The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System

A. F. Smeaton and C. J. van Rijsbergen*
Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

This paper presents the results of an experimental investigation into the effects that some forms of query expansion by term addition or term deletion, have on the retrieval effectiveness of a document retrieval system. The overall search strategy used by a user is an iterative process whereby a set of user-judged relevant documents at any point in the search is used to refine and improve on the remainder of the user's search. At some point during the search, the set of relevant documents found so far can be used to modify the original query, either by the addition or deletion of search terms. This process is called 'query modification' or 'query expansion'. A number of different types of query modification strategies are tried and the results obtained are presented and analysed.

1. INTRODUCTION

A document retrieval system (DRS) can be defined as an information system which stores and manipulates a database of document references, and retrieves from this database in response to a user's query, a set or list of document references. Documents include such things as technical papers, newspaper articles, book reviews, office memos, etc. Each document is usually represented by a set of index terms whose meaning roughly reflects the content of the document.

Users of a DRS may not always know exactly what their information need is, or be able to express it exactly, consequently their query generally is not an exact representation of their information need. Therefore if a document retrieval system is to be effective, some way must be found to discover the information need of a user. There is a fair consensus now that this can only be done through some trial and error process. Traditionally an intermediary person has been used to formulate a user's search, thereby allowing the information need to emerge through the interaction between the user and the intermediary.

In most commercial and experimental on-line IR systems a user is allowed to 'browse' through a thesaurus or some other term–term similarity structure, and can then modify his or her search by adding to or deleting search terms from the original query. This process is called *query modification* (or query expansion), and it may involve the *removal* or the *deletion* of search terms.

In recent years much work has been done in deriving retrieval strategies based on certain mathematical models.[1,2] All these strategies have in common the fact that they use relevance feedback (see below) to estimate the parameters of the weighting function used to implement the search. What they generally do *not* do is select the terms that are to be used in the search, the search terms are assumed given as fixed once and for all by the user. What we propose to do in this paper is to present the results of an experimental investigation into the effects that some forms of query modification have on the overall retrieval effectiveness of a document retrieval system based on one of the mathematical models referred to above. The query modification strategies we propose are guided by theoretical considerations, although our final choice is based on intuition.

Before we present our experimental results and their analysis, we will outline the basic retrieval strategy used in the experiments.

2. RELEVANCE FEEDBACK

The overall retrieval strategy from the user's point of view has an underlying assumption of term independence. In previous work[3] one of the authors has experimented with term dependence but found it extremely difficult to derive effective search strategies. Our approach here is to use term dependence for the selection of further search terms, but to use term independence in constructing the form of the weighting function. This approach was originally suggested in Ref. 4. The practical realization of the strategy is best described diagramatically as in Fig. 1.

From the user's search statement are derived a set of query terms which are then used in a simple retrieval strategy such as Inverse Document Frequency (IDF) weighting, to provide a sample of possibly relevant documents called the *initial sample*. Inverse Document Frequency weighting is a retrieval mechanism whereby search terms are assigned weights inversely proportional to their frequency of occurrence throughout the document collection, and documents are ranked according to the sum of the weights of the search terms that index them. The size of the initial sample is not important for our experiments, although it must not be too small, so we shall use a fixed size. The user then examines the documents in this sample and judges which ones are relevant. If no documents from the initial sample are found to be relevant, the user should reformulate his query. If, on the other hand, the user is satisfied that he has found all the documents relevant to his information need then he can terminate his search. Finally, and most likely, if he has found some, but not all of the relevant documents, he can enter an iterative feedback loop, adding extra documents to a constantly increasing set of

* To whom all correspondence should be addressed.

CCC–0010–4620/83/0026–0239$04.00

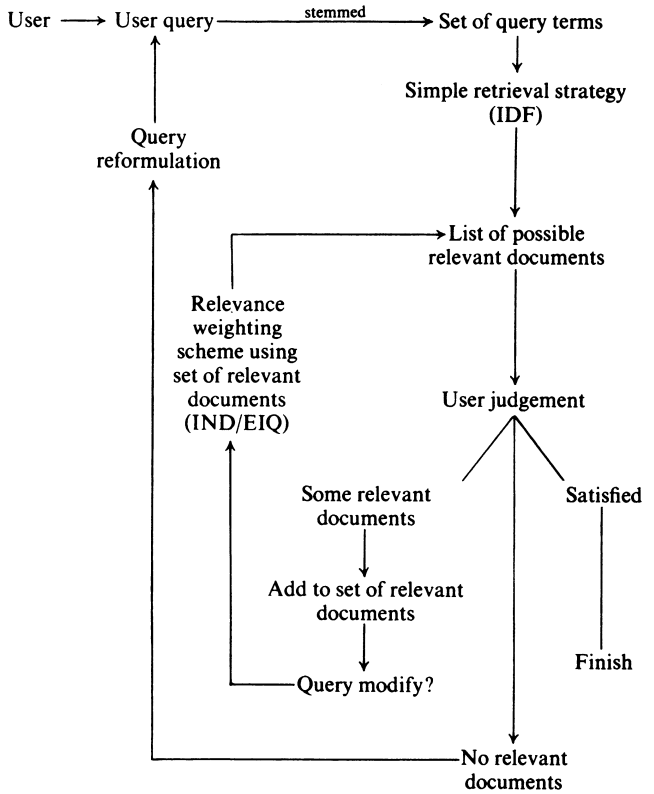

Figure 1

relevant documents. The increasing set of known relevant documents is used to refine the user's search using a relevance weighting scheme. Such a method uses the information contained in the known relevant documents found by the user in the search so far, to modify the weights assigned to each of the set of search terms.

One of the options that a user has when entering the feedback loop of Fig. 1, is query expansion or modification. A user can add to (or delete from) the set of search terms to further refine his search and 'home in' on a more exact representation of his information need. It is this aspect of the overall strategy in Fig. 1, that we will be concerned with.

3. IMPLEMENTATION PROBLEMS

In our experiments we will use two relevance weighting formulae. Both these formulae rely on the assumption of term independence, and they will be known as the IND and EIQ formulae. They are both implemented in the same manner.

For each search term in a query, a 'weight' indicating that term's 'degree of importance' is calculated using one of the relevance weighting formulae. Each document in the collection is then assigned a numerical 'score' comprised of the sum of the 'weights' of the search terms assigned to that document. The top-ranked documents, based on their numerical scores, are then presented to the user.

Each search term has associated with it 4 parameters, defined as:

R = number of relevant documents found by user
r = number of relevant documents found and indexed by the search term
N = number of documents in collection
n = number of documents indexed by the search term

The IND and EIQ relevance weighting formulae are defined as:

$$\text{IND} = \log \frac{r/(R-r)}{(n-r)/(N-n-R+r)} \tag{1}$$

$$\text{EIQ} = r \log \frac{rn}{RN} - (n-r) \log \frac{(n-r)n}{(N-R)n} - (R-r) \log \frac{(R-r)N}{N-n} + (N-R-n+r) \log \frac{(N-R-n+r)}{(N-R)(N-n)} \tag{2}$$

The basis for the form of these two functions may be found in Ref. 3. The IND formula is a form of the independence weight first derived by Robertson and Sparck Jones[5] and can be shown to be optimal. The EIQ formula has an *ad hoc* basis and is suboptimal; it was first used by Harper and van Rijsbergen[3] to overcome certain estimation problems. The IND formula should *theoretically* yield better retrieval performance than EIQ but does not always do so because of estimation difficulties. When only a small number of relevant documents are available for sampling (i.e. R. is small) as is usually the case, and $r = 0$ for one or more of the search terms, then problems arise when computing the IND weight since log0 is undefined. When computing EIQ under the same conditions, we get 0log0 which can be set equal to zero.

There are other ways of overcoming the difficulties associated with estimating IND when the log becomes undefined. The most common solution is to use a heuristic estimation rule which adds 0.5 to some parts of the IND formula.[5] It has been shown, however,[6,7] that this causes gross overestimation of some of the probabilities involved, which degrades the overall retrieval performances when using the IND weighting formula. However, in the absence of any realistic alternative, we have used the '0.5' rule in our experiments.

4. QUERY MODIFICATION

Query modification (or expansion) strategies were experimentally investigated in the late 1960s and early 1970s. At that time, the most common expansion strategy was to cluster index terms together before the user submitted a query, by constructing a similarity matrix of terms, and from this matrix to identify groups of similar objects based on some definition of what type of cluster was being used. Queries were then expanded by simply adding cluster-related terms. An example of such a strategy can be found in Ref. 8. Other, similar types of query expansion were investigated by Sparck Jones.[9] This early work in the area of query modification was relatively crude and unsophisticated and because of this, overall results were disappointing.

Barker *et al.*[10] have also conducted some query modification experiments which they have called 'refining a user search profile', based on the frequency of

occurrence of terms within a collection. Their results showed query modification to be useful, but their overall retrieval strategy was unsophisticated. Sparck Jones and Webster[11] have also reported on some query modification tests and conclude that when relevance weighting (IND or EIQ) is being used and a 'fair' amount of relevance information is available, relevance modifications (query modification) 'may be positively advantageous compared with unexpanded requests'.

More recently, van Rijsbergen *et al.*[12] have re-investigated query expansion but once again results have been disappointing. Their results can be attributed to the fact that their method added too many extra terms. They have remarked:

> 'This method . . . is admittedly only very crude, but it constitutes a first step in the direction of a more refined approach'

In our experiments we have used three sources of extra search terms. The first source of terms that we have used was the Maximum Spanning Tree (MST) constructed over the entire term vocabulary. The MST has been described in detail by Harper and van Rijsbergen.[3] It is a term–term dependence structure, somewhat similar to a thesaurus but derived from statistical associations between terms. The overall structure is that of a tree where every term in the collection points to, or is connected to, at least one other index term.

We have used the MST for query modification purposes by locating the original query terms in the MST and selecting some of the connections to some of the query terms, as extra search terms. Because the term–term connections in the MST represent some of the strongest statistical links between index terms, any terms connected in this tree, to a query term, could be as good a search term as the query term.

The second source of extra terms for query modification that we have used is the Nearest Neighbours (NN) to the query terms, using an appropriate similarity measure. The NN to a query term is the term most strongly related, statistically, to the given term, and hence NN modifications should yield better retrieval performances than MST modifications because NN connections will be stronger than MST connections, by definition.

The final and most promising source of extra search terms that we have experimented with has been the lists of index terms from the relevant documents found by the user in the search so far. Non-query terms which index relevant documents seem to offer the best possibilities as extra search terms because they allow a modification of the query which reflects the user's dynamic information need more accurately. MST or NN modifications use terms related to the original query terms, which in effect do not cater for a user who has expressed his information need badly through his original query. Query modifications which use terms from known relevant documents, would also eliminate the possibility of the $r = 0$ limiting case arising for the extra search terms added in, and thus the associated estimation problems would be reduced.

5. EXPERIMENTAL SETUP

5.1 Test collection

The test collection of documents and queries that we have used in our retrieval experiments is the NPL test collection originally prepared by Vaswani and Cameron at the National Physical Laboratory in England, and used recently in some of the Information Retrieval literature.[7] Documents and test queries from the collection were indexed by having the stopwords (AND, BY, WITH, etc.) removed and applying a stemming and conflation algorithm[13] to the remaining text. In the case of documents, both titles and abstracts were used in the indexing process, while the queries consist of a natural language statement of the user's information need.

Some basic statistics about the NPL collection are given in Table 1.

Table 1

	Documents	Terms	Queries	Relevance assessments
Numbers	11,429	7491	93	93
Max. Length	105	2511	13	84
Min. Length	1	1	2	1
Av. Length	19.96	30.45	7.14	22.40

This table shows that there are 11,429 documents and 7491 terms in the collection, which makes it a medium to large sized collection of documents. Each document is indexed by an average of 19.96 terms, and each term, in turn, indexes an average of 30.45 documents. By comparison with test collections used in other retrieval experiments, the NPL collection has quite a high level of indexing exhaustivity.

Of the 93 queries and relevance assessments, each query has an average of 22.40 relevant documents. These relevant documents were found by searching, not the entire collection, but for each query, the pooled output from some different search strategies. Thus the relevance assessments for the NPL collection are not completely exhaustive, but the known relevant documents probably constitute about 80% of all documents[14] relevant to the queries.

One interesting statistic about the NPL collection is that of the total of 664 query terms in the 93 test queries, 107 of these (16%) do not occur in any of the documents assessed as relevant to the query in question. (In the UKCIS test collection used by Harper[6] this figure was even higher, 51%.) This can be interpreted as saying that 16% of the NPL query terms are either negative discriminators of relevant and non-relevant documents (i.e. chosen by the user in a negative sense), or are simply the wrong terms, possibly chosen through lack of knowledge of the document collection on the part of the user. Since the latter of these possibilities is the most likely, this justifies adding terms related to the query terms, to the set of search terms. The effect that deleting some of the original query terms, might have on the overall retrieval effectiveness, can only be tested experimentally, and this will be done in our experiments.

Another interesting statistic from the NPL collection is that of the 2083 documents relevant to the 93 test queries, 49 of these (2.4%) are not indexed by any of the query terms of the pertinent queries. This gives even further justification to adding to the set of query terms, to

try to incorporate index terms which might occur in the relevant documents and so give those relevant documents a higher position in the collection ranking.

5.2 Evaluation

The evaluation of a document retrieval system can fall into one of two categories—evaluation of effectiveness and evaluation of efficiency. Efficiency deals with such things as the number of CPU cycles or disc accesses per retrieved document, and will not be considered further in this paper. Effectiveness deals with the quality of the material retrieved for the user in terms of the numbers of relevant documents retrieved at various points during the retrieval.

The measurement of retrieval effectiveness is one of the major research issues in Information Retrieval. The most commonly used pair of measures in the literature have been precision and recall[4] which are inversely proportional, and are computed for each query in the collection. They are defined as:

$$\text{Precision} = \frac{\text{No. documents relevant and retrieved}}{\text{No. documents retrieved}} \quad (3)$$

$$\text{Recall} = \frac{\text{No. documents relevant and retrieved}}{\text{No. documents relevant}} \quad (4)$$

N.B. Precision and Recall are usually given as percentages.

To measure the overall effectiveness of a document retrieval strategy, precision–recall figures are computed for each query in the test collection being used, and averaged to give an overall set of precision–recall pairs. The averaging technique that we will use to evaluate the retrieval effectivenesses obtained with various query modification strategies is called recall cutoff evaluation, and is described by Harper and van Rijsbergen.[3] This mechanism gives a set of precision values at standard recall values of 10%, 20%, ... 100%, by a method of pessimistic interpolation. This allows direct comparison between figures obtained using various query modification strategies.

One disadvantage of using the recall cutoff method of evaluation is that it presupposes that the full document collection has been ranked. As a consequence of this, the high recall figures should generally be ignored. Although this method of evaluation may not be the best available, it is the one that has been used previously in experimental work similar to what we will present later, and so we can do direct comparisons with the results of work by Croft and Harper,[15] Harper and van Rijsbergen,[3] Harper[6] and others.

One of the most difficult points associated with the evaluation of retrieval effectiveness is comparing two or more sets of performance figures. Several significance testing mechanisms have been proposed but none have proved universally acceptable. What we will do in this paper is use an *ad hoc* grading system with the following approximate scale for the difference between two sets of results:

Significant difference
Noticeable difference
Slight difference
Marginal difference
Same

5.3 Introduction to query modification experiments

In Section 2 of this paper we outlined a general retrieval strategy derived from probabilistic models of retrieval. Within this general framework the notion of query modification was expressed as an optional extra to be included in some, but not any specific, iteration of the overall strategy. In order to test the retrieval effectivenesses obtained with various query modification or expansion strategies, all the query modifications have to be carried out within the same experimental environment in order to be able to make comparisons. The retrieval setup which we have used will now be described.

An initial sample of documents was selected from the collection using inverse document frequency (IDF) weighting of the query terms. The size of this initial sample was 10 documents. Table 2 provides a breakdown of this initial sample.

Table 2

Number of queries: 93
Number of queries with all relevant documents: 1
Number of queries with no relevant documents: 9
Number of queries in evaluation: 83

distribution of queries by number of relevant documents in sample:

Number of rels:	1	2	3	4	5	6	7	8	9	10
Number of queries:	17	17	10	9	9	10	4	2	3	2

Of the 93 queries of the NPL collection, 83 have not retrieved all of their relevant documents, after the initial sample has been judged by the user. Table 2 shows that these 83 queries have an average of 3.77 relevant documents in their initial samples, whereas Table 1 has shown that overall, the queries have an average of 22.40 relevant documents each.

It is at this point in the overall retrieval strategy that we have incorporated query modification in our experiments. When the user has examined the 10 documents from the initial sample and has found some, but not all of the relevant documents to the query in question, then various forms of query modification by term addition or term deletion have been tried and the retrieval effectivenesses obtained with these modification strategies are reported in the next section.

Since this paper is concerned with the retrieval effectivenesses obtained using forms of query modification after the initial sample has been examined, what is of interest is the comparative effectivenesses obtained after the collection has been re-ranked. Therefore our results concentrate on the *relative* performances obtained with various forms of query modification, by presenting the precision–recall figures obtained from the *residual ranking* of the collection—after the documents from the initial sample have been removed. This type of evaluation has been used previously by Harper and van Rijsbergen[3] and Harper.[6]

For all the modification strategies dealt with in this paper, the benchmark result will be the precision–recall figures obtained by having no modification of the query at all, i.e. the search terms are the query terms. The following are the precision values of the residual ranking

at standard recall points (10, 20, ... 100) for no modification of the query, using the IND weighting formula in the first case and the EIQ formula in the second. These figures will be the benchmarks for the results we will present in the next section.

IND: 41.5 31.8 26.0 21.0 18.1 14.7 11.5 8.3 5.3 2.5
EIQ: 40.0 29.9 24.5 19.8 18.1 15.0 11.6 8.0 4.8 2.1

6. EXPERIMENTAL RESULTS

In this section we will present some of the experimental results we have obtained during our research into the retrieval effects of query modification. A fuller and more complete set of results can be found in Ref. 16. Each modification strategy will be given a numerical code and modifications will be referred to by their codes after their results have been presented. Each set of results will consist of two parts—the first part will be the precision figures for when the documents have been reranked using IND weighting, and will be denoted by '/IND' after the numerical code. The second set of figures will have been obtained for EIQ document reranking and will be denoted '/EIQ'. All the query modification results presented in this paper are given as an Appendix.

6.1 Adding MST-derived search terms

The results of this query modification strategy will now be presented first. Here we shall add $q/2$, q and $2q$ extra MST-derived terms (q is the number of original query terms) to allow us to see how different numbers of extra search terms affect retrieval.

01: Add each query term's strongest MST connection.* If this is already a search term, add the second strongest connection.

01/IND: 35.5 28.9 24.2 19.8 16.3 14.2 11.5 8.1 4.8 2.7
01/EIQ: 41.5 32.7 26.6 22.1 18.7 15.2 11.6 8.1 4.1 2.3

02: Double the size of the query by adding the strongest of the MST connections to any of the query terms.

02/IND: 30.3 23.8 18.2 15.4 13.0 10.9 9.1 7.3 4.5 2.3
02/EIQ: 40.1 31.1 25.0 21.1 17.9 14.2 11.3 9.1 5.2 2.8

03: Add the $q/2$ terms with the strongest of the MST connections to any of the query terms, where q is the size of the original query.

03/IND: 34.7 27.0 22.1 17.8 15.3 12.9 10.3 8.1 5.0 2.4
03/EIQ: 39.7 31.4 25.6 21.2 17.6 14.5 10.7 8.4 4.9 2.2

The results obtained with query modification strategies 01–03 show that the more terms that are added as search terms, the worse performance becomes in comparison to unexpanded queries, when the IND formula is used to rerank documents. When the EIQ formula is used, however, all the performance figures obtained with these modification strategies show a slight improvement over unexpanded queries.

* A query term's strongest MST connection is the index term connected to that query term, in the MST, with the highest value of similarity measure.

One possible explanation as to why the IND results are degraded when query modification is used could be because all query terms are treated as equal in modifications 01–03 whereas in fact 17% of the query terms do not appear in the relevant documents. Two possible ways of measuring the discrimination power of an index term are to use the values of the IND and EIQ weights as computed for each search term. Some modifications in which this has been done, are now given.

04: Rank the terms of a query by their IND weights and for each term in the top half of the ranking, add its strongest MST connection not already a search term.

04/IND: 34.9 30.0 25.3 21.1 17.2 14.9 11.9 8.1 4.9 2.7
04/EIQ: 40.8 32.9 27.1 22.5 19.2 15.7 11.8 8.0 4.8 2.3

05: Rank the terms of a query by their EIQ weights and for each term in the top half of the ranking, add its strongest MST connection not already a search term.

05/IND: 35.0 29.8 25.2 21.2 17.3 14.9 11.8 8.1 5.0 2.7
05/EIQ: 41.8 32.7 27.0 22.4 19.2 15.8 12.0 8.0 4.9 2.3

These results confirm that using the IND or EIQ index term weights as term discrimination value indicators improves performance, because both 04 and 05 yield better retrieval figures than 03. All three modification strategies add the same number of extra terms and therefore can be compared directly. This improvement in retrieval effectiveness can be observed for both IND and EIQ document reranking functions.

The retrieval performances of query modification strategies 04 and 05 when the document reranking is IND, however, are like all previous modification strategies tried so far in that the performance is degraded compared to unexpanded queries. Query modifications 04 and 05 have been degraded less than any other query modifications so far. When document reranking is by EIQ weighting, 04 and 05 yield results which are better than any other EIQ weighting figures.

6.2 Adding randomly selected terms

In an attempt to understand why the performances of IND weighting has led to degradation in performance when query modification is used, the next modification strategy adds index terms which are randomly selected from the collection vocabulary. These extra search terms are therefore completely unrelated to the query.

06: Add in $q/2$ extra search terms randomly selected from the collection vocabulary where q is the original number of query terms.

06/IND: 39.3 31.2 25.4 20.2 17.5 14.1 11.2 8.2 5.2 2.5
06/EIQ: 40.1 29.9 24.5 19.7 18.2 15.1 11.6 8.0 4.8 2.1

These results show a slight degradation of retrieval performance, although less of a degradation than with anything tried so far, when the IND formula is used to rerank documents, whereas the EIQ performance figures are about the same as with unexpanded queries.

6.3 Deleting terms

Since the modification strategies tried so far have yielded degradations in retrieval performance, the next logical

step is to try to delete terms from the original query. The term(s) chosen for deletion should be the ones with the poorest power of discrimination. Since IND and EIQ index term weights have been used to measure this in query modification strategies 04 and 05, they will be used again. As Table 1 has shown, some of the queries have only 2 index terms, so in order not to delete terms from 'shorter' queries, the following modification strategies were used.

07: If there are 4 or more query terms, delete the one with the lowest IND weight.

07/IND: 41.5 32.3 25.0 19.6 16.8 13.9 11.0 7.7 5.0 2.2
07/EIQ: 39.4 30.8 24.4 19.0 17.5 14.7 11.4 7.8 4.7 1.9

08: If there are 4 or more query terms, delete the one with the lowest EIQ weight.

08/IND: 41.5 31.6 25.7 20.3 17.4 14.0 11.2 8.0 5.0 2.4
08/EIQ: 40.0 30.1 24.8 19.4 17.8 14.8 11.6 8.0 4.8 2.3

These query modifications (07/08) have deleted one term from 78 of the 83 queries in the evaluation. The performances of IND document weighting for both modification strategies 07 and 08 show slight degradations compared to unexpanded queries, whereas the performances of EIQ document weighting are about the same as with unexpanded queries.

6.4 Adding NN terms

An alternative way to using the MST, of finding terms related to a given term, is to find a term's nearest neighbour. The nearest neighbour to a given term would be the index term most statistically related to that term and hence nearest neighbour modifications should yield better retrieval performance than MST modifications. The results obtained with such modifications are given below.

09: To each query term, add its NN. If this is already a search term, add the second NN.

09/IND: 35.4 28.3 23.8 20.1 16.7 14.4 11.4 8.2 5.1 2.9
09/EIQ: 40.8 32.5 27.1 22.2 18.8 15.3 11.5 8.2 4.9 2.5

The retrieval performance results using modification strategy 09 can be directly compared with those of modification 01, the equivalent MST modification, since both strategies add the same number of extra terms. Such a comparison shows MST modifications slightly better than (or not as degraded as) NN modifications for both IND and EIQ document weighting.

6.5 Adding terms from relevant documents

Some results of query modification strategies which obtain extra terms from known relevant documents will now be presented. At the start of the second iteration of the loop in the overall stragey of Fig. 1, each of the 83 queries in the evaluation has an average of 73 non-query terms in the term lists of the known relevant documents. This is obviously too many to add, so some measures of the discrimination powers of these extra terms will be used and the top-ranked terms will be selected for addition.

The discrimination power indicators that we will use will be the IND weights for the terms, the EIQ weights for the terms (as used before), the r values for the terms (the number of times the terms occur in the sample of relevant documents) and the Porter formula[17] defined as $Nr - Rn$. The number of extra search terms to be added in all the relevant document query modifications will be the same, namely $q/2$.

10: Rank non-query terms indexing at least one known relevant document, by their IND weights, and add the top $q/2$ of these (q is the original number of query terms).

10/IND: 37.6 29.2 23.8 19.5 16.6 14.5 11.1 8.1 5.1 2.5
10/EIQ: 38.1 29.9 23.9 19.3 17.6 14.5 11.4 7.8 4.7 2.0

11: Rank non-query terms indexing at least one known relevant document, by their EIQ weights, and add the top $q/2$ of these (q is the original number of query terms).

11/IND: 37.6 27.6 24.0 19.1 16.8 13.9 11.1 8.3 5.0 2.4
11/EIQ: 38.1 28.2 23.9 18.6 16.7 13.8 11.2 8.1 4.5 2.1

12: Rank non-query terms by their r values and add the top $q/2$ of these to the set of query terms.

12/IND: 38.5 29.4 23.2 18.9 15.9 13.2 10.6 8.3 5.0 2.5
12/EIQ: 38.9 30.0 23.6 18.9 16.7 12.7 10.5 8.0 4.6 2.1

13: Rank non-query terms by their $Nr - Rn$ values and add the top $q/2$ of these to the set of query terms.

13/IND: 38.8 29.5 23.2 18.9 16.3 13.2 10.6 8.3 5.0 2.5
13/EIQ: 38.8 30.0 23.8 19.0 16.8 13.8 10.5 8.1 4.6 2.1

All these performance figures show degradations when compared to unexpanded queries. MST and NN modifications do not perform better than the equivalent relevant document modifications under either IND or EIQ weighting for the reranking of the collection, which is as we expected. All the relevant document modifications tried indicate that the IND, EIQ, r and Porter formulae perform roughly the same as discrimination power indicators.

In the next Section we will try to give an understanding as to why our experimental results were not what we expected.

7. CONCLUSIONS

Probably the most important result from our experimental investigation into query modification is the fact that none of the strategies we have tried have yielded a significant or even noticeable improvement in retrieval effectiveness. Despite this, we have found some very interesting results.

A consistent pattern to emerge from the results presented here and in other experiments[16] is that the more extra search terms there are, the worse the degradation in retrieval effectiveness becomes. This result has also been found in experiments by Robertson *et al.*[7] As an example, modifications 03, 01 and 02 add successively more extra search terms under the same basic modification strategy, yet there is a very noticeable degradation in the retrieval performance figures as more terms are added. This occurs much more noticeably when the document reranking is done by IND rather than EIQ weighting. In fact, throughout all our query modification

results, EIQ weighting gives consistently better retrieval than IND. This is an important point which we shall return to later.

An interesting comparison of results can be made between modifications 01 and 09 which add the same number of extra search terms but use the MST and nearest neighbours (NNs), respectively. These results, and those obtained elsewhere[16] show that MST modifications yield marginally better retrieval than NN-based modifications. This result is unexpected since NN terms are, overall, statistically more associated to query terms, than are the MST derived terms. A similar type of unexpected result was found when comparing NN modifications with modifications which derive extra terms from known relevant documents. It was expected that relevant document modifications would yield better retrieval than NN or MST modifications, and this has indeed occurred. In fact the complete ranking of performance figures obtained with various query modification strategies was quite surprising and looks something like this:

	No modification
Retrieval	Randomly selected terms
performance	Terms from known relevant documents
degraded	Terms from the MST
↓	Terms as nearest neighbours

There are a number of parameters which have contributed to this overall set of results. The most important of these concerns the problem of estimation of probabilities from small samples. It has been shown elsewhere[5] that in order to cater for limiting cases in the estimation of probabilities, heuristic estimation techniques must be used. In addition to this, because the amount of available data for estimating probabilities is so small in the document retrieval situation (43 of the 93 queries have less than three relevant documents from a sample of 10), probabilities cannot be estimated accurately. For document retrieval without query modification, the detrimental effect that the estimation problem has on retrieval cannot be measured readily, as there is no real yardstick, but when this situation is compounded by the addition of extra search terms, there is a notable decline in retrieval performance which confirms that the poor estimations have had a significant effect on the overall results.

Another observation which fits into our explanation of poor results is that the modifications which in theory were what would have yielded the best retrieval (i.e. MST and NN modifications) in fact yielded the greatest degradation in retrieval results. This is because the detrimental effects that poor probability estimations have on retrieval are 'clustered' into an area around the query, whereas the poor estimation effects of, say, randomly selected extra search terms, are more 'scattered' throughout the collection. When extra randomly selected search terms are added to the query, the weights assigned to these terms have little net effect on the overall ranking of the collection *in the high precision area*. This is because randomly selected search terms will not usually co-occur with the original search terms and hence including these extra search terms will hardly have any effect on the top-ranked documents which are highly ranked because they are indexed by more than one original search term. Adding terms related to the query, i.e. MST or NN terms, means that the weights assigned to these extra search terms have a more significant influence on the top-ranked documents because these extra terms will co-occur with the original search terms more often. Because the EIQ document reranking formula has few estimation problems compared to the IND formula, it is only natural to expect that the EIQ function will perform consistently better than the IND function when query modification is incorporated into retrieval. This can be observed from the results presented in this paper and also results from Ref. 3.

One other point that could have had an effect on our results is that in our experiments we used one modification strategy for all queries in the collection. Since the queries are quite diverse in nature (i.e. 2 to 13 terms and 1 to 84 relevant documents) it follows that a modification strategy which may be good for one type of query may not be good for another type. To devise different modification strategies for different types of query, while being a quite valid proposal, is, however, beyond the scope of this paper.

A further point that we would like to make with respect to the estimation problem is that our query modifications could have been implemented too soon within the overall retrieval strategy. Sparck Jones and Webster[11] have concluded that only when a 'fair' amount of relevance information is available does any query modification yield an advantage over unmodified queries. It may be that when enough relevant documents have been identified by the user, probability estimates may be more accurately estimated and hence query modification may become a more feasible operation in improving document retrieval. This notion is partly supported by results from Ref. 16 and we intend to investigate this possibility as further research.

In conclusion we would like to re-emphasize that the estimation problem both with respect to small samples and limiting cases, is far more serious than has been indicated in previous IR literature. It is our contention that unless some better estimation rules are found which can accurately estimate probabilities from small samples of relevant documents, then the theoretical advantages of using some retrieval strategies with sound mathematical backgrounds, may never materialize into significant improvements in retrieval effectiveness.

REFERENCES

1. S. E. Robertson, The probability ranking principle in IR. *Journal of Documentation* **33**, 294–304 (1977).
2. G. Salton, Mathematics and information retrieval. *Journal of Documentation* **35**, 1–35 (1979).
3. D. J. Harper and C. J. van Rijsbergen, An evaluation of feedback in document retrieval using cooccurrence data. *Journal of Documentation* **34**, 189–216 (1978).
4. C. J. van Rijsbergen, *Information Retrieval* 2nd Edn., Butterworths, London (1979).
5. S. E. Robertson and K. Sparck Jones, Relevance weighting of search terms. *Journal of the ASIS* **27**, 129–146 (1976).
6. D. J. Harper, Relevance feedback in document retrieval systems: an evaluation of probabilistic strategies. *PhD. Thesis*, Cambridge University (1980).

7. S. E. Robertson, C. J. van Rijsbergen and M. F. Porter, Probabilistic models of indexing and searching. *Information Retrieval Research*, pp. 35–56, Butterworths, London (1981).
8. J. Minker, G. A. Wilson and B. H. Zimmermann, An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval* **8**, 329–348 (1972).
9. K. Sparck Jones, *Automatic Keyword Classification for Information Retrieval*. Butterworths, London (1971).
10. F. H. Barker, D. C. Veal and B. K. Wyatt, Retrieval experiments based on chemical abstracts condensates. *Research Report No. 2*, UKCIS, Nottingham, England (1974).
11. K. Sparck Jones and C. A. Webster, Research on relevance weighting 1976–1979. *British Library Report 5553*, Cambridge, England (1980).
12. C. J. van Rijsbergen, D. J. Harper and M. F. Porter, The selection of good search terms. *Information Processing and Management* **17**, 77–91 (1981).
13. M. F. Porter, An algorithm for suffix stripping. *Program* **14**, 130–137 (1980).
14. K. Sparck Jones and C. J. van Rijsbergen, Information retrieval test collections. *Journal of Documentation* **32**, 59–75 (1976).
15. W. B. Croft and D. J. Harper, Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* **35**, 285–295 (1979).
16. A. F. Smeaton, The retrieval effects of query expansion on a feedback document retrieval system. *Technical Report 2*, Department of Computer Science, University College Dublin (1982).
17. M. F. Porter, A proposal for writing a probabilistic IR system. *Information Technology: Research and Development* **1**, 131–156 (1982).

Received October 1982

APPENDIX—RESULTS SUMMARY

IND results:

01/IND:	35.5	28.9	24.2	19.8	16.3	14.2	11.5	8.1	4.8	2.7
02/IND:	30.3	23.8	18.2	15.4	13.0	10.9	9.1	7.3	4.5	2.3
03/IND:	34.7	27.0	22.1	17.8	15.3	12.9	10.3	8.1	5.0	2.4
04/IND:	34.9	30.0	25.3	21.1	17.2	14.9	11.9	8.1	4.9	2.7
05/IND:	35.0	29.8	25.2	21.2	17.3	14.9	11.8	8.1	5.0	2.7
06/IND:	39.3	31.2	25.4	20.2	17.5	14.1	11.2	8.2	5.2	2.5
07/IND:	41.5	32.3	25.0	19.6	16.8	13.9	11.0	7.7	5.0	2.2
08/IND:	41.5	31.6	25.7	20.3	17.4	14.0	11.2	8.0	5.0	2.4
09/IND:	35.4	28.3	23.8	20.1	16.7	14.4	11.4	8.2	5.1	2.9
10/IND:	37.6	29.2	23.8	19.5	16.6	14.5	11.1	8.1	5.1	2.5
11/IND:	37.6	27.6	24.0	19.1	16.8	13.9	11.1	8.3	5.0	2.4
12/IND:	38.5	29.4	23.2	18.9	15.9	13.2	10.6	8.3	5.0	2.5
13/IND:	38.8	29.5	23.2	18.9	16.3	13.2	10.6	8.3	5.0	2.5

EIQ results:

01/EIQ:	41.5	32.7	26.6	22.1	18.7	15.2	11.6	8.1	4.1	2.3
02/EIQ:	40.1	31.1	25.0	21.1	17.9	14.2	11.3	9.1	5.2	2.8
03/EIQ:	39.7	31.4	25.6	21.2	17.6	14.5	10.7	8.4	4.9	2.2
04/EIQ:	40.8	32.9	27.1	22.5	19.2	15.7	11.8	8.0	4.8	2.3
05/EIQ:	41.8	32.7	27.0	22.4	19.2	15.8	12.0	8.0	4.9	2.3
06/EIQ:	40.1	29.9	24.5	19.7	18.2	15.1	11.6	8.0	4.8	2.1
07/EIQ:	39.4	30.8	24.4	19.0	17.5	14.7	11.4	7.8	4.7	1.9
08/EIQ:	40.0	30.1	24.8	19.4	17.8	14.8	11.6	8.0	4.8	2.3
09/EIQ:	40.8	32.5	27.1	22.2	18.8	15.3	11.5	8.2	4.9	2.5
10/EIQ:	38.1	29.9	23.9	19.3	17.6	14.5	11.4	7.8	4.7	2.0
11/EIQ:	38.1	28.2	23.9	18.6	16.7	13.8	11.2	8.1	4.5	2.1
12/EIQ:	38.9	30.0	23.6	18.9	16.7	12.7	10.5	8.0	4.6	2.1
13/EIQ:	38.8	30.0	23.8	19.0	16.8	13.8	10.5	8.1	4.6	2.1