XII. Query Splitting Using Relevant Documents Instead of Queries in Relevance Feedback

T. Leventhal and R. Miller

Abstract

Iterative search procedures permit the user of an information retrieval system to "update" a query following the display of some type of information. Relevance feedback attempts to phrase a theoretically best way of discriminating between relevant and nonrelevant items. [1] The use of relevance feedback has many applications in information retrieval. [2] This paper attempts to explore the use of query splitting as a method of relevance feedback. The SMART retrieval system is used. All experiments are performed on the Cranfield 200 thesaurus collection. Evaluation of the results of these experiments are based on recall and precision tables. Suggestions are made at the end of the paper for further investigation of the uses of query splitting in information retrieval systems.

1. Introduction

Many experiments have been done with user interaction in automated information retrieval systems. The results of these experiments have proven that user interaction increases retrieval performance. [3] Previous work done by Borodin, Kerr, and Lewis [4] concerned itself with the problem of query splitting. The query splitting algorithm used at the time involved a relevance judgment of the five documents ranking highest with the original query. Document-document correlations of the relevant documents retrieved

were used to group these documents. If the correlations exceeded a certain constant, the documents were put into the same group; otherwise, they were separated. A new query was formed for each group using the formula:

$$Q^{j}_{i+1} = Q_i + \sum_k r_k - \sum_k n_k$$

where Q^{j}_{i+1} is the (i+1)th query for group j, Q_i is the original query, r_k is a relevant document in group j, and n_k are the two highest ranking nonrelevant items retrieved by Q_i. Certain modifications of this formula were made for cases where no relevant documents were retrieved. This process could be repeated for each subsequent query.

Work done by Crawford and Melzer [2] also led to the problem dealt with in this paper. A modification was used of the general query update formula

$$Q_{i+1} = \alpha Q_i + \beta Q_o + \gamma \sum_{i=1}^{N_1} R_i + \delta \sum_{i=1}^{N_2} N_i$$

where Q_{i+1} is the new query being formed, Q_i is a query formed prior to Q_{i+1}, R_i are relevant documents, and N_i are high ranking nonrelevant documents. Here, all the coefficients were set equal to zero, except γ. Thus, only relevant documents were used as a new query.

As will be seen in the next section, the problem undertaken by this paper is actually a combination of the work done by these two groups of authors. Query splitting will be done using only relevant documents in the new queries.

2. Motivation and Assumptions

Regular query splitting (that done by Borodin, Kerr, and Lewis) and the use of relevant documents in feedback both, in general, increase retrieval performance as measured by recall and precision. [2,4] By combining these, it is hoped that this increase in performance remains.

It is assumed that the person using the system would be able to tell whether the documents retrieved by his query are relevant or not. Sometimes, a query retrieves documents which fall into separate groups. There may be other documents which are also relevant and which would fall into these groups, had they been retrieved. The retrieval of these additional relevant documents is the purpose of this project. An example is shown in Fig. 1.

It is important that this method of query splitting should only be used where it would contribute to retrieval performance. Many conditions must hold to make this feedback process at all practical. First, the person requesting a search must be able to make very good relevance judgments. He must also be able to tell if the relevant documents retrieved by his first query would serve as better queries than his original. Query splitting should only be used if the documents retrieved by the original query fall into distinct groups. If this is not the case, splitting would result in an overlapping of the searches. A classical example of this is the query about aerodynamics of birds. Here, a decision as to the use of query splitting depends upon the nature of the document collection. If it contains books specifically about aerodynamics of birds, then query splitting should not be used. Splitting should be used when the person making the request would have to settle for books about birds or books about aerodynamics.

Another major assumption involved is that the original query does not do very well in retrieving relevant documents. If the original query

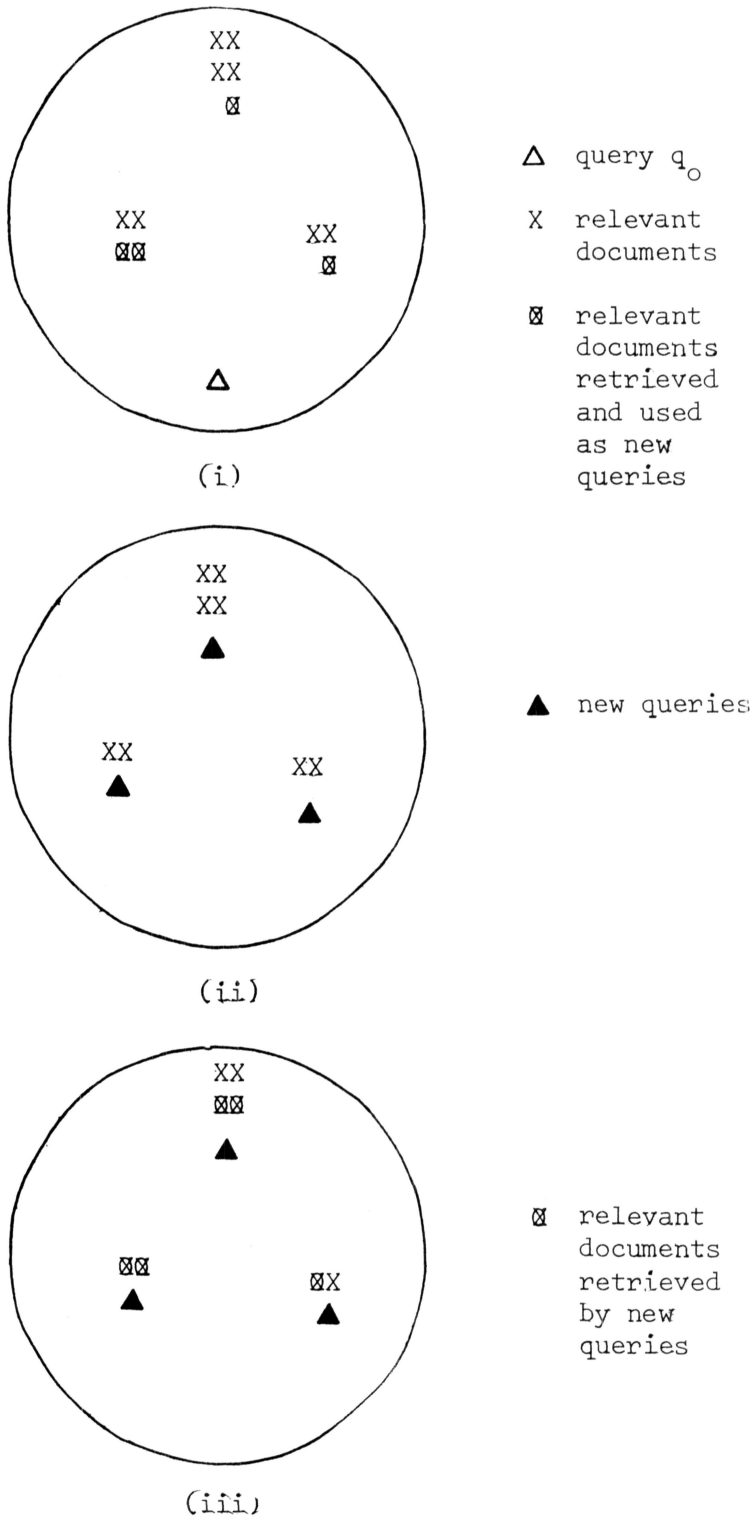


Illustration of Query Splitting Using Only Relevant Documents as Feedback

Fig. 1

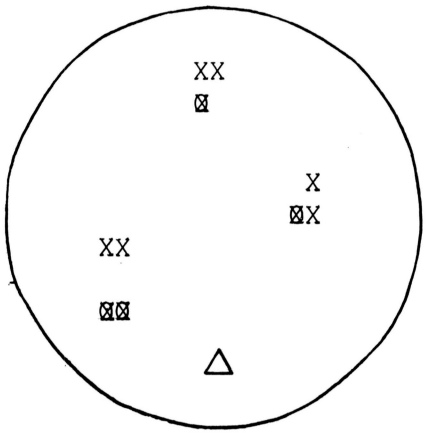
does do a good job, then any type of feedback could only result in a little improvement at best. If the original query retrieves no members of a group of relevant documents, then it is very unlikely that query splitting will help retrieve members of that group.

What this project proposes is actually a stronger splitting of the query than was done by Borodin, Kerr, and Lewis. Whereas their method moved the split queries closer to, but not inside, groups of relevant documents, this project puts the split queries right inside the groups such as shown in Fig. 2. It is hoped that this method increases the retrieval performance of the system.

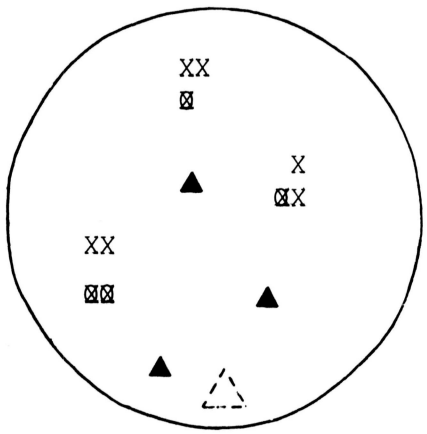
3. Implementation

The experiments are performed on the Cranfield 200 collection which contains 200 documents and 42 queries. The thesaurus form of the document and query vectors is used. First a full search is done on the collection using all 42 queries. This search is carried out under the SMART retrieval system. At least the top twenty ranking documents are displayed for each query. The ranks of these documents are determined by the magnitude of the cosine correlation with each query. Relevance judgments are made for the documents retrieved. (This has been done for the Cranfield 200 collection).

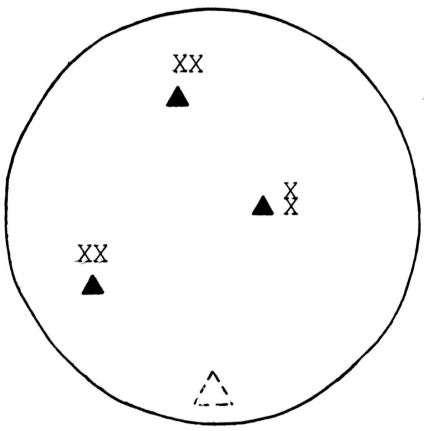
If a query retrieves two or more relevant documents, the cosine correlation between these documents is determined. This is done using a control subroutine called by MASTER. The subroutine reads in the query number and the relevant documents retrieved for that query. It then locates the document vectors, uses INNER to find their cosine correlations, and prints them. If the original query retrieves no relevant documents or only one in the top twenty, other methods of feedback are used to increase



(i) original query and relevant documents retrieved by it



(ii) location of new split queries by method of Borodin, Kerr, and Lewis



(iii) location of new split queries by proposed method

Comparison of Query Splitting Methods

Fig. 2

retrieval performance.

Once the correlations between relevant documents have been determined, the documents are split into groups for each query. These groups will be used as the new queries. The splitting is done by comparing each relevant document against the other relevant documents retrieved by the same query and seeing if their cosine correlations are above 0.5. If it is, the documents are put into the same groups. If not, they are put into separate groups. A document may be in more than one group. For example, if the correlation matrix for documents 34, 35, and 36 was the following:

	34	35	36
34	1	0.60	0.38
35	0.60	1	0.58
36	0.38	0.58	1

the groups formed would be (34,35) and (35,36). This grouping could be done by programming but since the size of the collection used is small, they are done by hand in the present case. Programming could save time for grouping documents of larger document and query collections.

After the groups are formed, new queries are generated using CRDCEN. If only one document is a member of a group, then the new query is the document itself. The relevant documents for this new query are the same as the relevant documents for the old query from which it was retrieved. If two or more documents are members of a group, then a new query is formed by adding the weights of the concepts of each document. Again, the relevant documents for this new query are the same as the relevant documents for the old query.

SMART is again used to do a full search of the collection with the

new queries. Final rankings are obtained and average recall-precision graphs are done by the computer.

The results obtained by the new split queries and an evaluation of these results are contained in the next section.

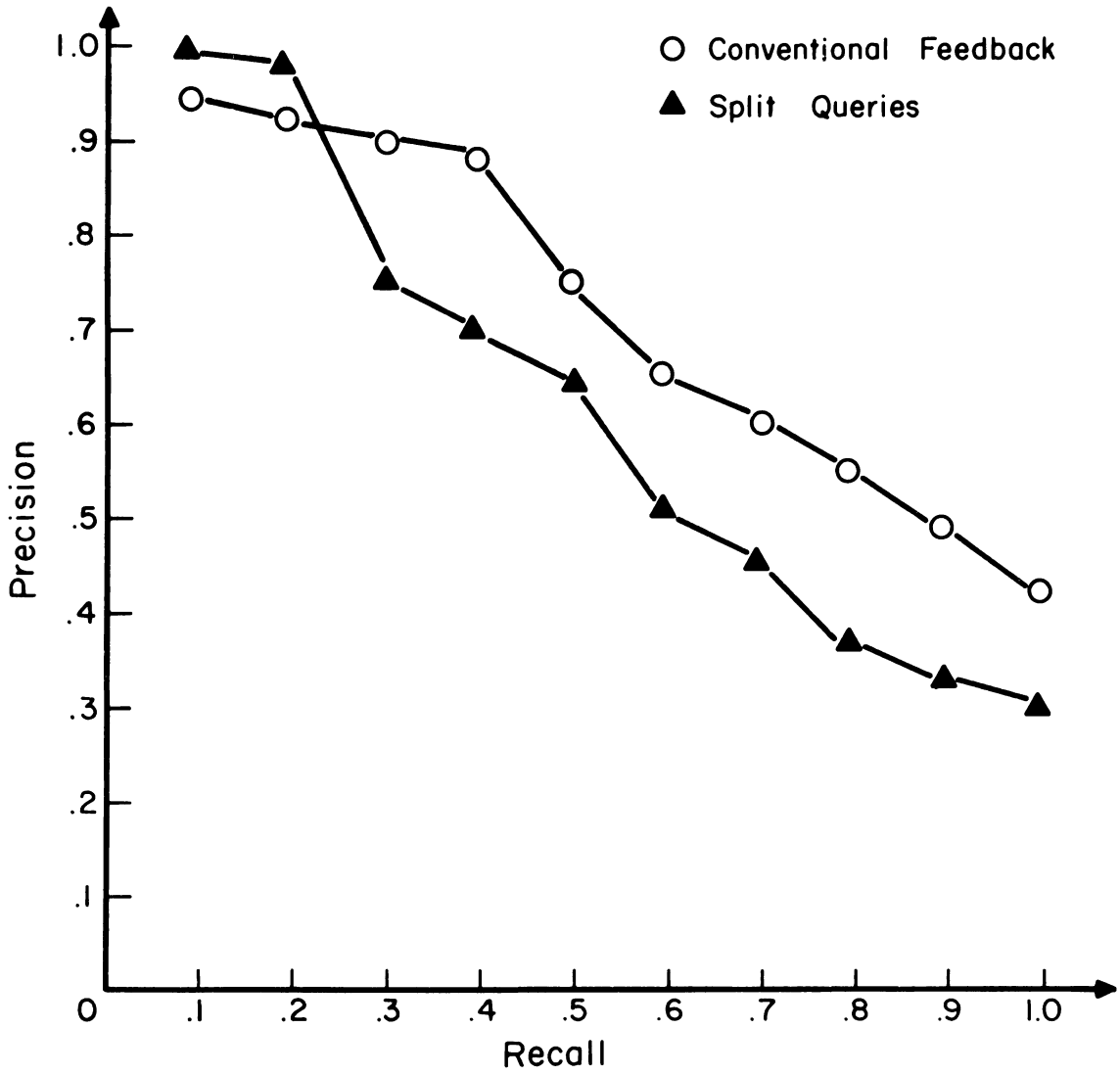
4. Evaluation and Results

Evaluation is provided by comparing the results obtained by the first iteration of a conventional relevance feedback run against the results of the split query run. The conventional run uses the sum of the two highest ranking relevant documents retrieved plus the initial query as a new query.

Only twenty-four of the forty-two queries in the Cranfield collection produce groups of relevant documents for query splitting. In some instances all of the relevant documents associated with a query are used to split the query. For this reason only sixteen of the queries provide any meaningful basis for evaluation.

Fig. 3 is a comparison of the recall-level averages of twenty-four queries for the split query search and the conventional feedback search. The higher precision at low recall for the split queries is caused by the use of relevant documents as new queries. These relevant documents are always retrieved first by the split queries. This does not help the user because he is already aware of these documents. At higher recall, the conventional feedback run shows higher precision than the split query run.

A better method of evaluation is to use the residual document space for recall-precision graphs. The residual document space contains only those documents which were not shown to the user for relevance judgments after the initial search. This type of evaluation is done for individual



Comparison of Conventional Feedback with Split Queries

Fig. 3

queries in the second part of this section.

For each of the sixteen queries which are split, there are two or more lists of ranked documents that are retrieved. A single ranking of documents is obtained for each original query by merging the lists adcording to document-query correlations. If a document appears in more than one list, it receives a rank according to its higher correlation.

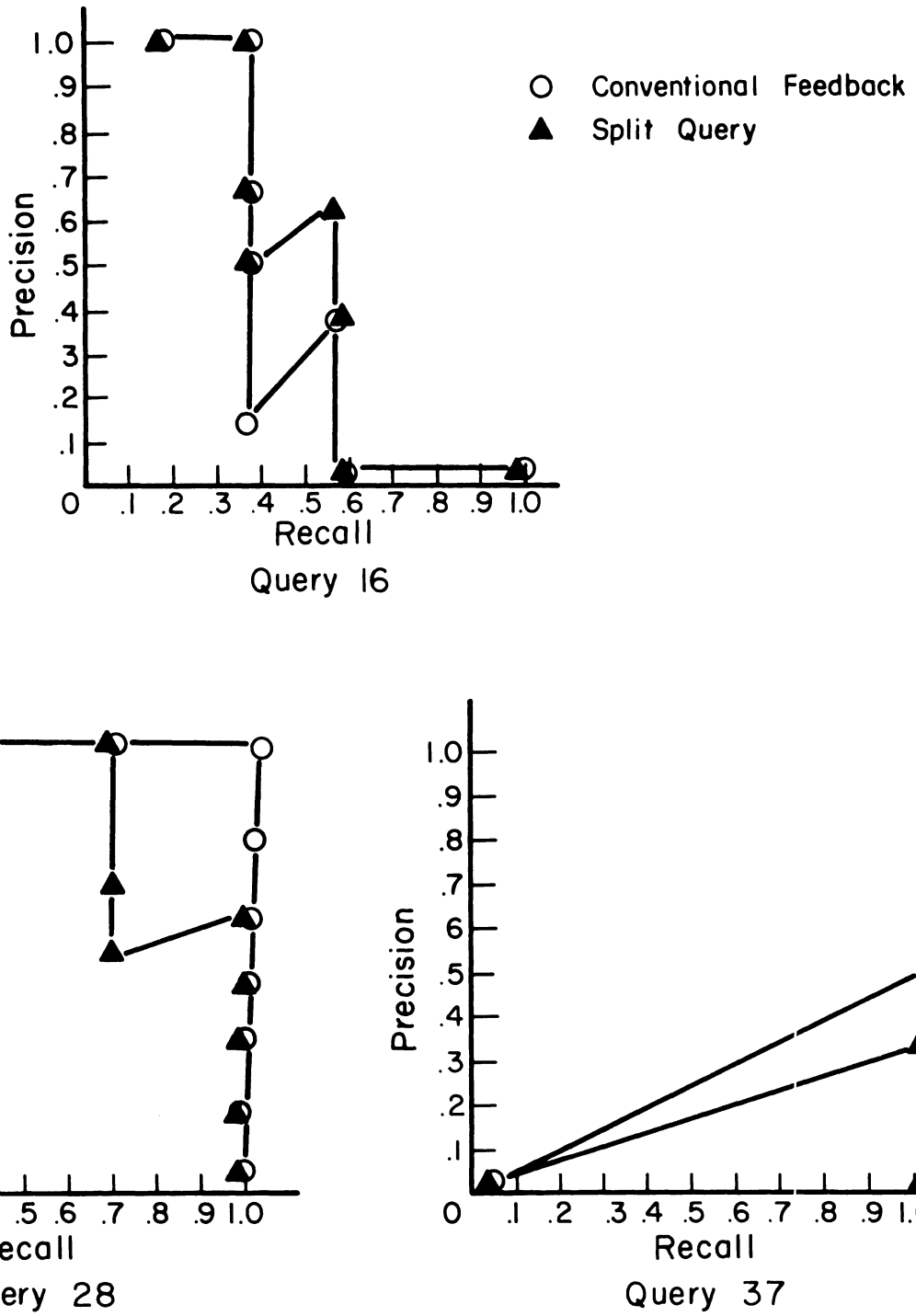
In some cases a query can be split in more than one way depending upon the number of documents initially shown to the user. This variation is tested in two ways for the merged lists of documents. The increase in relevant documents obtained by the split query run over the conventional feedback run is noted as a function of the number of documents given to the user after the initial search. This is done for the case where no documents are dropped from the document space. The same thing is done for the residual document collection. These results are tabulated in Fig. 4, consisting of three tables resulting from ten, fifteen, and twenty documents being initially shown to the user. The numbers at the top of each column tell whether the increase is measured for the highest ranking documents among the top five, ten, fifteen, etc. The tables show that query splitting seems to give good results when the entire document space is used for evaluation. When the documents which have been previously seen are dropped from the document space (which is a better criterion for evaluation), very little improvement is seen. In most cases, query splitting does no better than the conventional feedback technique. In some cases, it even does worse.

Fig. 5 contains a sampling of recall-precision graphs. These graphs are based on residual collection evaluation for the merged rankings of the split queries. In some cases, one method is better than the other, but most

Query Number	Increase in relevant (no documents dropped from document space)					Increase in relevant (previously seen documents dropped)		
	5	10	15	20	25	5	10	20
7	+1	--	--	--	--	--	--	--
13	--	--	--	+1	+1	--	--	--
15	+1	+2	+1	+1	--	+1	--	--
16	-2	--	--	-1	--	--	--	--
18	+1	--	-1	--	--	-1	--	--
24	+1	+1	--	--	--	+1	--	--
28	--	--	--	--	--	--	--	--
31	-1	--	--	--	--	--	-1	--
39	+1	+1	--	--	--	--	--	--
Totals	+2	+4	--	+1	+1	+1	-1	--
User Initially Sees Ten Documents								
2	--	--	--	--	--	--	--	--
5	+1	+1	--	-1	-2	-1	-1	--
7	+1	--	--	--	--	--	--	--
8	+1	+1	--	--	--	--	--	--
11	--	--	--	--	--	--	--	--
34	--	+1	+1	--	+1	+1	--	--
37	--	--	--	--	--	--	--	--
Totals	+3	+3	+1	-1	-1	--	-1	--
User Initially Sees Fifteen Documents								
4	+2	+3	+3	+3	+3	+1	+1	--
11	+1	+1	+1	+1	+1	--	--	--
13	+1	+1	+1	+1	+2	--	--	--
16	--	+2	+2	+1	+1	--	+1	--
Totals	+4	+7	+7	+6	+7	+1	+2	--
User Initially Sees Twenty Documents								

Increase in Number of Relevant Documents Retrieved

Fig. 4



Sample Recall—Precision Graphs
on Residual Document Space

Fig. 5

of the time the differences are so slight that it really does not make any difference.

5. Conclusions

Using just the documents as split queries instead of using them to modify the split queries does not cause an appreciable change in the results of the project done by Borodin, Kerr, and Lewis. From this it can be concluded that when the documents are added to the query to modify it, they swamp the original query. The original query takes on a minor role.

When using a small homogeneous collection such as the Cranfield 200, the results obtained by query splitting are not significantly different from those of the conventional feedback search. One reason for this is that the groups of documents existing in the document space may be very small. Sometimes only one document belongs to a group. If this is the case, then there can be no additional relevant documents retrieved in that group. Another reason, as stated before, is that sometimes no member of a group of relevant documents is retrieved by the initial query.

One suggestion for further research in query splitting is to use different document and query collections. The Cranfield 200 is a very small homogeneous collection of documents. Query splitting may be more practical when used on a larger, more varied collection, a situation which is possibly more common. Research can also be done using high ranking nonrelevant documents as negative feedback for split queries. A very important area of study is determining some sort of query splitting use algorithm. Knowing when to use query splitting and when not to use it is extremely important in getting good results.

References

1 G. Salton, Automatic Information Retrieval, Class Notes, Computer Science 435, Cornell University, 1969.

2 R. G. Crawford and H. Z. Melzer, The Use of Relevant Documents Instead of Queries in Relevance Feedback, Report No. ISR-14 to the National Science Foundation, Section IX, Department of Computer Science, Cornell University, 1968.

3 E. Ide, User Interaction with an Automated Retrieval System, Report No. ISR-12 to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, 1967.

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

5 R. Dietz, T. Horowitz, and W. Riddle, Relevance Feedback in an Information Retrieval System, Report No. ISR-11 to the National Science Foundation, Section IV, Department of Computer Science, Cornell University, 1966.

6 G. Salton, Automatic Information Organization and Retrieval, McGraw Hill, Inc., New York 1968.