Search and Retrieval Experiments in
Real-Time Information Retrieval

G. Salton

Technical Report
No. 68-8

Department of Computer Science
Cornell University, Ithaca, N. Y.

Search and Retrieval Experiments in
Real-Time Information Retrieval

G. Salton*

Abstract.

Future operating document retrieval systems may be based on fully-automatic information analysis methods instead of manual indexing, and on real-time search procedures which allow the user to interact with the system during the search process.

Performance characteristics are first given for fully-automatic information retrieval systems, and comparisons are made with presently operating partly-manual systems. Thereafter, various user-controlled search strategies are described, and the potential of these strategies in improving systems performance is discussed. The evaluation results for the real-time retrieval procedures are used to derive design criteria for future automatic information systems.

1. Introduction

Throughout the world, the design and operations of large-scale information systems has become of concern to an ever-increasing segment of the scientific and professional population. Furthermore, as the amount and complexity of the available information has continued to grow, the use of mechanized or partly mechanized procedures for various information storage and retrieval tasks has also become more widespread. As a result, a number of large information systems are now in operation in which at least the search operations — that is, the comparison of incoming search

*Dept. of Computer Science, Cornell University, Ithaca, New York.

This study was prepared for presentation at the IFIP Congress 68', and was supported in part by the National Science Foundation under grant GN-495.

requests with stored information items is carried out automatically. Typical examples in the United States are the NASA Scientific and Technical Information Facility, and the Modlars system at the National Library of Medicine.

While these operational information systems are thus able rapidly to search vast storage files, often containing many hundreds of thousands of items, most of the operations other than the search itself are performed manually with the help of human experts. In particular, all the content analysis and indexing operations, leading to the assignment of suitably chosen combinations of index terms to the stored documents and to incoming search requests are normally performed by specialists who know the given subject area, as well as the performance characteristics of the retrieval environment within which they operate.

As will be seen in the next section, many of the information systems which base their operations on manual indexing but largely automatic search methods are quite successful in isolating from the large mass of largely irrelevant stored material, many of the items which prove pertinent to the users' information needs. Nevertheless, the feeling that manual systems and procedures should be replaced by suitably chosen automatic methods has continued to grow, and a number of fully-automatic information storage and retrieval systems have been designed and put into operation, at least on an experimental basis. The SMART system represents one such effort to replace the intellectual indexing by sophisticated automatic text analysis procedures, and thereby to produce a retrieval environment in which all document and query handling procedures are performed automatically [1,2,3].

In the next section, the performance characteristics of presently operating information systems are briefly outlined and a comparison is made with the performance of the alternative fully-automatic environment. Various procedures are then described leading to an improvement in systems performance, and conjectures are made concerning the design of future automatic information systems.

2. Performance Characteristics of Information Systems

Many different criteria may suggest themselves for measuring the performance of an information system. In the present context, the system effectiveness is assumed to depend on its ability to satisfy the users' information needs by retrieving wanted material, while rejecting unwanted items. Two measures have been widely used for this purpose, known as recall and precision, and representing respectively the proportion of relevant material actually retrieved, and the proportion of retrieved material actually relevant [4]. (Ideally, all relevant items should be retrieved, while at the same time, all nonrelevant items should be rejected, as reflected by perfect recall and precision values equal to 1).

It should be noted that both the recall and precision figures achievable by a given system are adjustable, in the sense that a relaxation of the search conditions often leads to high recall, while a tightening of the search criteria leads to high precision. Unhappily, experience has shown that on the average recall and precision tend to vary inversely since the retrieval of more relevant items normally also leads to the retrieval of more irrelevant ones. In practice, a compromise is usually made, and a performance level is chosen such that much of the relevant material is retrieved, while the number of nonrelevant items which are also retrieved is kept within tolerable limits. As an example, the MEDLARS system at the National Library of Medicine is said to achieve an average recall of 0.58 and a precision of 0.50; that is, an average user may expect to retrieve 58 percent of the relevant material contained in the system, while only one half of the retrieved material is actually unwanted. Obviously, other operating points are achievable by altering the indexing or search techniques, producing either higher recall normally with low precision, or higher precision with lower recall.

In order to make it possible to design more effective retrieval systems, it is important to be aware of the reasons for system failure under presently existing operating conditions. Tables 1 and 2 summarize the performance for 18 queries processed by the MEDLARS system. It may be seen from Table 1 that a large proportion of the recall failures of the system (that is, failures to retrieve relevant material) is due to the lack of sufficient user-system interaction, while Table 2 reveals that many of the precision failures (that is, failures to reject irrelevant material) are caused by lack of specificity in the document indexing. In both cases, the manual analysis process is at least partially to blame, since the query analyses as well as the document indexing are performed by experts who cannot — in the absence of specific information concerning the user population — always be aware of the appropriate indexing requirements.

In the automatic SMART system, an attempt is made to replace the intellectual indexing effort used in the conventional situations by a fully-automatic computer analysis of the document and query texts. Specifically, a variety of language analysis procedures — including suffix cut-off methods, thesaurus look-up, phrase generation methods, statistical term associations, syntactic analysis, and others — are used to reduce document and query texts into analyzed concept (or term) vectors. The concept vectors attached to the documents are then matched with the vectors derived from the search requests, and the documents, arranged in decreasing query correlation order, are submitted to the user as answers to the query.

One might expect that an automatic text analysis of the type used in the SMART system would necessarily produce retrieval results which are much inferior to those obtained in a system based on manual indexing. In actual fact, the automatic environment makes it possible to use for analysis purposes relatively large sections of text, such as abstracts or summaries, thus insuring a high degree of indexing exhaustivity; furthermore, the importance of some assigned terms can be enhanced by automatic weighting methods, leading to a more sophisticated matching process between analyzed queries and documents than is normally possible. As a result, the benefits of the index language control supplied in the more conventional retrieval situation by the human indexers appear to be balanced by a deeper and more complex type of analysis available in an automatic environment; this is reflected by evaluation results which indicate that the search effectiveness of the fully-automatic systems is not inferior to that obtainable at present in a partly-manual system.

Consider as an example, the performance characteristics of the SMART system, presented in the recall-precision graphs of Fig. 1. The left-hand graph (Fig. 1(a)) shows the performance of an automatic word-stem matching procedure, using weighted word stems extracted from document abstracts and from search requests respectively, for three different document collections in the areas of documentation (ADI), aerodynamics (CRAN), and computer engineering (IRE). The performance for the three collections is generally comparable, ranging from an average precision of 0.80 at recall of 0.10, to an average precision between 0.10 and 0.30 for a recall of 1. Fig. 1(a) also exhibits the large distance which remains to be covered between the performance of the automatic word-stem matching system and that of an ideal system (in the upper right-hand corner of the graph) where both recall and precision values are close to 1.

A variety of more refined language analysis procedures can be used in an attempt to improve the performance of automatic information systems [4,5]. The use of a stored thesaurus capable of recognizing synonymons and other closely related terms is one such process which can be incorporated into an automatic content analysis system. Fig. 1(b) represents average performance for 780 documents in the computer field for both the word-stem matching process, and a thesaurus look-up method in which all word stems are first replaced by thesaurus group entries (concepts) prior to the comparison between query and documents. It may be seen that the synonym recognition implicit in the thesaurus procedure serves to improve the retrieval performance by approximately ten percent. Other language analysis methods produce some further improvement, but do not come close to the ideal performance area [5].

A comparison between the automatic SMART procedures using abstract processing, and the conventional operational retrieval situations based on the matching of manually assigned index terms shows that the retrieval effectiveness achievable is quite comparable in the two cases. Table 3(a) shows "normalized recall" and "normalized precision" values averaged over 42 queries for 200 documents in aerodynamics formerly used as part of the Aslib-Cranfield tests [4]. Columns 1 and 3 of Table 3(a) give the performance for the manually assigned index terms both without and with the use of word normalization through the thesaurus. Columns 2 and 4 do the same for the SMART word stem and thesaurus procedures. It may be seen that in each case, the evaluation measures exhibit the same order of magnitude, the stem process being slightly better for the index terms, and the thesaurus process favoring slightly the automatic abstract process.

The same general conclusions can be derived from the recall-precision comparisons of Table 3(b), which reflect the average performance of 18 queries and 273 documents in biomedicine. Column 1 indicates the performance for the MEDLARS manual indexing process used at the National Library of Medicine, while columns 2 to 4 contain figures for three automatic methods incorporated into the SMART system. It is seen that, for the small sample collection, somewhat better average recall is obtained with the SMART methods, and somewhat better precision through the manual indexing.

The description of the test environment used to produce the data of Table 3 must remain outside the scope of the present study [5,6]. However, if the data given are assumed to be representative of performance differences between presently operating semi-manual retrieval systems and the fully-automatic analysis systems of the future, then a conclusion that the performance level is comparable seems reasonable. Moreover, in both cases, the performance is far below that which an ordinary user submitting a search query to a retrieval system might be hoping to achieve.

In the next few sections, various interactive search techniques are described which may help in raising the search effectiveness closer to the optimal area in the upper right-hand corner of the recall-precision graph.

3. User Feedback Retrieval Methods

A) General Methodology

One of the main hopes in obtaining a retrieval performance which goes beyond that presently reached under normal operating conditions, is to include the customer in the search process. In particular, fewer errors are likely to be made if the information obtained from the users is not restricted to the search request proper, but is supplemented by a variety of special user need indications, or by evaluation data about the acceptability of items previously retrieved by the system in answer to the search requests. User-system interaction is now current for many computer applications, often implemented by special input-output console devices, with the help of operating systems which enable the system to render more or less simultaneous service to a large class of users.

In an information retrieval environment, user interaction may take the form of simple dictionary display routines which can be used to present to the user selected dictionary excerpts as an aid in formulating the original search requests, or in reformulating queries which were originally inadequate [7,8]. Alternatively, more sophisticated methods may be used in which the reformulation of the search requests is automatically performed based on feedback information obtained from the user population [9,10]. The relevance feedback process incorporated into the SMART system is particularly well adapted to a time-sharing computer organization with simple console equipment, since it requires only a minimum of interaction with the user, and places most of the burden on internally stored routines.

Specifically, an initial search is first performed for each request received, and a small amount of output, consisting of some of the highest scoring documents, is presented to the user. Some of the retrieved output is then examined by the user who identifies each document as being either relevant (R) or not relevant (N) to his purpose. These relevance judgments are later returned to the system, and used automatically to adjust the initial search request in such a way that query terms, or concepts, present in the relevant documents are promoted (by increasing their weight), whereas terms occurring in the documents designated as nonrelevant are similarly demoted. This process produces an altered search request which may be expected to exhibit greater similarity with the relevant document subset, and greater dissimilarity with the non-relevant set.

The altered request can next be submitted to the system, and a second search can be performed using the new request formulation. If the system performs as expected, additional relevant material may then be retrieved, or, in any case, the relevant items may produce a greater similarity with the altered request than with the original. The newly retrieved items can again be examined by the user, and new relevance assessments can be used to obtain a second reformulation of the request. This process can be continued over several iterations, until such time as the user is satisfied with the results obtained.

A large number of feedback experiments have been performed with the SMART system in order to identify those methods which appear to be most effective in improving retrieval performance. These experiments are based on a query alteration algorithm described by the following equation:

$$q_{i+1} = \alpha q_i + \beta q_0 + \gamma \sum_{i=1}^{n_1} \underline{r}_i - \delta \sum_{i=1}^{n_2} \underline{s}_i + \sum_{i=1}^{n_3} \underline{w}_i \underline{d}_i + \sum_{i=1}^{n_4} \underline{v}_i \underline{c}_i \quad . \qquad (1)$$

Here, the $(i+1)^{st}$ query statement, q_{i+1}, is defined as a composite obtained from the ith query formulation q_i, the initial q_0, a set of n_1 documents, $\underline{r}_i$, identified by the user as relevant, a set of n_2 documents, $\underline{s}_i$, identified as not relevant, a set of n_3 documents, $\underline{d}_i$, supplied by the user without specific relevance indications, and a set of n_4 important concept numbers, $\underline{c}_i$, also obtained from the user. A number of parameters ($\alpha, \beta, \gamma, \delta, \underline{w}_i$, and $\underline{v}_i$) serve as weighting functions. These parameters are set equal to zero when the corresponding information items are not specified. Alternatively, they may be set equal to 1 or greater if the corresponding items are to be added to (or subtracted from) the present query statement in order to produce the new query.

Equation (1) represents a vector transformation in the document space, which moves the query close to the documents, or concepts, identified as important by the user, and away from the documents, or concepts, specified as unimportant. The effect of various types of transformations is examined in the next few subsections.

B) Positive Feedback

One of the simplest feedback methods is obtained by using previously retrieved relevant items to update the search requests. Fig. 2(a) represents the output obtained for the Cranfield collection by retrieving, in each case, 5 documents at a time, asking the user to identify any relevant items, and adding these relevant items to the search requests to obtain new query formulations. Fig. 2 presents averages over 42 search requests for initial runs, as well as first and second feedback iterations, using equation (1) in the following simplified form:

$$q_{i+1} = q_i + \sum \underline{r}_i \quad . \tag{2}$$

Fig. 2(a) shows that an average improvement of twenty per cent is obtained in the precision between initial run and first feedback iteration. Some further improvement is produced by the second feedback run, particularly in the high recall region [11,12].

The output of Fig. 2(a) is obtained by a process illustrated for query 8 in Table 4(a). Five documents are first retrieved by the initial run, labelled respectively 187, 173, 39, 33, and 139. Document 39 is now identified as relevant (marked 'R' in Table 4(a)), and is used to modify the query. The modified query now produces the five documents labelled 39, 42, 179, 112, and 181. Relevant document 39 had already been retrieved in the initial run, but 42 is new, and after it is identified as relevant, a new query is constructed which, in turn, produces five additional documents, and so on, until the user obtains no further useful information. Obviously, the choice of cut-off, which in the example is set at five retrieved items, is arbitrary, and may in practice be made to depend on the user's wishes.

The example of Table 4(a) illustrates the fact that an improvement in recall and precision may be obtained from one search iteration to the next for two different reasons, known as the ranking effect and the feedback effect, respectively [13]. The former is exemplified by document 39 whose rank is improved from 3 to 1 between the initial run and the first feedback iteration, following its identification as a relevant item. The feedback effect proper is illustrated by relevant document 42, which was not initially retrieved, but jumps from rank 9 to 2 as a result of the feedback action.

One may argue that the ranking effect, which reflects improvements in rank of items already retrieved, should be disregarded, since the user may not care to look at retrieved items more than once. The feedback effect alone may be measured by freezing the ranks of all items as they are retrieved. This is reflected in the output of Table 4(b), where the first 5 ranks are maintained throughout, the first 10 ranks after iteration one, the first 15 after iteration two, and so on. Document 42 is now seen to improve only from rank 9 to 6, since the top 5 ranks are preempted. The corresponding recall-precision graph shown in Fig. 2(b) shows that the feedback effect produces less improvement than the ranking effect.

In a practical retrieval system, a compromise might be made between the output of Table 4(a) and 4(b) by taking documents previously identified as either relevant or not relevant out of the system, but leaving the others alone. For the example of Table 4, this would delete document 39 after the initial run, and documents 39 and 42 following the first iteration.

The output of Fig. 2 and Table 4 clearly shows that user feedback information can help in improving retrieval performance by moving the query close to the area identified as important by the user. Occasionally, however, it happens that some relevant items are lost as other new ones are retrieved. Consider, for example, the relevant document 41 which initially receives rank 11 in the output of Table 4. After the first iteration, it gains from rank 11 to 9, but after document 42 is identified as relevant, item 41 decreases from 9 to 15. That is, as the query approaches document 42, it recedes from 41. This situation is shown graphically in Fig. 3, where query 8 first approaches document 39, and then document 42, while at the same time, it gets away from document 41. In Fig. 3, the size of the correlation coefficient between query and document is represented as varying inversely with the physical distance between them.)

It is clear that in a situation such as the one illustrated in Fig. 3, the search request covers several subject areas. To obtain good retrieval performance, the query must then be split into two parts, one to reflect the subject area of documents 39 and 42, and the other the area of document 41. Experimental query splitting algorithms are presently under construction.

C) Negative Feedback

The feedback process illustrated in the last subsection, operates with a fixed number of retrieved documents, and uses only positive information (that is, relevant documents identified by the user) for feedback purposes. In a practical system, one may anticipate that some users would be willing to look at more information than others before supplying feedback data. This may be the case particularly in situations where nothing useful is retrieved within the top 5, or top 10 items. In such a case, two different procedures are available, known respectively as variable and negative feedback.

In the variable process, the cut-off value, which determines the number of documents to be retrieved for the user before feedback is returned to the system, varies from query to query. One possible strategy might consist in first retrieving five items; if the user is now in a position to supply a relevance judgment, retrieval stops, and the query is updated in preparation for a subsequent search operation; if not, additional documents are retrieved (in decreasing correlation order with the query) until such time as a relevant item is identified by the user.

The recall-precision performance of a typical variable feedback system is shown in Fig. 4(a) for the 200 document Cranfield collection, averaged over 42 search requests. In the output of Fig. 4(a), the cut-off is chosen directly after the first relevant item, with a maximum possible cut-off of 15. It is seen that the variable cut-off process operates much more successfully than the standard system based on a fixed cut-off of 5 items (the second iteration of a standard cut-off process is shown superimposed on the graphs of Fig. 4). Furthermore, out of 42 queries used in the Cranfield tests, only two were found without any relevant items in the top 15 ranks, the average number of items examined being only four.

While the variable process thus shows promise, particularly at the high-precision end of the performance scale, some queries may always be submitted for which relevant items are difficult to identify without an exhaustive examination of a large part of the collection. Under these circumstances, a negative strategy can be used, designed to inform the system about what the user does not wish to retrieve. One of the simplest strategies consists simply in identifying the top document as nonrelevant. The query will then be modified in such a way as to decrease its similarity with the item identified as nonrelevant; at the same time, the assumption is that this modification may increase its similarity with the relevant items in the collection [11].

Consider, as an example, the output for query 1 processed against the collection of 200 Cranfield documents. The standard feedback strategy, which consists in incrementing the concept weights from documents identified as relevant, is illustrated in Table 5(a). A cut-off of 5 documents is again assumed. It is seen from the table that no relevant document is ever identified within the top 15 documents, so that the feedback procedure is unavailing in that case. The relevant documents 21, 22, and 1 remain at their initial ranks 32, 33, and 200 respectively.

On the other hand, when the negative feedback strategy is used in the form

$$q_{i+1} = q_i + \sum \underline{r}_i - s_i \quad , \qquad (3)$$

the altered query retrieves relevant document 22, which, in turn, is used to retrieve the remaining relevant items (21 and 1). The process is illustrated in the output of Table 5(b), and the graphical representation of Fig. 5. The first nonrelevant item (no. 125) is first identified by the user; as the query moves away from document 125, it approaches item 22, whose rank improves from 33 to 6 between initial runs and first iteration. When item 22 is identified as relevant after the first iteration, the other relevant items 1 and 21 increase their positions from ranks 199 and 200 to 11 and 12 (the top ten ranks are frozen by this time in the output of Table 5(b)).

The negative feedback strategy profits from the fact that in a reasonably homogeneous document space, the various alternative subject areas should be reachable in a small number of negative feedback steps [14]. The case of query 1 previously discussed renders plausible the large average improvement shown for the 42 Cranfield queries in Fig. 4(b) between the negative feedback strategy of equation (3) and the positive strategy of equation (2). Fig. 4(b) shows the first iteration curve using negative feedback, compared with the second iteration positive strategy. The improvements in retrieval effectiveness exhibited in Fig. 4 for the variable cut-off and negative feedback methods make it appear that the two strategies might be combined in an actual operational situation.

The feedback procedures described in the present section are likely substantially to improve retrieval performance, over the presently achievable performance levels. In some cases, the improvement may be as high as 30 percent in the high precision area and 20 percent in the high recall area. The recall-precision graphs included in this study demonstrate, however, that the distance to the theoretically desirable ideal system is still large. A perfect system, exhibiting both high recall and high precision may well not be reachable in the near future.

This is particularly true in view of the fact that, in practice, time limitations make it impossible to compare the complete document file with each search request, since system responses must be furnished to the user population in real time. In the SMART system, the documents in a collection are grouped into clusters, and only those document clusters are searched which are situated close to the search request [12,15]. Such a partial search system, illustrated in Fig. 6, brings with it a built-in ceiling in the attainable recall — since some relevant items may not be included in the document groups which are actually searched. Such partial search systems, operating in conjunction with user feedback strategies remain to be studied experimentally, before being implemented in practice. In the meantime, the search and retrieval methods included in the present study may serve as a preview of the potential of the automatic retrieval systems of the future.

References

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

[2] M. E. Lesk, Operating Instructions for the SMART Text Processing and Document Retrieval System, Scientific Report No. ISR-11 to the National Science Foundation, Section II, Department of Computer Science, Cornell University, June 1966.

[3] G. Salton et al., Scientific Reports on the SMART System, Nos. ISR-11, ISR-12, ISR-13, Department of Computer Science, Cornell University, June 1966, June 1967, and December 1967.

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

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

[6] G. Salton and D. Williamson, Notes on the SMART-MEDLARS Study, to be included in Scientific Report No. ISR-14 to the National Science Foundation, Department of Computer Science, Cornell University, 1968.

[7] R. M. Curtice and V. Rosenberg, Optimizing Retrieval Results with Man-machine Interaction, Center for the Information Sciences Report, Lehigh University, Bethlehem, Pa., 1965.

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

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

[10] J. J. Rocchio and G. Salton, Information Search Optimization and Iterative Retrieval Techniques, Proceedings of the AFIPS Fall Joint Computer Conference, Vol. 27, Spartan Books, November 1965.

References
(contd)

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

[12] G. Salton, Search Strategy and the Optimization of Retrieval Effectiveness, Proceedings of the FID/IFIP Conference on Mechanized Information Storage, Retrieval, and Dissemination, Rome, June 1967; also in Scientific Report No. ISR-12 to the National Science Foundation, Section V, Department of Computer Science, Cornell University, June 1967.

[13] H. A. Hall and N. H. Weiderman, The Evaluation Problem in Relevance Feedback, Scientific Report No. ISR-12. to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, June 1967.

[14] J. Kelly, Negative Response Relevance Feedback, Information Storage and Retrieval, Scientific Report No. ISR-12 to the National Science Foundation, Section IX, Department of Computer Science, Cornell University, June 1967.

[15] R. T. Grauer and M. Messier, An Evaluation of Rocchio's Clustering Algorithm, Scientific Report No. ISR-12 to the National Science Foundation, Section VI, Department of Computer Science, Cornell University, June 1967.

Type of Recall Failure (resulting in relevant document not retrieved)	Number of Instances	Percentage of Total
A. Indexing Errors		
1. Important term missing from document specification (indexer omission)	6	11%
2. Indexing insufficiently exhaustive (some aspect of document content not reflected in the indexing)	6	11%
3. Indexing insufficiently specific (index terms used are too broad)	9	17%
B. Search Errors		
4. Important term missing from query specification (searcher omission)	8	15%
5. Lack of sufficient user-system interaction (searcher misunderstands user need)	21	40%
6. Missing items because of selective print-out (retrieved item not shown to user due to sampling)	3	6%

Typical Recall Failures in Conventional Retrieval System

(18 queries processed by Medlars retrieval system)

Table 1

Type of Precision Failure (resulting in non-relevant document retrieved)	Number of Instances	Percentage of Total
A. Indexing Errors		
1. Indexing too exhaustive (minor aspect of document content reflected in indexing)	17	11%
2. Indexing not sufficiently specific (indexing language used is too general	53	34%
3. Important index term missing or inappropriately or incorrectly used	7	5%
4. Incorrect term relationship specified by indexing rules	8	5%
B. Search Errors		
5. Search formulation insufficiently specific	18	11%
6. Lack of sufficient user-system interaction	24	15%
7. Search formulation insufficiently exhaustive (not enough Boolean formulation)	22	14%
8. Miscellaneous errors (on the part of searcher or requestor	7	5%

Typical Precision Failures in Conventional Retrieval System

(18 queries processed by Medlars)

Table 2

	Stem Indexing	SMART Stem Abstracts	Thes-3 Indexing	SMART Thes-3 Abstracts
Normalized Recall	0.890	0.864	0.873	0.884
Normalized Precision	0.683	0.670	0.694	0.695

a) Comparison of Cranfield Index Term Match with SMART Abstract Processing (Cranfield 200 documents; 42 queries)

	Medlars Indexing	SMART Word Form	SMART Word Form	SMART Thesaurus
Recall	0.643	0.704	0.718	0.690
Adjusted Precision	0.625	0.571	0.570	0.611

b) Comparison of Medlars Index Term Match with SMART Abstract Processing (Medlars 273 documents; 18 queries, macro-average)

Comparison of Index Term Matching with Automatic Abstract Processing

Table 3

INITIAL			FEEDBACK ITERATIONS					
			1		2		3	
RANK	DOC	CORR	DOC	CORR	DOC	CORR	DOC	CORR
1	187	.2315	39R	.9641	39R	.9143	39R	.9143
2	173	.1949	42R	.6533	42R	.8871	42R	.8871
3	39R	.1949	179	.5200	179	.5804	179	.5804
4	33	.1949	112	.5095	112	.5609	112	.5609
5	139	.1771	181	.4429	181	.4600	181	.4600
6	185	.1743	188	.3865	188	.4090	188	.4090
7	36	.1714	97	.3559	97	.3757	97	.3757
8	188	.1702	45	.3529	45	.3735	45	.3735
9	42R	.1666	41R	.3022	2	.3296	2	.3296
10	199	.1621	2	.2925	173	.2703	173	.2703
11	41R	.1542	173	.2606	62	.2661	62	.2661
12	15	.1291	62	.2494	117	.2615	117	.2615
13	98	.1238	187	.2474	116	.2612	116	.2612
14	178	.1231	101	.2371	187	.2510	187	.2510
15	23	.1208	185	.2355	41R	.2430	41R	.2430

a) Feedback Results for Query 8 (all effects)

INITIAL			FEEDBACK ITERATIONS					
			1		2		3	
RANK	DOC	CORR	DOC	CORR	DOC	CORR	DOC	CORR
1	187	.2315	187	.2474	187	.2510	187	.2510
2	173	.1949	173	.2606	173	.2703	173	.2703
3	39R	.1949	39R	.9641	39R	.9143	39R	.9143
4	33	.1949	33	.0521	33	.0317	33	.0317
5	139	.1771	139	.0945	139	.0865	139	.0865
6	185	.1743	42R	.6533	42R	.8871	42R	.8871
7	36	.1714	179	.5200	179	.5804	179	.5804
8	188	.1702	112	.5095	112	.5609	112	.5609
9	42R	.1666	181	.4429	181	.4600	181	.4600
10	199	.1621	188	.3865	188	.4090	188	.4090
11	41R	.1542	97	.3559	97	.3757	97	.3757
12	15	.1291	45	.3529	45	.3735	45	.3735
13	98	.1238	41R	.3022	2	.3296	2	.3296
14	178	.1231	2	.2925	62	.2661	62	.2661
15	23	.1208	62	.2494	117	.2615	117	.2615
17	0	0	0	0	41R	.2430	41R	.2430
134	0	0	40R	.0362	0	0	0	0
143	0	0	0	0	40R	.0368	40R	.0368
200	40R	.0000	0	0	0	0	0	0

b) Feedback Results for Query 8 (feedback effect only)

"Increment Only" Strategies for Query 8
(Cranfield thesaurus; cosine numeric N=5)

Table 4

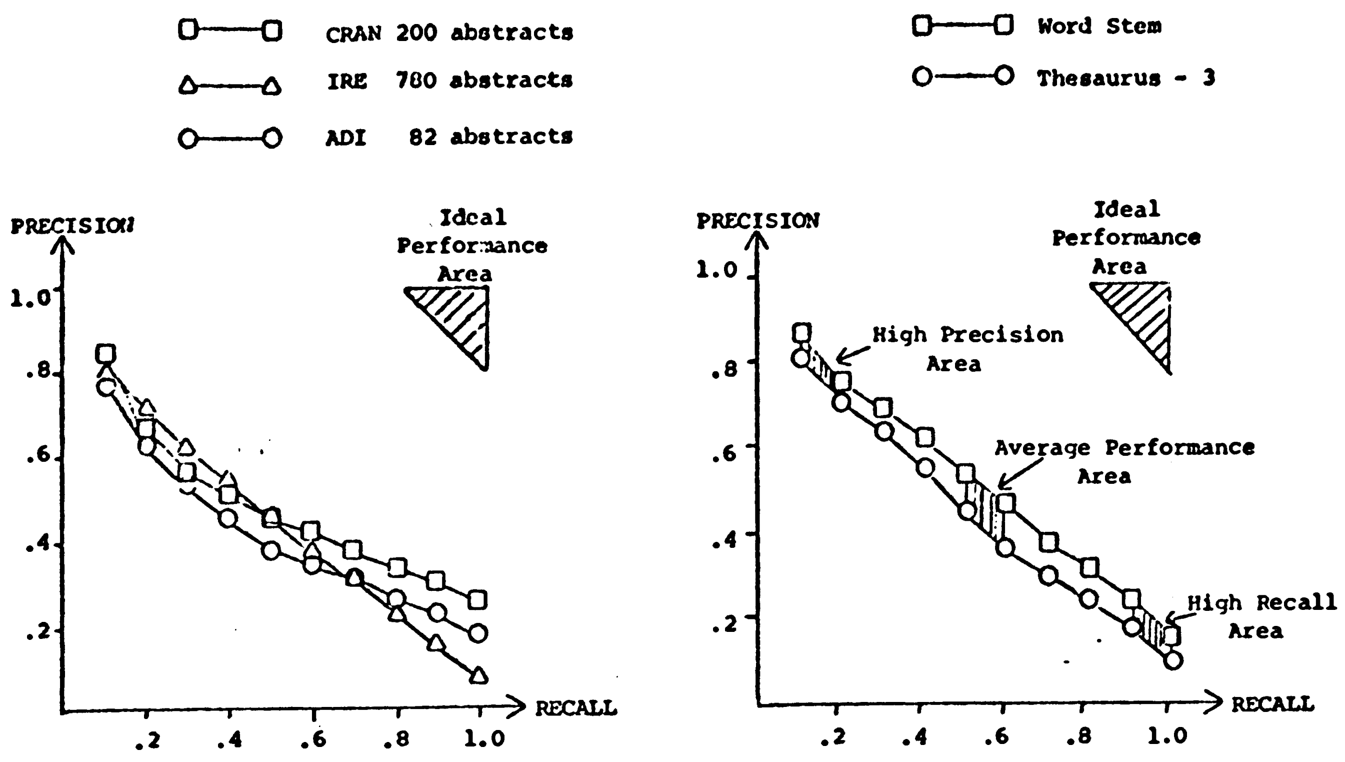
RANK	DOCUMENTS 0	1	2	3
1	125	125	125	125
2	12	12	12	12
3	159	159	159	159
4	64	64	64	64
5	66	66	66	66
6	123	123	123	123
7	67	67	67	67
8	65	65	65	65
9	24	24	24	24
10	29	29	29	29
11	190	190	190	190
12	25	25	25	25
13	127	127	127	127
14	136	136	136	136
15	85	85	85	85
32	21R	21R	21R	21R
33	22R	22R	22R	22R
200	1R	1R	1R	1R

a) "Increment Only" Strategy for Query 1
(Initial Run (0) and three Feedback Runs (1-3))

RANK	DOCUMENTS 0	1	2	3
1	125	125	125	125
2	12	12	12	12
3	159	159	159	159
4	64	64	64	64
5	66	66	66	66
6	123	22R	22R	22R
7	67	161	161	161
8	65	155	155	155
9	24	184	184	184
10	29	189	189	189
11	190	70	21R	21R
12	25	4	1R	1R
13	127	136	141	141
14	136	37	59	59
15	85	31	30	30
32	21R	0	0	0
33	22R	0	0	0
199	0	1R	0	0
200	1R	21R	0	0

b) "Decrement High" Strategy for Query 1
(feedback only evaluation)

"Decrement High" Strategy for Query 1
Table 5



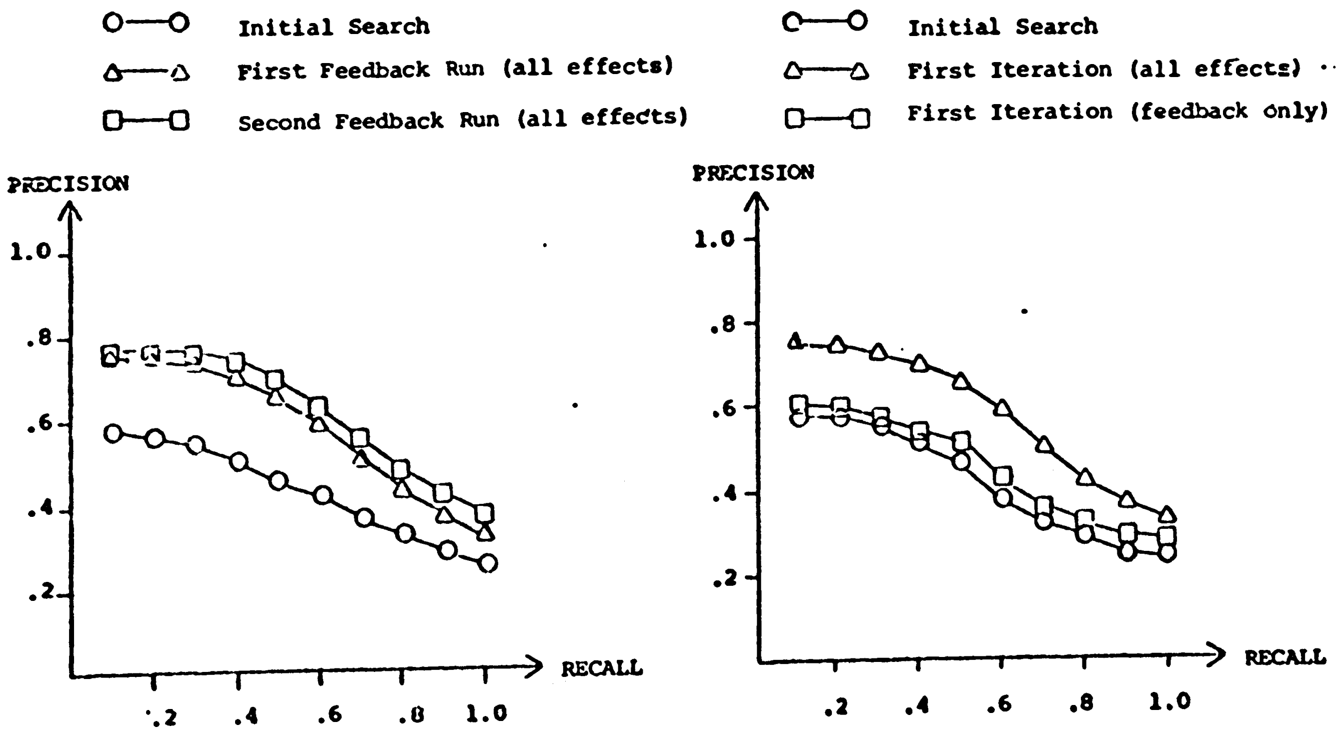
a) Word Stem Comparison for Three Collections (Cosine Numeric Correlation)

b) Word Stem Thesaurus Comparison (Averages over 780 IRE Documents) (Cosine Numeric Correlation)

Typical Performance Characteristics for Several Document Collections and Processing Methods

(cosine numeric correlation)

Fig. 1



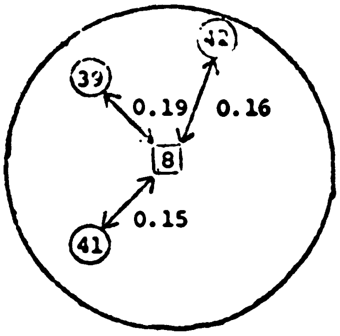
a) Several Feedback Iterations (all effects)

b) Comparison of Ranking and Feedback Effect

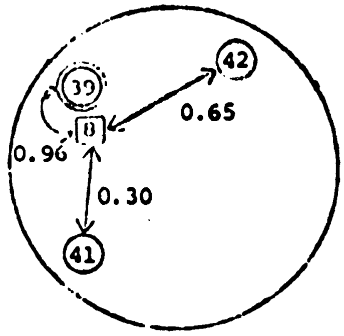
Feedback Performance — "Increment Only" Strategy $q_{i+1} = q_i + \sum r_i$

(Cranfield thesaurus run; cosine numeric correlation)

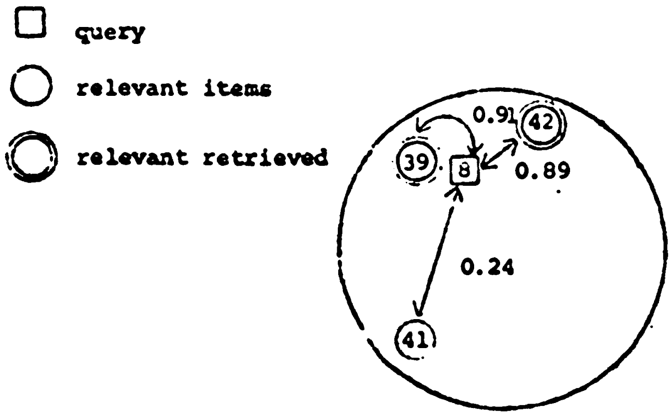
Fig. 2



a) Initial Search



b) First Feedback
(after retrieval of 39)

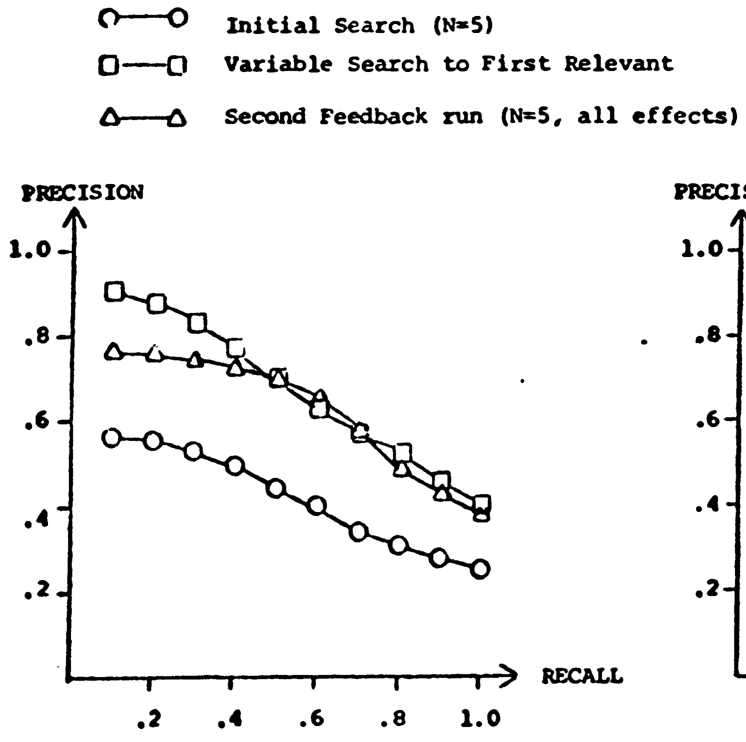


c) Second Feedback
(after retrieval of 39 and 42)

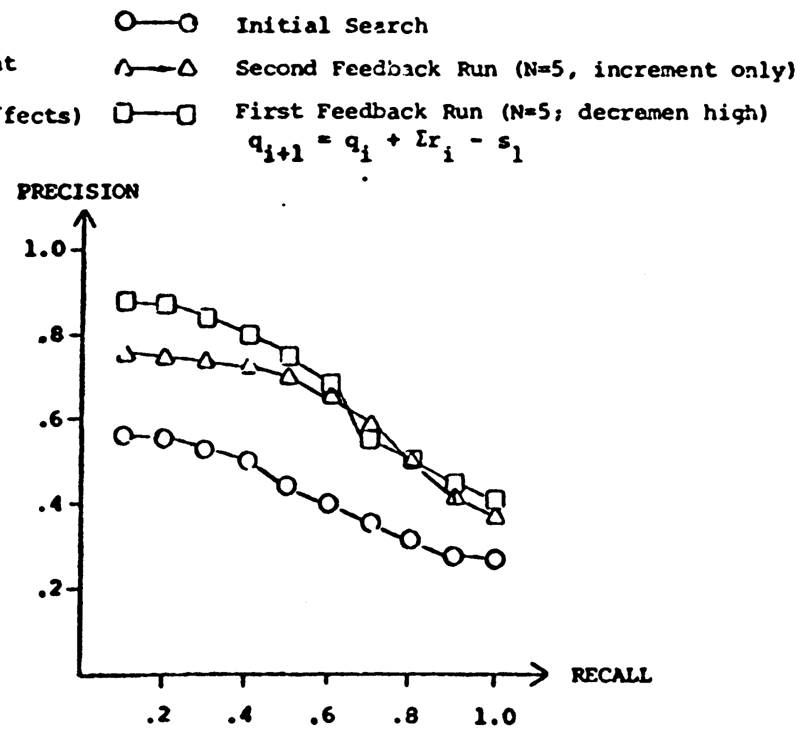
Feedback Illustration for Query 8

(Query Splitting)

Fig. 3



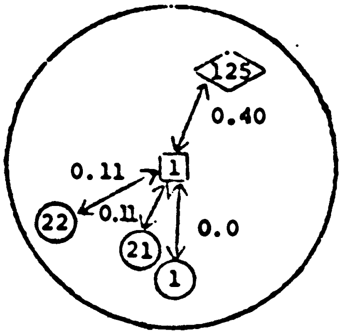
a) Variable Feedback (to first relevant in top 15; average search of 4)



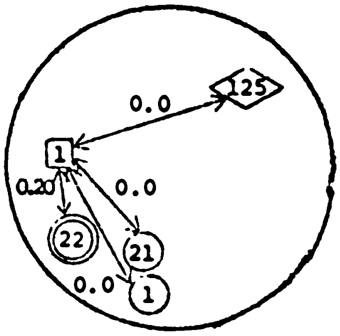
b) Comparison of Positive with Negative Feedback (N=5, all feedback effects)

Variable Cut-off and Negative Feedback
(Cranfield thesaurus, cosine numeric correlation)

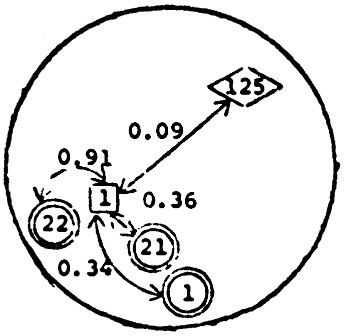
Fig. 4



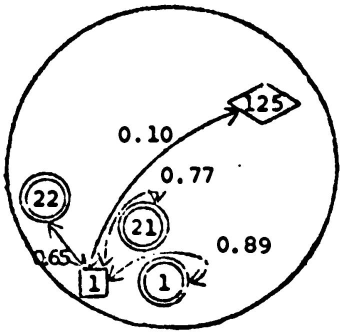
a) Initial Search
(no relevant retrieved)



b) First Negative Feedback
(document 125)



c) Second Positive Feedback
(document 22)



d) Third Positive Feedback
(documents 1, 21)

query

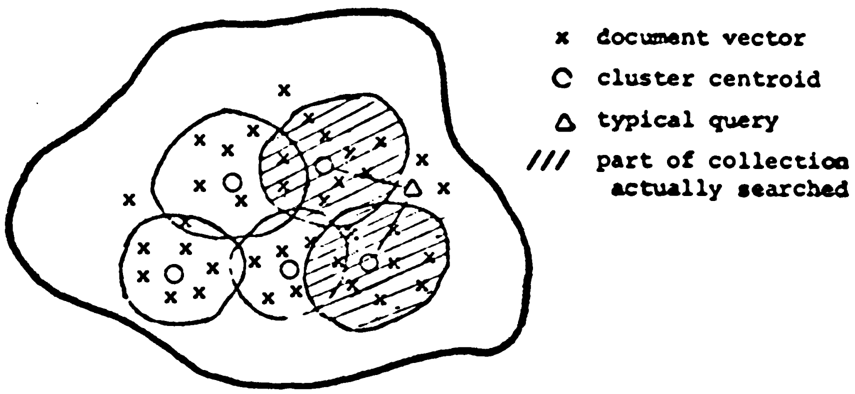
relevant items

relevant retrieved

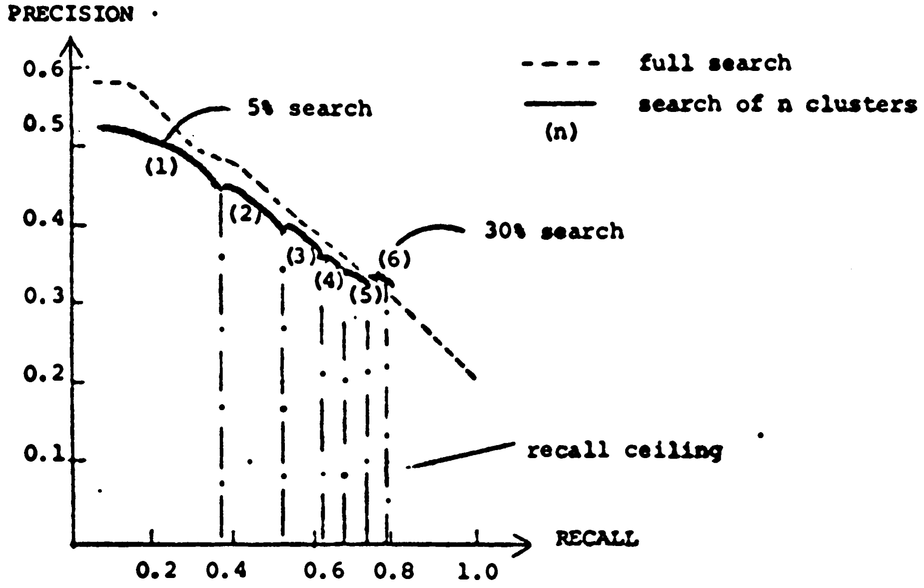
non-relevant item

Negative Feedback Illustration, Query 1

Fig. 5



a) Sample Clustered Document Space



b) Cluster Search Evaluation

(Cranfield 200 documents, 42 queries, abstract stem process, 23 clusters)

Fig. 6