=1}^{N_{JR}} (a_3 n_{D_i} + a_4 n_{J_i}) \, D_{JR_i} \right) \quad ,$$

ere q = the altered query

Documents 1, 3, 35, 36, 89 are relevant to query 1

Documents 2, 8, 35, 36, 89, 90 are relevant to query 2

Documents 4, 36, 90 are relevant to query 3

Document	J-r docs*	#**
1	3	1
	35	1
	36	1
	89	1
2	8	1
	35	1
	36	1
	89	1
	90	1
3	1	1
	35	1
	36	1
	89	1
4	36	1
	90	1
8	2	1
	35	1
	36	1
	89	1
	90	1
35	1	1
	2	1
	3	1
	8	1
	36	2
	89	2
	90	1
36	1	1
	2	1
	3	1
	4	1
	8	1
	35	2
	89	2
	90	2
89	1	1
	2	1
	3	1
	8	1
	35	2
	36	2
	90	1
90	2	1
	4	1
	8	1
	35	1
	36	2
	89	1

* Documents joint-relevant to the given document

** Number of times each joint-relevant document occurs in a list of relevant documents with the given document

Examples of Joint Relevance

Fig. 1

q_0 = the initial query

D_{R_i} = the relevant documents among the top n (here n=5), according to the ranking produced by the full search

N_R = the number of such documents D_{R_i}

D_{JR_i} = documents joint relevant to any of the D_{R_i}

N_{JR} = the number of such documents D_{JR_i}

n_{D_i} = the number of D_{R_i} to which a particular D_{JR_i} has been found to be joint relevant

n_{J_i} = total number of joint relevancy decisions of the particular D_{JR_i} with any of the D_{R_i}

a_0 =
a_1 =
a_2 = adjustable parameters
a_3 =
a_4 =

(One may choose to include in feedback only those D_{JR_i} which have n_{J_i} greater than a certain minimum value.)

An example of the use of this notation is given in Fig. 2.

The particular coefficients that have been tried for the Cranfield 424 collection are $a_0 = 100$, $a_1 = 100$, $a_2 = 100/(\sum_{i=1}^{N_{JR}} n_{D_i})$,), and $a_4 = 1$. Parameter a_2 is normalized because some documents

Documents 5, 6, 89, 312, and 400 shown to user.

He identifies 6, 89, and 400 as relevant.

Joint relevance lists for these documents:

6		89		400	
32	3	51	1	5	2
51	3	71	1	89	1
65	1	212	1	93	1
212	1	400	1	284	1
312	2				
400	2				

$D_{R_1} = 6$, $D_{R_2} = 89$, $D_{R_3} = 400$; $N_R = 3$; $N_{JR} = 10$

i	D_{JR_i}	n_{D_i}	n_{J_i}
1	5	1	2
2	32	1	3
3	51	2	4
4	65	1	1
5	71	1	1
6	93	1	1
7	212	2	2
8	284	1	1
9	312	1	2
10	400	2	3

Computation of Joint Relevance Parameters

Fig. 2

in the collection have an extremely large number of joint-relevant documents, while others have none.

The successive definitions of the query of Fig. 2 are illustrated in Fig. 3. The query used for the pure positive feedback search and that used for the joint relevance search are always identical except that in the latter certain concepts have increased weight and other concepts are added.

To obtain a final evaluation, the simple feedback and joint relevance runs are compared to each other (and to the full search) by the AVERAGE and VERIFY routines.

4. Evaluation

The run on the ADI collection shows enough difference between the two methods to merit a run on the Cranfield collection. The chi square probabilities were 0.0001 for the t-test; 0.0483 for the sign test without ties, 1.0000 with ties; and 0.0006 for the Wilcoxon test.

Recall-precision for the Cranfield 424 collection are displayed in Fig. 4. The higher precision at low recall for simple positive feedback is probably due to the inability of a vector loaded with many concepts to be very accurate in choosing the highest-ranking documents, although performing well on the whole. From the graph of Fig. 4, it is seen that the simple feedback method is more advisable than the particular joint relevance strategy tried when only the ranking of the documents at the top is important.

Of the relevant documents which were changed in rank by

Search 1

$$q = q_0$$

Search 2

$$q = q_0 + 1(6) + 1(89) + 1(400)$$

Search 3

$$\begin{aligned} q = 100q_0 &+ 100(6) + 100(89) + 100(400) \\ &+ \frac{100}{25}\,(2(5) + 3(32) + 4(51) + 1(65) \\ &+ 1(71) + 1(93) + 2(212) + 1(284) \\ &+ 2(312) + 3(400)) \end{aligned}$$

3(400), e.g., means document 400 is added in with weight 3.

Query Alteration

Fig. 3

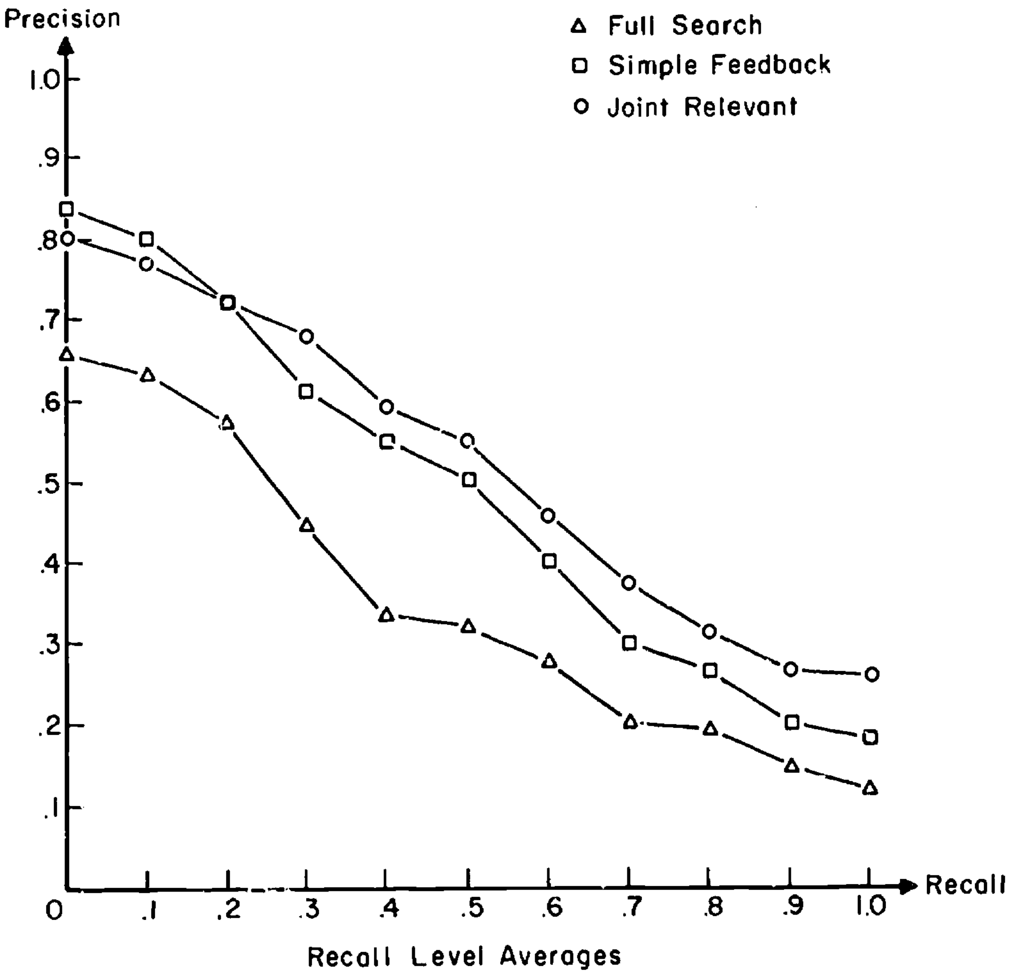

Precision
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
Recall
Recall Level Averages
△ Full Search
□ Simple Feedback
○ Joint Relevant

Fig. 4

the joint relevance process, 59 obtain lower ranks and 127 receive higher ranks. The probability under the null hypothesis of a chi square larger than that observed is 0.0000 for the t-test, Wilcoxon test, and sign test without ties; for the sign test using ties the probability is 1.0000. The large number of ties can be attributed to the lack of joint relevance information to be added into many of the queries.

Performance of the simple feedback and joint relevance searches are shown for several queries in Fig. 5. Sometimes the addition of joint relevance information does not substantially affect the effectiveness of the query one way or the other (e.g. query 54). Sometimes it actually moves the query away from relevant documents (query 26). But often it produces dramatic improvement (query 77). Sometimes the improvement is due to the direct addition of relevant documents (query 13), some of which would have been more effective had they had greater weight. Sometimes very few relevant documents are added, but the important concepts are nevertheless amplified by inclusion of joint relevance information (query 42). Sometimes the inclusion of both produces improvement (query 7). Sometimes the additions dilute the query (query 61).

The above analysis supports the conclusion that inclusion of joint relevance information, even if restricted to the weight of one relevant document only, produces significant improvement.

In evaluating this experiment one must keep in mind the differences between the experimental situation and an actual one. The results are biased positively by the fact that in an actual

Query 7		
	F	J
1	54R	54R
15	173	7R
41	7R	
64		123R
70		56R
91	123R	
102	56R	
107		121R
132		126R
137	126R	
138		124R
147		122R
160	121R	120R
168		125R
219	124R	
231	120R	
233	122R	
249	125R	

Query 13		
R	F	J
1	72R	72R
2	62R	62R
3	83	5R
4	91	6R
14	5R	234
54	5R	

Query 26		
R	F	J
1	58R	58R
3	378R	169
11	184	91R
33	34R	
42	91R	
49		378R
57	151R	
58		34R
96	78R	
98		151R
145		78R

Query 42		
R	F	J
1	294R	294R
3	80	293R
5	296	292R
7	287	290R
9	293R	291R
11	7R	289R
20	94	43R
39	292R	
50	43R	
104	289R	
120	290R	
235	291R	

Query 54		
R	F	J
1	124R	124R
2	125R	125R
8	55R	55R
13	411	136R
14	392	173R
21	82R	351
23	79	82R
28	30R	112
29	388R	61
37	173R	
38	136R	
39	10R	
48		10R
49		100R
50	101R	
54		30R
55	100R	101R
60		388R
61		164R
76		184R
80	164R	
81	38R	
86	118R	
87		391R
88	391R	
102	364R	
106	150R	
116	384R	
118		364R
128		38R
142		118R
150		150R

Query 61		
R	F	J
1	141R	142R
2	142R	141R
3	143R	143R
5	145R	145R
6	221R	144
7	140	221R

Query 71		
R	F	J
1	89R	418R
2	344	89R
3	264	420R
5	418R	268
38	420R	

Key:

R = Rank

F = Simple Feedback Search

J = Joint Relevance Search

The D_{JR_i} are given in the following form:

33(292) implies that document 292 was added in with weight 33

Query 7: 9(7), 5(11), 5(40), 9(48), 100(54), 9(56), 9(98), 9(120), 5(121), 9(122), 9(123), 9(124), 9(125), 5(126)

Query 13: 33(5), 33(6), 33(62), 100(72)

Query 26: 33(7), 100(58), 33(80), 33(169)

Query 42: 33(92), 33(293), 100(294), 33(307)

Query 54: 8(7), 4(48), 8(54), 4(55), 8(56), 4(59), 4(98), 12(120), 4(121), 8(122), 8(123), 100(124), 8(125), 4(126), 4(308), 4(310), 4(311)

Query 61: 100(141), 100(142), 100(143), 100(144), 100(145)

Query 71: 100(401)

Search Results for Joint Relevance Modification

Fig. 5

application not all the documents relevant to a given query would be shown to a user, identified as relevant, and added to the joint relevance lists. They are biased negatively by the fact that relevance information obtained by a query in the second half of the collection might often help a new query subsequently submitted in the second half; this effect could not be taken into account in the experimental design. A sounder though more laborious experiment would have been to run the entire query collection against the document collection, while updating the joint relevance lists after each query. Still more significant results would have been obtained had the joint relevance lists been composed of only those documents which a user might see and identify as relevant. However, such experiments are difficult to perform without adequate system support.

5. Conclusions

The assumptions of part 2 are found to be largely justifiable, although the importance of the number of past joint relevance decisions should be further investigated. The danger of biasing the system toward one type of query is avoided, since the two halves of the query collection are fairly similar. The experiments are not extensive enough to detect isolation of documents. As expected, loose and strong connections are established by general and specific queries, respectively. The joint relevance procedure does take advantage of document groups. And a partial but important answer to the weighting problem is that greater emphasis should be placed on the joint relevant documents, although waysmust be found to counteract the negative effects of such an increase. Pe:haps this effect may be partially counteracted as the number of queries run through a system increases.

In a new experiment, low-weight concepts might be eliminated from altered queries. Certainly better values for a_0, a_1, a_2, a_3, and a_4 should be found. There may be possibilities for the use of joint-relevance information in negative feedback. Incorporation of the best known feedback strategies into the joint-relevance query alteration equation should be attempted. Perhaps high-frequency occurrence in joint-relevance lists of a document already known to be relevant should lead to a higher weighting of such a document.

The experimental data indicate that the use of joint relevance information is a valuable tool in information retrieval, that more testing of procedures for using this information is in order, and that the nature of the tradeoff between computational complexity and effectiveness of additional information must be determined for such procedures.

References

[1] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill, Inc., New York 1968.

[2] T. L. Brauen, Document Vector Modification in On-Line Information Retrieval Systems, Information Storage and Retrieval, Report ISR-17 to the National Science Foundation, Department of Computer Science, Cornell University, September 1969.

[3] A. Borodin, L. Kerr, F. Lewis, Query Splitting in Relevance Feedback Systems, Information Storage and Retrieval, Report ISR-14 to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, October 1968.

[4] V. R. Lesser, A Modified Two-Level Search Algorithm Using Request Clustering, Information Storage and Retrieval, Report ISR-11 to the National Science Foundation, Section VII, Departmentof Computer Science, Cornell University, June 1966.

[5] S. Worona, Query Clustering in a Large Document Space, Information Storage and Retrieval, Report ISR-16 to the National Science Foundation, Section XV, Department of Computer Science, Cornell University, September 1969.