The potential and actual effectiveness of interactive query expansion

Mark Magennis and Cornelis J. van Rijsbergen
Department of Computing Science,
University of Glasgow,
Glasgow G12 8QQ
United Kingdom
(email: mark3@dcs.gla.ac.uk)

In query expansion, terms from a source such as relevance feedback are added to the query. This often improves retrieval effectiveness but results are variable across queries. In interactive query expansion (IQE) the automatically-derived terms are instead offered as suggestions to the searcher, who decides which to add. There is little evidence of whether IQE is likely to be effective over multiple iterations in a large scale retrieval context, or whether inexperienced users can achieve this effectiveness in practice. These experiments address these two questions. A small but significant improvement in potential retrieval effectiveness is found. This is consistent across a range of topics. Inexperienced users' term selections consistently fail to improve on automatic query expansion, however. It is concluded that interactive query expansion has good potential, particularly for term sources that are poorer than relevance feedback. But it may be difficult for searchers to realise this potential without experience or training in term selection and free-text search strategies.

1 Background and motivation

In free-text retrieval systems, queries can often be improved by adding extra terms that appear in relevant documents but which were not included in the original query. This is called query expansion. If the terms are provided by the system, as a result of relevance feedback for example, then it is called automatic query expansion. The potential of automatic query expansion for improving retrieval effectiveness has been investigated by many researchers, trying various sources of new terms with variable results.

Using relevance feedback terms has produced the greatest improvements (Harman 1992; Salton and Buckley 1990), although on occasion it has also been shown to degrade performance (Robertson et al. 1981; Smeaton and van Rijsbergen 1983). The effectiveness of relevance feedback query expansion depends on many factors, including the document ranking functions used (Smeaton and van Rijsbergen 1983), the document collection and queries (Salton and Buckley 1990), and the number of terms that are used (Buckley et al. 1994; Harman 1988).

Exploiting co-occurrence of terms in the document collection has generally achieved little or no effect on overall retrieval performance (Minker et al. 1972; Peat and Willett 1991; Smeaton and van Rijsbergen 1983). However, as with other query expansion term sources, some individual queries may be improved by adding co-occurring terms whilst others are degraded (Harman 1988).

SIGIR 97 Philadelphia PA, USA

Stemming algorithms such as those devised by Porter (Porter 1980) and Lovins (Lovins 1968), have been shown to offer small overall improvements in most situations but with great variation across queries, some being improved greatly, others being degraded (Harman 1987; Harman 1988; Hull 1996; Keen 1992).

Kristensen used a manually-constructed thesaurus in a limited domain (economics & environment) (Kristensen 1993). Adding loosely-defined synonyms, related terms, and narrower terms, resulted in a large improvement in overall recall at the expense of a small drop in precision. Voorhees & Hou used a general purpose thesaurus (WordNet) as a source of related concepts resulting in the improvement of some queries but degradation of others (Voorhees and Hou 1993).

These studies show that automatic query expansion is capable of producing large improvements in retrieval effectiveness, particularly when relevance feedback terms are used, but that the effectiveness varies greatly, particularly across queries. Although nearly all of these experiments have used only a single application of query expansion improvements can continue to be made over many iterations (Harman 1992).

Relevance feedback may produce poor query expansion terms for a number of reasons. The sample of relevant documents may be very small, terms may be extracted from non-relevant sections of otherwise relevant documents, and some relevant terms may also attract non-relevant topics that are already over-emphasised. It seems reasonable to assume that a searcher, given a list of the query expansion terms, will be able to distinguish the good terms from the bad terms. This is the assumption underlying the use of interactive query expansion. In interactive query expansion the potential query expansion terms are shown to the searcher as suggestions. The searcher then decides which to add and which not to add. This technique can be used with any source of

potential query expansion terms and it has been implemented in a number of operational systems.

CUPID uses relevance feedback to suggest search terms (Porter 1982). MUSCAT uses relevance feedback to suggest word stems and UDC (Universal Decimal Classification) numbers (Porter and Galpin 1988). One version of INSTRUCT used term clustering statistics and morphological processing of the query words to suggest related word stems (Wade and Willett 1988). CITE uses a dictionary and the MeSH thesaurus at the pre-search stage to suggest initial query terms which are related to the terms entered by the user. (Doszkocs 1983). However, none of these implementations have been evaluated to determine whether the interactive query expansion facility leads to improved retrieval effectiveness.

The earliest evidence for the potential of interactive query expansion was provided by Harman's experiments using the Cranfield 1400 test collection (Harman 1988). Her results showed that lists of candidate query expansion terms produced by a number of different methods could be improved upon by the selection of a subset that a searcher might be expected to be able to choose. Relevance feedback was used to produce a list of 20 candidate query expansion terms. Searchers' term selections were simulated by using only those candidate terms that occurred in at least one of the relevant unretrieved documents, on average 12 of the original 20. According to Harman, this simulated a "perfect" choice by the user.

The result was an improvement in retrieval effectiveness. The simulated interactive query expansion terms produced an average of 2.3 relevant documents from the next 20 retrieved, compared with 1.9 for the unfiltered list and 1.3 for no feedback at all. Overall the simulated interactive query expansion improved 122 queries and degraded 11, whilst the automatic query expansion improved 91 and degraded 24. Even greater relative improvements were achieved by simulating searchers' selections from candidate query expansion terms produced by a stemming algorithm (Lovins) and a synonym thesaurus since the suggested terms were much poorer.

These results show that worthwhile improvements are possible from interactive query expansion in the restricted context represented by the Cranfield collection. However, it is not possible to say whether searchers can achieve these improvements in practice, or whether the potential is the same in a more realistic large scale retrieval context.

Experiments using real searchers to carry out interactive query expansion have so far failed to show any significant improvements over automatic query expansion (Araya 1990; Hancock-Beaulieu et al. 1995). There is, however, some evidence to suggest that searchers may prefer the interactive method regardless of whether it improves performance (Koenemann and Belkin 1996) and this may in itself be a case for using it.

The experiments described in this paper aim to determine the potential and actual effectiveness of multiple iterations of interactive query expansion in a large scale realistic search context. The potential effectiveness, compared with automatic query expansion, is measured using a method similar to Harman's but with an improved simulation of good term selections. It is assumed that experienced users of interactive query expansion would be able to reach this level of performance. The 'experienced user' performance is compared with the performance of inexperienced interactive query expansion users in the same setting.

2 Experimental set-up

2.1 Experimental tasks

To make it possible to compare the retrieval effectiveness of automatic, experienced user interactive, and inexperienced user interactive query expansion, a single task context is used. However, there simply aren't any experienced users of multiple iteration interactive query expansion to be found and training experimental subjects would be too costly, both in time and in financial payments. Experienced users' selections therefore have to be simulated in some way.

The simulation of experienced user interactive query expansion requires 2,401 multiple-iteration searches for each query (see section 3.2 for details). It is clearly not possible to do this with real searchers so the searches, including the relevance judgement phase, are automated by using a standard test collection. Using a test collection raises problem for the real user interactive query expansion tests, however. To allow direct comparison with the retrieval performance of automatic query expansion the same documents, topics, and relevance judgements have to be used. But the interactive query expansion users are not then involved in their own tasks. Instead, they are being asked to search on predetermined topics that they may poorly understand. Steps that are taken to deal with this problem are described in section 4.

The task contexts used for this and all subsequent performance measurement experiments are taken from the TREC exercise. For these experiments, a small portion of the TREC data is used, consisting of 173,252 articles which appeared in the Wall Street Journal newspaper between 1987 and 1992.

The topics used for these experiments are from TREC-3 (topics 151-200). An example is shown in Figure 1. They are long enough to provide a full, unambiguous description of what is and is not relevant, so an experimental subject with no prior knowledge of the subject area should be able to follow them without difficulty. The topic titles are used as the initial queries because they resemble the range of queries that one would expect users to input. Two extreme examples are topic 168 "Financing AMTRAK", and topic 181 "Abuse of the elderly by family members, and medical and nonmedical personnel, and initiatives being taken to minimize this mistreatment".

Topic: Financing AMTRAK

Description: A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK)

Narrative: A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to AMTRAK would also be relevant.

Figure 1. TREC-3 topic number 168

Each search consists of multiple iterations of query modification and retrieval. Each iteration starts with the retrieval of the best 20 documents using the current query. These are assessed for relevance and 20 relevance feedback terms are generated. A subset of these terms, derived according to the technique being tested, is then added to the query and the next iteration begins.

2.2 The experimental system

The document collection is indexed using single, lower case, non-hyphenated words as index terms. A stoplist is used to reject words that are too common to be used as discriminators for retrieval purposes, such as *the* and *to*, and words that appear in more than half of all documents, such as *wall*, *street*, and *today*. Stemming is not used since interactive searchers may not be able to understand the meaning of stems and it provides little or no expected gain in retrieval effectiveness (Hull 1996).

The query-document matching function uses the inverse of the frequency ofoccurence of a query term in the document collection to measure the term's rarity. Individual documents are scored by summing the inverse document frequency (IDF) weights of all the query terms that appear in it. Ranking the documents according to their scores and making a cut-off gives the relevance feedback set. It is now well accepted that, if the corpus generally consists of quite long documents like those in the TREC Wall Street Journal corpus, retrieval effectiveness is improved by including in the matching function the frequency of occurence of each query term in the document (Harman 1992). However, this search engine doesn't use document term frequencies so the retrieval effectiveness might be expected to be lower than the state of the art. This may affect the query expansion, since the number of relevant documents available to provide feedback terms will be lower. However, the difference is not large enough to pose any methodological problem.

Relevance feedback terms are produced by ranking all the terms contained in the set of relevant documents using the relevance feedback term weighting function and making a cut-off. The relevance feedback term weighting function used by the IR engine is the Robertson and Spark Jones F4 function (Robertson and Sparck Jones 1976):

$$W_t = \log_2 \frac{\frac{r}{(R-r)}}{\frac{(n-r)}{(N-n-R+r)}}$$

where W_t = the weight for query term t, r = the number of relevant documents for the query containing term t, R = the total number of relevant documents for the query, n = the number of documents in the collection containing term t, and N = the total number of documents in the collection.

3 Automatic query expansion performance

The number of terms used for query expansion is known to affect performance. Harman found that, in the Cranfield 1400 test collection using a good term-weighting formulae, effectiveness improved as the relevance feedback term cut-off increased up to 20 terms, then gradually degraded with the addition of further terms (Harman 1988; Harman 1992). A completely different result was achieved by Buckley et al. using relevance feedback query expansion in a routing environment (Buckley et al. 1994). Performance continued to improve as the cut-off was increased, never degrading. The differences between the results of these two experiments illustrates the problems of interpreting results achieved in one situation in a different situation. There are many possible differences that could have an affect (some of these are discussed in (Buckley et al. 1995)). It is therefore difficult to predict how query expansion performance will vary with the cut-off in the context of these experiments, although various factors suggest that the situation will more closely resemble Harman's than Buckley's.

The cut-off that gives the best average retrieval effectiveness over a representative range of searches is found empirically. The effectiveness of using automatic query expansion with this cut-off provides the most stringent baseline for the performance of the interactive techniques.

For each of the 20 topics relevance feedback was performed on the 20 documents retrieved by the initial query, producing a ranked list of 20 potential query expansion terms. All terms ranked at or above a given cut-off were used for query expansion and another 20 documents were retrieved. This was repeated for four iterations of query expansion, thus retrieving a total of 100 documents for the search. Searches were carried out using all cut-offs between 0 and 20, 0 being no query expansion. The mean precision at 100 documents retrieved for each cut-off is shown in Table 1 and graphically in Figure 2.

Cut-off	0	1	2	3	4
Mean Precision	16.05	24.95	28.45	28.55	29.05

Cut-off	5	6	7	8	9
Mean Precision	29.5	29.85	29.4	28.45	28.47

Cut-off	10	11	12	13	14
Mean Precision	29.45	28.05	28	28.25	27.65

Cut-off	15	16	17	18	19
Mean Precision	28.05	28.3	28.15	28	27.8

Cut-off	20
Mean Precision	27.35

Table 1. Mean precision after four iterations of query expansion (100 documents retrieved) using different cut-offs

The general trend is similar to Harman's results - a sharp increase up to a peak, followed by a more gradual decrease (Harman 1988; Harman 1992). But the peak comes at only 6 added terms and the decrease looks anything but smooth. Although all cut-offs give a large improvement over no query

expansion there is little material difference in the scores for all the cut-offs between 2 and 20.

The mean precision at cut-off 6 is only just over 9% better than at cut-off 20, giving an average 2.5 extra relevant documents in the hundred retrieved. It is debatable whether a user would consider this an appreciable difference. It is also debatable whether this difference reflects a real difference in effectiveness rather than a statistical sampling error. Wilcoxon's matched-pairs signed-ranks test reveals no significant difference at p=.05 between the scores for cut-offs 6 and 20. The mean recall achieved by the best cut-off was 45.2% after 100 documents retrieved. Because some of the topics have many more than 100 relevant documents and only 100 documents are retrieved in the experiment, 100% recall is not attainable for some topics. The best possible mean recall over all topics that could be achieved by 100 documents retrieved is 64.58%. A recall of 45.2% is therefore perfectly acceptable.

The *precision/retrieval* graph in Figure 3 shows how five different cut-offs perform throughout the search. A cut-off of 5 is slightly unusual in that its precision rises sharply in the first iteration of query expansion when 40 documents have been retrieved. But by 60 and 80 documents retrieved it has been overtaken by cut-offs 6 and 10 which follow the smoother pattern shown by all other cut-offs. Cut-off 1 also follows this smooth pattern but with less overall effectiveness. The curve for cut-off 0 indicates the effectiveness of the unexpanded search.

An important aspect of the results is that the similarity in mean precision hides differences that are very variable from query to query. For some topics the best overall cut-off, which is 6, retrieves up to 10 times as many relevant documents as no query expansion. For others there is little or no improvement. In some cases the lack of improvement is due to there being few relevant documents in the dataset or the very good performance of the original query. But there are also some failures. For example, with topic 200 the unexpanded search finds only 28 of the 63 relevant documents but this is increased to only 34 by query expansion, and with topic 196 both searches achieve only about 50% recall.

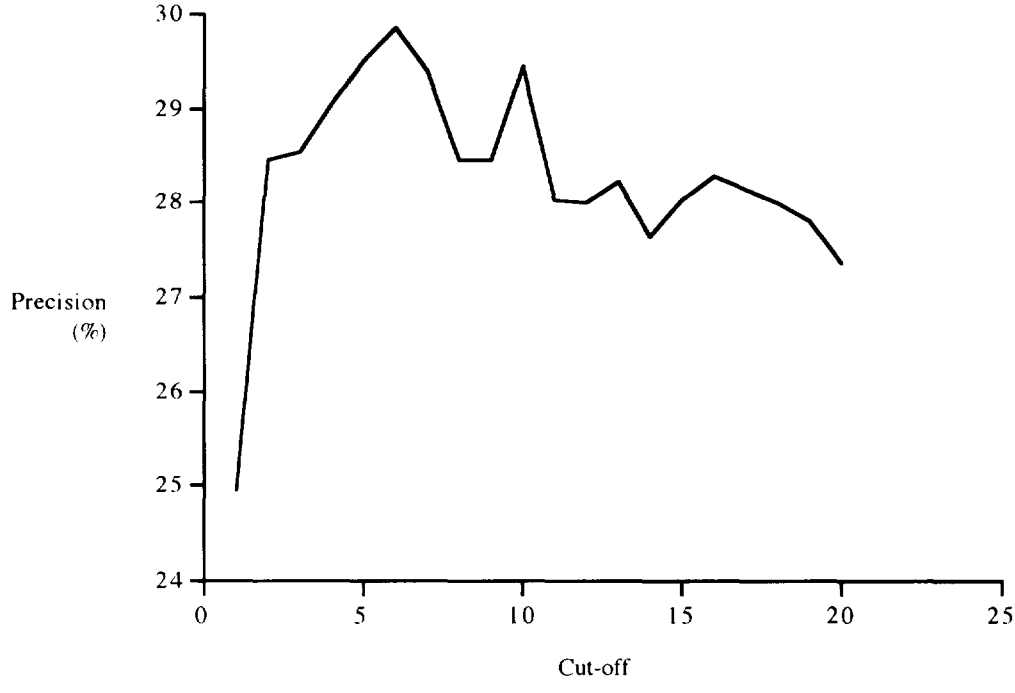

Figure 2. Mean precision after four iterations of query expansion (100 documents retrieved) using different cut-offs

The variability in the performance of different cut-offs between topics is well illustrated by Figure 4 which shows the precision at 100 documents retrieved for each cut-off on three selected topics. For topic 189 the low cut-offs are the most effective whilst for topic 190 they are the least effective. Topic 192 is different again, with relatively constant precision across all cut-offs.

The overall results show that automatic query expansion offers a large improvement in retrieval performance in this situation and that the best cut-off to use is relatively small compared with previous findings in other situations. As long as the cut-off is in the right region it is not critical to get the absolute best. There are good reasons to expect interactive query expansion to be capable of much more improvement than this since it goes further than just selecting the best overall cut-off. An interactive query expansion user can vary the cut-off between searches and between iterations of a single search. By making non-contiguous term selections the user also effectively re-ranks the relevance feedback terms before making a cut-off. This gives a clue about how to simulate experienced user performance - by re-ranking the relevance feedback terms according to their utility and making the best cut-off on each iteration of each search.

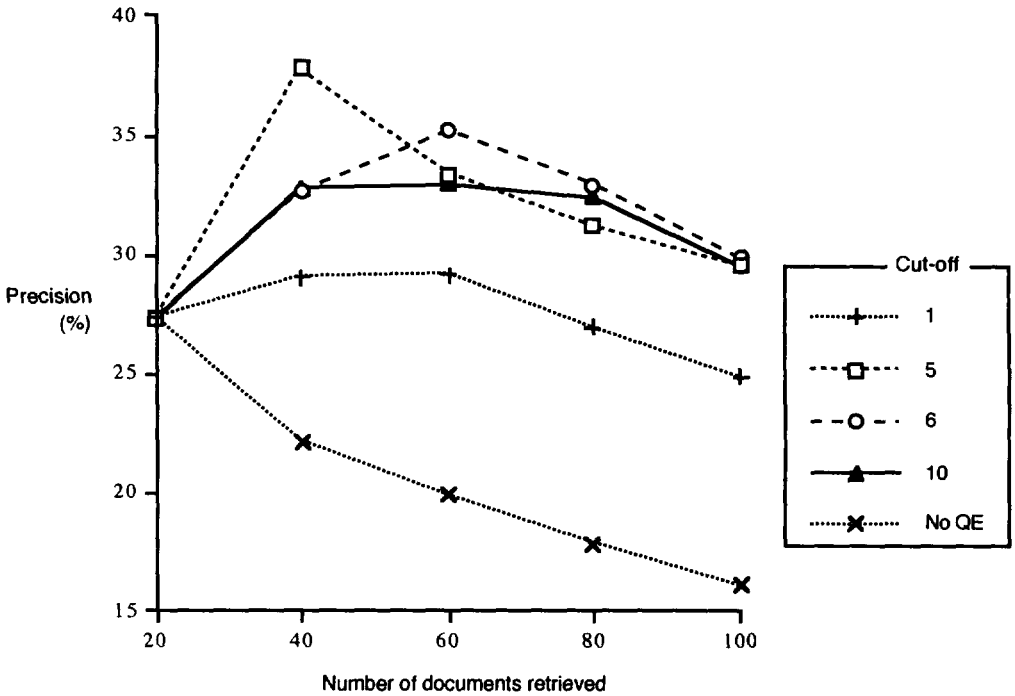

Figure 3. Mean precision after each of four iterations of query expansion for different cut-offs

4 Experienced user performance

This next experiment is an attempt to estimate how much retrieval effectiveness can be improved by experienced interactive query expansion users.

4.1 Simulating experienced users' term selections

Harman attempted to measure the potential of interactive query expansion by simulating users' selections. The approach was to devise a computable theoretical definition of 'good' terms and select all of those (Harman 1988). Harman's definition of a good term was *any term that appears in at least one of the unretrieved relevant documents*. This is a fair definition since it recognizes that the task is to select terms that will attract the target documents and it is certain that a term that doesn't appear in any of the target documents will not attract them. However, although attracting target documents is necessary, the real task is to attract them without also attracting unwanted documents. The IR system has a built-in function that weights terms according to their utility for discriminating one set of documents from another - the relevance feedback term weighting function. The utility of each of the suggested terms is therefore measured by using the normal relevance feedback term weighting function with the target documents in place of the relevant documents. The terms are then re-ranked by their resulting utilities and the selection is made by applying some cut-off to this ranked list. Harman's method is a special case of this in which the cut-off is made at the point where the utility becomes zero.

An approximation of the best cut-off point is found empirically, by trying all possible combinations of 5 different cut-offs over 4 iterations of query expansion. The number of searches required to test more than five cut-offs would have taken longer than the time available. The results for automatic query expansion suggest that the best numbers of terms to select are more likely to be in the range 0 to 10 than in the range 10 to 20, so it is sensible to use more samples at the lower end of the range. Searches are therefore carried out using every combination of the cut-offs 0, 3, 6, 10, and 20, over 4 query expansion iterations.

So experienced users' interactive query expansion performance is simulated by the following method:

- On each query expansion iteration rank the 20 relevance feedback terms according to the utilities calculated by applying the relevance feedback term weighting function to the unretrieved relevant documents
- Select the highest ranked terms by applying a cut-off to this ranking and add those to the query
- Run a search using every possible combination, over four query expansion iterations, of the cut-offs 0, 3, 6, 10, and 20
- The best retrieval effectiveness measures experienced user performance

There are various ways in which this simulation can be criticised. The first is that users, even experienced ones, do not have the precise knowledge of term distribution statistics that the relevance feedback term weighting function uses. It is therefore unlikely that a user, given a set of 20 terms, could rank them in utility order. Secondly, effective query expansion is not as simple as selecting terms that distinguish the target documents from all

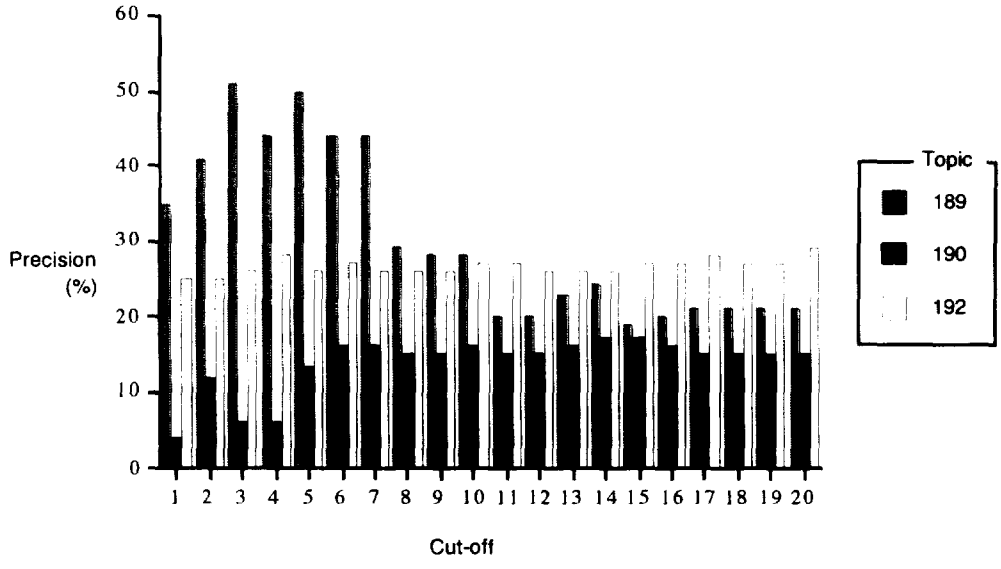

Figure 4. Precision after four iterations of query expansion using different cut-offs for three selected topics

the other documents in the database. It is also a matter of rejecting the dangerous terms that also attract similar but non-relevant topics. Experienced users are likely to realise this and make their selections accordingly. Thirdly, it is not clear that a strategy of targeting all the non-relevant documents will result in the best retrieval performance. Lastly, it is unlikely that even experienced users would be able to consistently make the best cut-off. This objection is reduced by the fact that the simulation doesn't necessarily make the best cut-offs itself since it only tries 5 out of the 21 possible cut-offs. In conclusion, there are some reasons to believe that experienced users might make better term selections than this simulation and other reasons to believe that they might make worse selections. Despite these objections, it is reasonable to assume that the simulation method will result in good term selections with retrieval effectiveness sufficiently similar to experienced user performance.

4.2 Effectiveness of experienced users' term selections

The retrieval results for searches simulating experienced users' term selections are shown as precision at 100 documents retrieved (Table 2). Also shown for comparison are the results for the best fixed cut-off searches and the results using Harman's method of selecting any term that appeared in at least one of the unretrieved relevant documents.

Compared with the best overall automatic query expansion search, which used the top 6 terms, precision is improved for every topic except one. The overall improvement is therefore unequivocal and doesn't require significance testing. Mean precision is increased from 30% to 35% giving the user an extra 5 relevant documents per search on average. This improvement is also very consistent. However, the experienced user simulation is outperformed on one topic by the best automatic query expansion search. The scores achieved on topic 189 for both the experienced user simulation and Harman's method are surprisingly low at 38% and 24% precision respectively.

		Automatic		Simulated interactive	
Topic	No QE	All 20 terms	Top 6 terms	Harman	Experienced user
181	1	1	1	2	2
182	16	35	36	39	42
183	64	73	78	76	81
184	4	25	24	27	27
185	23	42	45	49	58
186	13	13	14	19	22
187	15	43	47	46	59
188	3	6	3	6	6
189	5	21	44	24	38
190	4	15	16	15	18
191	7	32	31	32	37
192	21	29	27	27	30
193	2	14	15	18	18
194	4	8	9	10	12
195	3	27	27	35	36
196	41	42	44	44	54
197	22	33	37	36	41
198	42	44	59	54	69
199	3	8	6	8	8
200	28	36	34	37	42
Mean	16.05	27.35	29.85	30.2	35.0

Table 2. Precision achieved after 4 iterations of query expansion by various query expansion methods

The best automatic query expansion search for that topic, using a cut-off of 2, achieves 51% precision. This shows that the experienced user simulation does not always produce the best selection. Overall, Harman's method is not particularly good, achieving a mean precision only just better than the best automatic query expansion search.

5 Inexperienced user performance

The results of the previous experiments show the potential of multiple iteration interactive query expansion for improving retrieval performance. The improvement that experienced users can be expected to achieve is not great in the experimental search situation compared with that which has been achieved by other researchers in other situations. It is nevertheless worthwhile. This experiment aims to discover whether or not inexperienced users can achieve this improvement.

Retrieval performance is measured in the same search situation (tasks and system) that was used for the previous experiments. The only difference is the addition of an interactive document viewing and term selection phase.

5.1 Experimental procedure

Five subjects were recruited from the undergraduate and postgraduate population of the University's Computing Science department. Before searching, the system was demonstrated by the experimenter who carried out a single feedback iteration of a test search. Subjects then carried out a single feedback iteration of a test search themselves so that they could get used to the system and reveal any problems they might have. Each subject then searched on 4 randomly assigned topics of the 20 used in the experiment, going through four iterations of query expansion after the initial search. There was no time limit and searches took between 20 and 90 minutes with a mean of 1 hour.

On each iteration the titles of all the retrieved articles were displayed and subjects were able to view their full texts with occurrences of current query terms highlighted. They were asked to indicate which of the articles were relevant by clicking checkboxes. However, their judgements were ignored by the system and the official TREC relevance judgements were used for relevance feedback. This ensured that the retrieved documents, and therefore the suggested terms, were identical throughout the whole search to those used in the simulation of experienced users. The experimental instructions given to subjects purposely mislead them by suggesting that their relevance judgements were being used to produce the term suggestions. This was done to ensure that the subjects would become familiar with the topic and the results of the search. It was felt that if the subjects knew that their judgements were being ignored they would stop bothering to look at the articles properly. When the subjects had made their judgements they were checked against the official TREC judgements and if any were in disagreement the subjects were asked to reconsider them after re-reading the topic description. Again they were mislead into thinking that the system was merely 'making sure' rather than that it had an in-built list of relevant documents. The purpose of this prompting was to make sure the subjects fully understood the requirements of the search topic.

In the query expansion phase the 20 relevance feedback terms were displayed to the subject in a selection list. Any number from 0 to 20 could be selected. By default no terms were selected. Subjects could view examples of how the terms were used in relevant documents. This was done to enable them to overcome the problem of unfamiliar subject vocabulary. The selected terms were added to the query which was then resubmitted.

The instructions given to the subjects were carefully written to avoid giving any suggestions regarding how many and what type of terms to select.

5.2 Results of inexperienced user searches

The inexperienced users' term selections generally failed to improve on automatic query expansion. The complete results are shown in Table 3. The mean precision of 27.25 was lower than that achieved by any automatic cut-off between 2 and 20. However, Wilcoxon's matched-pairs signed-ranks test reveals that the difference is never significant at p=.05. Compared with the best overall automatic cut-off of 6, precision was improved for 8 topics and degraded for 10, with no change for the other 2. However, there were only two topics for which subjects managed to make a small improvement over all the automatic query expansion cut-offs.

Topic	Top 6 terms	Best automatic cut-off	Experienced user	Inexperienced user
181	1	1	2	1
182	36	40	42	38
183	78	81	81	60
184	24	26	27	22
185	45	53	58	54
186	14	14	22	13
187	47	47	59	42
188	3	6	6	4
189	44	51	38	45
190	16	17	18	17
191	31	35	37	28
192	27	29	30	28
193	15	16	18	12
194	9	10	12	7
195	27	29	36	8
196	44	52	54	42
197	37	37	41	37
198	59	60	69	40
199	6	8	8	8
200	34	36	42	39
Mean	29.85	32.4	35.0	27.25

Table 3. Precision achieved after 4 iterations of query expansion by various query expansion methods

Apart from topic 195, on which the inexperienced user did particularly badly, there was no marked variation in the lack of

improvement. This shows that the effects of interactive query expansion for inexperienced users is steady across topics, a result that mirrors what was found for the experienced user simulation. There is not enough data to draw any conclusions about differences between individual subjects although, given the relative lack of topical variation, it is unlikely that there would be much.

6 Results and discussion

Automatic query expansion has been shown to be capable of producing large improvements in retrieval effectiveness, particularly when relevance feedback terms are used, but the effectiveness varies greatly, particularly across queries. Interactive query expansion seems very promising, both as a way of fixing the problems automatic query expansion and of improving on its successes. It has been implemented in a number of operational systems and small scale simulation experiments have shown that it has potential for improving retrieval effectiveness. However, there has previously been no evidence to suggest whether the potential is the same in a more realistic large scale retrieval context, or whether inexperienced users can achieve this potential in practice.

The experiments described in this paper aimed to determine both the potential and the actual effectiveness of multiple iterations of interactive query expansion in a large scale realistic search context. The main results are as follows:

- Automatic query expansion using relevance feedback terms offers a large overall improvement in retrieval performance in this situation. The improvement is very variable across topics, however, with some showing little or no improvement. The best number of terms to use is relatively small compared with previous findings in other situations. Six terms were found to work best overall although the exact number is not critical.
- Interactive query expansion, as it might be performed by an experienced user, offers a small but significant further improvement. This improvement is very consistent across a range of topics.
- Inexperienced users of interactive query expansion did not make good term selections and failed to improve on automatic query expansion. This lack of improvement is also very consistent across a range of topics.

These experiments investigate an attempted improvement on what is arguably the most successful technique available in a free-text retrieval system for improving retrieval effectiveness - relevance feedback query expansion. Other sources of query expansion terms that have not been shown to be as useful may have more potential for improvement by interactive filtering. Sources such as semantically- or morphologically-based thesauri are complementary to relevance feedback, providing different types of terms. Improvements that can be made using interactive query expansion on these sources may prove more worthwhile.

The lack of improvement by the inexperienced users suggests that interactive query expansion may be difficult to use well. It adds a great deal of complexity on top of the automatic method. A free text search system using relevance feedback for automatic query expansion is very simple for searchers. All they have to do is provide an initial query and judge the relevance of the documents that are presented to them. There is no complex functionality to learn and the user interface can be minimal. Whatever methods it uses for searching are hidden from the searcher who has no chance to influence it once the initial query has been input. There is therefore no search logic or strategy involved in using it. This is often recognised as one of its benefits, making it particularly suitable for novice or infrequent searchers or those who lack the time, ability, or motivation to learn to use it more effectively.

Using interactive query expansion instead of automatic query expansion gives the searcher a greater degree of control over the search. There is some evidence to suggest that searchers may prefer the interactive method regardless of whether it improves performance (Koenemann and Belkin 1996) and this may in itself be a case for using it. It does, however, require extra effort from searchers. The extra term selection stage not only takes time and involves added user interface functions but also presents the searcher with a task that involves decision making and strategy. They must decide which terms to select and which to reject, knowing that their choices will have a direct bearing on the success of the search.

The majority of online-searching up till now has been done by librarians using commercial search systems. These systems offer an extensive functionality that requires complex query languages. In order to use these systems effectively librarians are given extensive training. In addition to this training, there is a huge amount of literature available concerning functions, tactics, and strategies for using these systems and exploiting their facilities. In contrast, there is no such training or advice available for using interactive query expansion in a free text search environment and there is nothing in the literature about strategies for free-text searching. Without good strategies and careful reasoning it is unlikely that a searcher will be able to use techniques such as interactive query expansion effectively.

The question of how searchers use, or could use, interactive query expansion is therefore an important research topic.

Acknowledgement

The authors would like to thank Dr. Alan Smeaton and the School of Computer Applications, Dublin City University, for providing machine and desk space during the preparation of this paper.

References

Araya, J.E. (1990). Interactive query reformulation and feedback experiments in information retrieval, PhD thesis (available as Technical Report 90-1115). Cornell University, Ithaca, New York.

Buckley, C., Salton, G. and Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. *Proceedings of 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, 292-300.

Buckley, C., Singhal, A., Mitra, M. and Salton, G. (1995). New retrieval approaches using SMART: TREC 4. *Proceedings of Fourth Text REtrieval Conference (TREC-4)*.

Doszkocs, T.E. (1983). CITE NLM: Natural-language searching in an online catalog. *Information Technology and Libraries*, 2, 364-380.

Hancock-Beaulieu, M., Fieldhouse, M. and Do, T. (1995). An evaluation of interactive query expansion in an online library catalogue with a graphical user interface. *Journal of Documentation*, 51(3), 225-243.

Harman, D. (1987). A failure analysis on the limitation of suffixing in an online environment. *Proceedings of 10th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New Orleans.

Harman, D. (1988). Towards interactive query expansion. *Proceedings of 11th annual international ACM SIGIR conference on research and development in information retrieval*, Grenoble, 321-331.

Harman, D. (1992). Relevance feedback revisited. *Proceedings of 15th annual international ACM SIGIR conference on research and development in information retrieval*, Copenhagen, 1-10.

Hull, D.A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 70-84.

Keen, E.M. (1992). Presenting results of experimental retrieval comparisons. *Information Processing and Management*, 28(4), 491-502.

Koenemann, J. and Belkin, N.J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. *Proceedings of CHI 96 International conference on Human Computer Interaction*, Vancouver, B.C., Canada, 205-212.

Kristensen, J. (1993). Expanding end-users' query statements for free-text searching with a search-aid thesaurus. *Information Processing and Management*, 29(6), 733-744.

Lovins, J.B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11, 22-31.

Minker, J., Wilson, G.A. and Zimmerman, B.H. (1972). An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8, 329-348.

Peat, H.J. and Willett, P. (1991). The limitations of term co-occurence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42, 378-383.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Porter, M.F. (1982). Implementing a probabilistic information retrieval system. *Information Technology: Research and Development*, 1(2), 131-156.

Porter, M.F. and Galpin, V. (1988). Relevance feedback in a public access catalogue for a research library: Muscat at the Scott Polar Research Institute. *Program*, 22(1), 1-20.

Robertson, S.E. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3), 129-146.

Robertson, S.E., van Rijsbergen, C.J. and Porter, M.F. (1981). Probabilistic models of indexing and searching. *Information Retrieval Research*, Butterworths, London, 35-56.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41, 288-297.

Smeaton, A.F. and van Rijsbergen, C.J. (1983). The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26, 239-246.

Voorhees, E.M. and Hou, Y.W. (1993). Vector expansion in a large collection. *Proceedings of First Text REtrieval Conference (TREC-1)*, 343-351.

Wade, S.J. and Willett, P. (1988). INSTRUCT: A teaching package for experimental methods in information retrieval. Part III. Browsing, clustering and query expansion. *Program*, 22(1), 44-61.