Re-examining the Potential Effectiveness of Interactive Query Expansion

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde
Glasgow. G1 1XH
Ian.Ruthven@cis.strath.ac.uk

ABSTRACT

Much attention has been paid to the relative effectiveness of interactive query expansion versus automatic query expansion. Although interactive query expansion has the potential to be an effective means of improving a search, in this paper we show that, on average, human searchers are less likely than systems to make good expansion decisions. To enable good expansion decisions, searchers must have adequate instructions on how to use interactive query expansion functionalities. We show that simple instructions on using interactive query expansion do not necessarily help searchers make good expansion decisions and discuss difficulties found in making query expansion decisions.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: - search process, relevance feedback.

General Terms

Experimentation, Human Factors

Keywords

Query expansion, Evaluation

1. INTRODUCTION

Query expansion techniques, e.g. [1, 5], aim to improve a user's search by adding new query terms to an existing query. A standard method of performing query expansion is to use relevance information from the user – those documents a user has assessed as containing relevant information. The content of these relevant documents can be used to form a set of possible expansion terms, ranked by some measure that describes how useful the terms might be in attracting more relevant documents, [13]. All or some of these expansion terms can be added to the query either by the user – *interactive query expansion* (**IQE**) – or by the retrieval system – *automatic query expansion* (**AQE**).

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.

One argument in favour of AQE is that the system has access to more statistical information on the relative utility of expansion terms and can make better a better selection of which terms to add to the user's query. The main argument in favour of IQE is that interactive query expansion gives more *control* to the user. As it is the user who decides the criteria for relevance in a search, then the user should be able to make better decisions on which terms are likely to be useful, [10].

A number of comparative user studies of automatic versus interactive query expansion have come up with inconclusive findings regarding the relative merits of AQE versus IQE. For example, Koenemann and Belkin [10] demonstrated that IQE can outperform AQE for specific tasks, whereas Beaulieu [1] showed AQE as giving higher retrieval effectiveness in an operational environment. One reason for this discrepancy in findings is that the design of the interface, search tasks and experimental methodology can affect the uptake and effectiveness of query expansion techniques.

Magennis and Van Rijsbergen [12] attempted to gauge the effectiveness of IQE in live and simulated user experiments. In their experiments they estimated the performance that might be gained if a user was making very good IQE decisions (the *potential* effectiveness of IQE) compared to that of real users making the query modification decisions (the *actual* effectiveness of IQE). Their conclusion was that users tend to make sub-optimal decisions on query term utility.

In this paper we revisit this claim to investigate more fully the potential effectiveness of IQE. In particular we investigate how good a user's query term selection would have to be to increase retrieval effectiveness over automatic strategies for query expansion. We also compare human assessment of expansion term utility with those assessments made by the system.

The remainder of the paper is structured as follows. In section 2 we discuss the motivation behind our investigation and that of Magennis and Van Rijsbergen. In section 3 we describe our experimental methodology and data. In section 4 we investigate the potential effectiveness of IQE and in section 5 we compare potential strategies for helping users make IQE decisions. In section 6 we summarise our findings.

2. MOTIVATION

In this section we summarise the experiments carried out by Magennis and Van Rijsbergen, section 2.1, some limitations of these experiments, section 2.2, and discuss the motivation behind our work, section 2.3

2.1 The potential effectiveness of IQE

In [11, 12] Magennis and Van Rijsbergen, based on earlier work by Harman [7], carried out an experiment to estimate how good IQE could be if performed by expert searchers.

Using the WSJ (1987-1992) test collection, a list of the top 20 expansion terms was created for each query, using terms taken from the top 20 retrieved documents. This list of possible expansion terms was ranked by applying the F4, [14], term reweighting formula to the set of *unretrieved* relevant documents. This set of documents consists of the relevant documents *not* yet seen by the user. This set could not be calculated in a real search environment as retrieval systems will only have knowledge of the set of documents that *have* been seen by the user. However, as query expansion aims to retrieve this set of documents, they form the best evidence on the utility of expansion terms.

Using these sets of expansion terms, Magennis and Van Rijsbergen simulated a user selecting expansion terms over four iterations of query expansion. At each iteration, from the list of 20 expansion terms, the top 0, 3, 6, 10 and 20 terms were isolated. These groups of terms simulated possible sets of expansion terms chosen by a user. By varying which group of terms was added at each iteration, all possible expansion decisions were simulated. For example, expansion by the top 3 terms at feedback iteration 1, the top 10 terms at feedback iteration 2, etc. The best simulation for each query was taken to be a measure of the best IQE decisions that could be made by a user; the *potential* effectiveness of IQE.

2.2 Limitations

One of the benefits of an approach such as that taken by Magennis and Van Rijsbergen is that it is possible to isolate the effect of query expansion itself. That is, by eliminating the effects of individual searchers and search interfaces the results can be used as baseline figures with which to compare user search effectiveness. However, in [11, 12], Magennis noted several limitations of this particular measurement of the potential effectiveness of IQE.

i. only certain combinations of terms are considered, i.e. the top 3, 6, 10 or 20 terms. Other combinations of terms are possible, e.g. the top 4 terms, and these may give better retrieval performance.

ii. real searchers are unlikely to use add a *consecutive* set of expansion terms, i.e. the top 3, 6, 10 or 20 terms suggested by the system. It is more likely that searchers will choose terms from throughout the list of expansion terms. In this way, users can, for example, avoid poor expansion terms suggested by the system.

iii. the ranking of expansion terms is based on information from the *unseen* relevant documents; ones that the user has not yet viewed. In a real search environment the expansion terms will be ranked based on their presence or absence in the documents seen and assessed relevant by the user.

iv. only one document collection was used. Differences in the creation of test collections, the search topics used and the documents present in the test collection may affect the results of their conclusions and restrict the generality of their conclusions.

2.3 Aims of study

In our experiments we aim to overcome these limitations to create a more realistic evaluation of the potential effectiveness of interactive query expansion. In particular we aim to investigate how good IQE could be, how easy it is to make good IQE decisions and investigate guidelines for helping users make good IQE decisions. We also investigate what kind of IQE decisions are actually made by searchers when selecting new search terms. In the following section we describe how we obtain the query expansion results analysed in the first part of this paper. These experiments are also based on *simulations* of interactive query expansion decisions.

3. EXPERIMENTAL SETUP

In this section we outline the experimental methodology we used to simulate query expansion decisions. The experiments themselves were carried out on the Associated Press (**AP** 1998), San Jose Mercury News (**SJM** 1991), and Wall Street Journal (**WSJ** 1990-1992) collections, details of which are given in Table 1. These collections come from the TREC initiative [16].

Table 1: Collection statistics

	AP	SJM	WSJ
Number of documents	79919	90257	74520
Number of queries used	32	39	28
Average words per query[1]	2.9	3.7	2.9
Average number of relevant documents per query	37.8	58.6	30.3

For each query we use the top 25 retrieved documents to provide a list of possible expansion terms, as described below. Although each collection comes with a list of 50 topic (query) descriptions, we concentrate on those queries where query expansion *could* change the effectiveness of an existing query. This meant excluding some queries from each test collection; those queries for which there are no relevant documents, queries where *no* relevant documents were retrieved in the top 25 documents (as no expansion terms could be formed without at least one relevant document), and queries where *all* the relevant documents are found within the top 25 retrieved documents (as query expansion will not cause a change in retrieval effectiveness for these queries).

In our experiments we used the *wpq* method of ranking terms for query expansion, [13], as this has been shown to give good results for both AQE and IQE, [4]. The equation for calculating a weight for a term using *wpq* is shown below, where

[1]Queries used in the experiments only. The query comes from the short *title* field of the TREC topic description.

the value r_t = the number of seen relevant documents containing term *t,* n_t = the number of documents containing *t*, R = the number of seen relevant documents for query *q,* N = the number of documents in the collection.

$$wpq_t = \log \frac{r_t/(R-r_t)}{(n_t-r_t)/(N-n_t-R+r_t)} \bullet \left(\frac{r_t}{R} - \frac{n_t-r_t}{N-R} \right)$$

Our procedure was as follows:

For each query,

i. rank the documents using a standard *tf*idf* weighting to obtain an initial ranking of the documents.

ii. use the relevant documents in the top 25 retrieved documents to obtain a list of possible expansion terms, using the *wpq* formula to rank the expansion terms.

iii. using the top 15 expansion terms, create all possible sets of expansion terms. For each query this gives 32 678 possible sets of expansion terms. This simulates all possible selections of expansion terms using the top 15 terms, including no expansion of the query. Each of these 32 678 sets of terms represents a *possible* IQE decision that could be made by a user.

iv. using each combination of expansion terms, add the combination to the original query and use the new query to rank the documents, again using *tf*idf*. That is, for each query, we carry out 32 678 separate *versions* of query expansion.

v. calculate the recall-precision values for each version of the query. Here we use a full-freezing approach by which we only re-rank the unseen documents – those not use to create the list of expansion terms. This is a standard method of assessing the performance of a query expansion technique based on relevance information, [3]

We only use the top 15 expansion terms for query expansion as this is a computationally intensive method of creating possible queries. In a real interactive situation users may be shown more terms than this. However, it does allow us to concentrate on those terms that are considered by the system to be the best for query expansion.

For each query in each collection, therefore, we have a set of 32 678 possible IQE decisions that could be made by a searcher. For each possible IQE decision we can assess the effect of making this decision on the quality of the expanded query. We use this information in several ways; firstly, in section 4, we compare the possible IQE decisions against three methods of applying AQE. We then, in section 5, examine potential strategies for helping searchers make good IQE decisions. In section 5 we also compare the possible IQE decisions against human expansion decisions.

4. COMPARING QUERY EXPANSION TECHNIQUES

In this section we examine the potential effectiveness of IQE against three possible strategies for applying AQE. In this section we compare how likely a user is to make better query expansion decisions using IQE than allowing the system to perform AQE. Our three AQE techniques are:

Collection independent expansion. A common approach to AQE is to add a fixed number of terms, *n*, to each query. Our first AQE technique simulates this by adding the top six expansion terms to all queries, irrespective of the collection used. The value of six was chosen without prior knowledge of the effectiveness of adding this number of terms to any of the queries in the test collections used.

Collection dependent expansion. The previous approach to AQE adds the same number of expansion terms to all queries in all collections. When using a specific test collection we can calculate a better value of *n*; one that is specific to the test collection used. To calculate *n*, for each collection, we compared the average precision over all the queries used in each collection after the addition of the top *n* expansion terms, where *n* varied from 1 to 15. The value of *n* that gave the optimal value of average precision for the whole query set was taken to be the value of *n* for each query in the collection.

These values could not be calculated in an operational environment, where knowledge of all queries submitted is unknown. However, it gives a stricter AQE baseline measure as the value of *n* is optimal for the collection used. The values for *n* are shown in Table 2, and is higher than the six terms added in the previous strategy.

Table 2: Optimal values of *n*

Collection	AP	SJM	WSJ
n	15	15	13

Query dependent expansion. The collection dependent expansion strategy adds a fixed number of terms to each query within a test collection. This is optimal for the entire query set but may be sub-optimal for individual queries, i.e. some queries may give better retrieval effectiveness for greater or smaller values of *n*. The query dependent expansion strategy calculates which value of *n* is optimal for individual queries. This may be implemented in an operational retrieval system by, for example, setting a threshold on the expansion term weights.

These three AQE methods act as baseline performance measures for comparing AQE with IQE.

4.1 Query expansion vs. no query expansion

We first compare the effect of query expansion against no query expansion; how good are different approaches to query expansion? In Table 3 we compare the AQE baselines against no query expansion: **t**he performance of the original query with no additional query terms. Specifically, we compare how many queries in each collection give higher average precision than no query expansion; the *percentage* of queries that are improved by each AQE strategy. Also included in this table, in bold figures, are the average precision figures given by applying the techniques.

As can be seen, all AQE strategies were more likely, on average, to improve a query than harm it. That is, all techniques

improved at least 50% of the queries where query expansion could make a difference to retrieval effectiveness.

The automatic strategy that is most *specific* to the query, the query dependent strategy, not only improves the highest percentage of queries– is most *stable* – but also gives the highest average precision over the queries - is most *effective*. Conversely the automatic strategy that is least effective and improves least queries is the one that is less tailored to either the query or collection – the collection independent strategy.

Table 3: AQE baselines and example IQE decisions.

Baseline	AP	SJM	WSJ
Collection independent	56%	72%	50%
	18.8	**23.8**	**18.4**
Collection dependent	72%	79%	53%
	19.0	**24.8**	**18.6**
Query dependent	75%	90%	86%
	20.1	**26.7**	**21.1**
IQE best	94%	97%	96%
	22.3	**29.1**	**22.4**
IQE middle	31%	38%	30%
	18.4	**22.9**	**18.1**
IQE worst	0%	0%	0%
	11.9	**15.8**	**14.0**

We can compare these decisions against possible IQE decisions. Firstly, in row 5 of Table 3, we show the percentage of queries improved, and average precision obtained, when using the *best* IQE decision for each query. This set of figures gives the best possible results on each collection when using query expansion. This is the highest potential performance of IQE using the top 15 expansion terms.

Comparing the performance of the best IQE decision against the AQE decisions, it can be seen that IQE has the potential to be the most *stable* technique overall in that it improves most queries. It also has the potential to be the most effective query expansion technique as it gives highest overall average precision. However this is only a *potential* benefit, as we shall show in the remainder of this paper it may not be easy for a user to select such an optimal set of terms.

For example, in the row 6 of Table 3 we show the performance of a middle-performing IQE decision. This is obtained, for each query, by ranking the average precision of all 32768 possible IQE decisions and selecting the IQE decision at position 16384 (half way down the ranking). This decision is one that would be obtained if a user makes query expansion decisions that were neither good nor poor compared to other possible decisions. This result shows that even fair IQE decisions can perform relatively poorly; improving less than half of queries and giving poorer retrieval effectiveness than any of the AQE strategies.

Finally, in row 7 of Table 3, we show the effect if a user was consistently making the worst IQE decisions possible, i.e. always choosing the combination of expansion terms that gave the lowest average precision of all possible decisions. Even though a user is unlikely to *always* make such poor decisions, these decisions are being made on terms selected from the top 15 expansion terms. So, although IQE *can* be effective it is a technique that needs to be applied carefully. In the next section we examine how likely a user is to make a *good* decision using IQE.

4.2 AQE vs. IQE

In this section we look at how difficult it is to select a set of expansion terms that will perform better than AQE or no query expansion. We do this by comparing how many of the possible IQE decisions will give better average precision than the AQE baselines. In Table 4 we show the results of this analysis. For each collection we show how many possible IQE decisions gave greater average precision than each of the three baselines (top row in columns 2-4) and how many of the decisions gave a *significantly* higher average precision than the baselines (bold figures in columns 2 – 4)[2].

Table 4: Percentage of combinations better than baselines

Baseline	AP	SJM	WSJ
No expansion	59%	69%	53%
	30%	**38%**	**21%**
Collection independent	45%	36%	41%
	9%	**11%**	**12%**
Collection dependent	47%	35%	43%
	13%	**9%**	**8%**
Query dependent	9%	10%	10%
	1%	**1%**	**2%**

What we are trying to uncover here is how likely a user is to make good IQE decisions over a range of queries. The argument for IQE, based on this analysis, is mixed. On the positive side over 50% of the *possible* IQE decisions give better performance than no query expansion, and over 20% of the possible decisions give significantly better performance (row 2). However, this also means that nearly half of the possible decisions will decrease retrieval performance[3] and most decisions will not make any significant difference to the existing query performance.

Compared against the best AQE strategy (query dependent), only a small percentage (9-10%) of possible decisions are likely to be better than allowing the system to make the query expansion decisions. Based on this analysis it

[2] Measured using a *t*-test ($p < 0.05$), holding recall fixed and varying precision. Values were calculated on the set of RP figures for each query not the averaged value.

[3] A small percentage (1%-3%) of possible decisions will neither increase nor decrease query performance.

appears that it may be hard for users to make very good IQE decisions; ones that are better than a good AQE technique.

The collection independent strategy is the most realistic default AQE approach as it assumes no knowledge of collections or queries. However, although 35%-45% of possible IQE decisions are better than the collection independent strategy, this still means that searchers are more likely to make a poorer query expansion decision than the system. This is only true, however, if users lack any method of selecting good combinations of expansion terms. In the next section we analyse potential guidelines that could be given to users to help them make good IQE decisions.

5. POSSIBLE GUIDELINES FOR IQE

In this section we try to assess possible instructions that could be given to users to help them make use of IQE as a general search technique.

5.1 Select more terms

One reason for asking users to engage in IQE is to give more evidence to the retrieval system regarding the information for which they are looking. Users, especially in web searches, often use very short queries [9]. Presenting lists of possible expansion terms is one way to get users to give more information, in the form of query words, to the system.

A useful guideline to give to users, then, may be to expand the query with as many useful terms are possible. In Table 5 we compare the size of IQE decisions that lead to an increase in retrieval effectiveness (*good* IQE decisions, Table 5, row 4) against those that led to a decrease in retrieval effectiveness (*poor* IQE decisions, Table 5, row 5). As can be seen, the size of the query expansion does not distinguish good decisions from poor decisions.

The size of the *best* IQE decisions (the average size of the combinations that gave the best average precision) is similar both to the average size of the good and poor combinations (Table 5, row 3). The sizes of the average of the best AQE decisions are also within a similar range (Table 5, row 2). So giving the system *more* evidence does not necessarily gain any improvement in effectiveness.

Table 5: Average size of query expansions

	AP	SJM	WSJ
Query dependent	6.63	7.10	8.46
IQE best	7.29	5.56	7.16
IQE good	7.35	7.60	7.27
IQE poor	7.61	7.27	7.44

5.2 Trust the system

A second approach might be to advise users to concentrate on the terms suggested most strongly by the system. These are terms that are calculated by the system to be the most likely to improve a query, and in our experiment are the terms with the highest *wpq* score. In Table 6, we present the average *wpq* value of the terms chosen in good and poor IQE decisions, and also in the best IQE and AQE strategies.

The average *wpq* value for terms in good (row 4) and poor IQE decisions (row 5) is relatively similar. This means that sets of terms with high *wpq* values are not more likely to give good performance than sets of terms with lower *wpq* values.

The average value for the best AQE decisions (row 2) is generally higher than that of the IQE decisions. This, however, results in part from the fact that the query dependent AQE strategy adds a consecutive set of terms taken from the top of the expansion term ranking. As these terms are at the top of the term ranking, they will naturally have a higher *wpq* value.

The average term score for the *best* IQE (row 3) decision is also higher than either the good or poor IQE decisions, so there is some merit in choosing terms that the system recommends most highly – those with high *wpq* values.

Table 6: Average *wpq* of terms chosen

	AP	SJM	WSJ
Query dependent	2.20	2.11	2.39
IQE best	2.94	1.91	2.26
IQE good	1.92	1.71	1.70
IQE poor	1.93	1.70	2.12

However, the lack of difference between the good and poor IQE decisions means we cannot *alone* recommend the user concentrates more closely on the terms suggested by the system. That is, highly scored terms are useful but the user must apply some additional strategy to select which of these terms to use for query expansion.

5.3 Use semantics

One of the more intuitive arguments in favour of IQE is that, unlike the statistically-based query expansion techniques, humans can exploit semantic relationships for retrieval. That is, people can recognise expansion terms that are semantically related to the information for which they are seeking and expand the query using these terms. However, investigations such as the one presented in [2] indicate that searchers can find it difficult to use semantic information even when the system supports the recognition and use of semantic relationships.

Consequently, in this section we outline a small pilot experiment designed to compare system recommendations of term utility against human assessment of the same terms.

5.3.1 System analysis of expansion term utility

The system, or *automatic*, analysis of an expansion term is based on the overall impact of adding that term to all possible IQE decisions that do not already contain the term. That is, we estimate the *likely* impact of adding a new expansion term *t* to an existing set of expansion terms.

For each query, each expansion term, *t*, belongs to 50% (16384) of the possible IQE decisions (and does not belong to 50% possible decisions, including no query expansion). In

effect these two sets of possible decisions are identical except as relates to *t*: adding *t* to each IQE decision in the latter set would give an IQE decision in the former set. By comparing the average precision of all IQE decisions that contain *t*, with the corresponding decisions that do not contain *t*, we can classify each of the top 15 expansion terms according to whether they are *good*, *neutral* or *poor* expansion terms. Good terms are those that are likely to improve the performance of a possible IQE decision (a set of expansion terms); neutral ones are those that generally make no difference and poor expansion terms are those that are likely to decrease the performance of a set of expansion terms.

We demonstrate this in Table 7, based on the TREC topic 259 '*New Kennedy Assassination Theories*' run on the AP collection. Each row shows what percentage of the 16384 possible decisions, not already containing the term in column 1, that are improved, worsened, or have no difference *after* the addition of the term. For example, the addition of the term *jfk* will always improve retrieval effectiveness. That is, adding the term *jfk* to any set of expansion terms will increase retrieval effectiveness. Conversely, adding the term *frenchi* will always reduce the retrieval effectiveness, and the addition of the term *warren*[4] will make no difference.

Table 7: Addition of expansion terms for TREC topic 259

Term	Improved	No difference	Worsened
jfk	100	0	0
oswald	3	0	97
dealei	37	4	59
kwitni	29	0	71
motorcad	64	4	32
marcello	100	0	0
warren	0	100	0
theorist	0	100	0
theori	18	0	82
depositori	67	0	33
documentari	40	19	41
belin	0	100	0
tippit	46	8	46
frenchi	0	0	100
bulletin	45	0	55

For simplicity, we classify terms simply by their predominant tendency. For the example in Table 7 the good terms are *jfk*, *motorcad*, *marcello*, and *depositori*. The poor terms are *oswald*, *dealei*, *kwitni*, *theori*, *documentari*, *frenchi* and *bulletin*, and the neutral terms are *warren*, *theorist* and *belin*. The term *tippit* is good and poor for an equal percentage of combinations and cannot be classified.

5.3.2 Human analysis of expansion term utility

The automatic classification of expansion term utility presented in the previous section was compared against a set of human classification of the same expansion terms.

We selected 8 queries from each collection and asked 3 human subjects to read the whole TREC topic and each of the relevant documents found within the top 25 retrieved documents. These were the relevant documents used to create the list of the top 15 expansion terms in the previous experiment. The subjects were given the full TREC topic description to provide some context for the search, and were shown the initial query that retrieved the documents. The subjects were then presented with the top 15 expansion terms. For each expansion term the subjects were asked whether they felt the term would be useful or not useful at retrieving additional relevant documents when added to the existing query[5].

We asked each subject to assess each of the 24 queries rather than distributing the queries across multiple subjects. This was to preserve any strategies the individual users may be employing when selecting expansion terms [8]. However, we did not ask the subjects to read the *non*-relevant retrieved documents as we felt this was too great a burden on the subjects.

The subjects' selection of expansion terms was compared against the automatic analysis from section 5.3.1 to compare the system classification against human classification of expansion term utility. The comparison was done in three ways; first we compare how good the subjects are at detecting good expansion terms, section 5.3.2.1, how good the subjects are at eliminating poor expansion terms, section 5.3.2.2, and examine the decisions made by the subjects, section 5.3.2.3.

5.3.2.1 Detecting good expansion terms

For each subject we examine first whether the subjects can detect *good* expansion terms; whether the subjects can recognise the expansion terms that are likely to be useful in combination with other expansion terms.

Table 8: Percentage of good expansion terms detected

	Subject 1	Subject 2	Subject 3
AP	73%	60%	63%
SJM	50%	40%	42%
WSJ	62%	32%	45%

In Table 8 we show the percentage of the good expansion terms, as classified in section 5.3.1, which were chosen by each subject as being possibly useful for query expansion. The

[4] From the Warren Commission which investigated the assassination of President Kennedy. This term and the term *theori* are the only ones to appear in the TREC topic description.

[5] If the subjects could not decide whether the term was useful/not useful, they could assign the term to the category 'cannot decide'.

subjects varied in their ability to identify good expansion terms, being able to identify 32% - 73% of the good expansion terms.

5.3.2.2 Eliminating poor expansion terms

If the subjects are not always good at detecting good expansion terms perhaps they are better are eliminating poor expansion terms? In Table 9 we show the percentage of expansion terms that were assessed as being poor by the system but good by the subjects. As in the previous section, the subjects' ability to correctly classify expansion terms varied with at least 25% of the poor expansion terms being rated as good by the subjects. The implication here is that subjects may have difficulty spotting poor expansion terms.

Table 9: Percentage of poor expansion terms classified as good by subjects

	Subject 1	Subject S2	Subject S3
AP	54%	36%	43%
SJM	39%	26%	35%
WSJ	38%	45%	39%

One reason for the poor classification of terms may be that the subjects are only choosing certain types of terms. In Table 10 we compare the cases where the system classification (column 2) agreed or disagreed with the subjects' classification (column 3) of terms.

Table 10: Comparison of system and subject classification

	System	User	S1	S2	S3
AP	**Good**	**Good**	692 (6.5)	570 (6.22)	666 (5.3)
	Poor	**Good**	622 (4.3)	914 (4.48)	601 (4.2)
	Good	**Poor**	429 (4.5)	830 (4.14)	578 (4.1)
	Poor	**Poor**	142 (1.7)	223(1.8)	178 (1.6)
SJM	**Good**	**Good**	1321 (7.5)	1831 (7.3)	1542 (7.7)
	Poor	**Good**	766 (3.7)	867 (3.7)	802 (3.7)
	Good	**Poor**	390 (2.8)	405 (3.8)	397 (3.5)
	Poor	**Poor**	53 (1.4)	253 (1.9)	179 (1.6)
WSJ	**Good**	**Good**	833 (5.2)	204 (2.2)	765 (4.5)
	Poor	**Good**	1496 (3.9)	682 (2.8)	881 (3.3)
	Good	**Poor**	285 (2.6)	598 (4.0)	270 (2.7)
	Poor	**Poor**	427 (1.8)	966 (3.0)	470 (2.3)

For each case we give the average collection occurrence of the terms and (the figure in parentheses) their average occurrence within the relevant documents. For example, for the terms on which subject 1 and the system agreed that the terms were useful, these terms appeared in an average of 692 documents in the AP collection and an average of 6.5 relevant documents.

Appearing in lots of relevant documents appears initially to correlate with an assessment of good expansion term utility. However the difference in relevant document occurrence between good/poor and bad/poor misclassification is often slight.

The most apparent pattern from Table 10 is that subjects tend to classify terms with a high *collection* frequency as being good expansion terms. Conversely terms with a low collection frequency are likely to be assessed as being poor expansion terms. This is not a universal pattern (Subject 2 on the WSJ collection for example does the opposite) but it is the main pattern and suggests that searchers may not being assessing which terms are useful but which terms are *recognisable*.

5.3.3 Subjects' reasons for expansion term selection

We discussed with each subject their reasons for their classification of expansion terms. Based on the subjects' reasons for classification and the later automatic classification, we can suggest three reasons for misclassification of expansion term utility.

i. *Statistical relationships are important as well as semantic ones*. Subjects tended to ignore terms if the terms appear to have been suggested for purely statistical reasons, e.g. numbers. In general this may be a sensible approach if the query does not mention specific numbers or dates. However, the documents in the static collections we used are only a sample of the *possible* documents on the topics investigated. In this case, strong statistical relationships may be useful for future retrieval.

ii. *Users cannot always identify semantic relationships*. Making good use of semantic information means being able to identify semantic relationships between the information need and the possible expansion terms. For specialised or unusual terms, the subjects could be unsure of the value of these terms unless the relationship between these terms and the information need was made clear in the documents.

However, being able to recognise why expansion terms have been suggested, and the searcher's ability to classify terms as useful or not, does not necessarily guarantee that the terms themselves will be seen as useful. Rather, we propose that searchers need more sophisticated support in assessing the potential quality of expansion terms.

iii. *Users cannot always identify useful semantic relationships for retrieval*. The difficulty most subjects experienced with selecting expansion terms is that, although they felt they could identify obvious semantic relationships, they could not identify which semantic relationships were going to attract more relevant documents. In short, the subjects felt they could not identify the effect of individual expansion terms on future retrieval. Instead the subjects concentrated mainly on terms they viewed as safe; those that were semantically related to the *topic* description rather than the retrieved relevant documents. That is, the subjects tended to concentrate on terms for *new* queries rather than modified or refined queries.

This type of decision-making can also be seen in other investigations, e.g. [15] which demonstrated that, although terms suggested from relevant documents *can* be useful terms,

they are often not used as a main source of additional search terms.

In a real interactive environment users can, of course, try out expansion terms, or add their own new terms, and see the effect on the type of documents retrieved. However, the lack of connection between expansion terms and documents used to provide those terms indicates that searchers may need more support in how to use query expansion as a general interactive technique.

6. CONCLUSIONS

In this paper we examined the potential effectiveness of interactive query expansion. This is mainly a simulation experiment and is intended to supplement rather than replace experimental investigations of real user IQE decision-making. There are several limitations to this work: for example, we only concentrated on altering the content of the query; future investigations will compare the results obtained here when we use relevance weighting in addition to query expansion. We also do not differentiate between queries although the success of query expansion can vary greatly across queries. We will consider this in future work, our intention here is to investigate the *general* applicability of query expansion.

The experimental results initially provided a comparison between AQE and IQE techniques. From Table 3, section 4.1, IQE has the potential to be an effective technique compared with AQE. One of the main claims for IQE is that searchers can be more adept, than the system, at identifying good expansion terms. This may be particularly true for certain types of search, e.g. in [6] Fowkes and Beaulieu showed that searchers preferred IQE when dealing with complex query statements. Subjects may also be better at targeting specific aspects of the search, i.e. focussing on parts of their information need.

However, the analyses presented here show that the potential benefits of IQE may not be easy to achieve. In particular searchers have difficulty identifying useful terms for effective query expansion. The implication is that simple term presentation interfaces are not sufficient in providing sufficient support and context to allow *good* query expansion decisions. Interfaces must support the identification of relationships between relevant material and suggested expansion terms and should support the development of good expansion strategies by the searcher.

7. REFERENCES

[1] Beaulieu, M. Experiments with interfaces to support query expansion. Journal of Documentation. 53. 1. pp 8-19. 1997.

[2] Blocks, D., Binding, C., Cunliffe, D., and Tudhope, D**.** Qualitative evaluation of thesaurus-based retrieval**.** Proceedings of the 6th European Conference in Digital Libraries. Rome. Lecture Notes in Computer Science 2458. pp 346-361. 2002.

[3] Chang, Y. K., Cirillo, C., and Razon, J. Evaluation of feedback retrieval using modified freezing, residual collection & test and control groups. The SMART retrieval system - experiments in automatic document processing. G. Salton (ed). Chapter 17. pp 355-370. 1971.

[4] Efthimiadis, E. N. User-choices: a new yardstick for the evaluation of ranking algorithms for interactive query expansion. Information processing and management. 31. 4. pp 605-620. 1995.

[5] Efthimiadis, E. N.. Query expansion. ARIST Volume 31: Annual Review of Information Science and Technology. Martha E. Williams (ed). 1996.

[6] Fowkes, H., and Beaulieu, M. Interactive searching behaviour: Okapi experiment for TREC-8. Proceedings of IRSG 2000. 22nd Annual Colloquium on Information Retrieval Research. Cambridge. Cambridge. 2002.

[7] Harman, D. Towards interactive query expansion. Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 321-331. Grenoble. 1988.

[8] Iivonen, M. Consistency in the selection of search concepts and search terms. Information Processing and Management. 31. 2. pp 180-186. 1995.

[9] Jansen, B. J., Spink, A., and Saracevic, T. Real life, real users, and real needs: A study and analysis of users on the web. Information Processing & Management. 36. 2. pp 207-227. 2000.

[10] Koenemann, J., and Belkin, N. J. A case for interaction: a study of interactive information retrieval behavior and effectiveness. Proceedings of the Human Factors in Computing Systems Conference (CHI'96). pp 205-212. Zurich. 1996.

[11] Magennis, M. The potential and actual effectiveness of interactive query expansion. PhD thesis. University of Glasgow. 1997.

[12] Magennis, M., and van Rijsbergen, C. J. The potential and actual effectiveness of interactive query expansion. Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 324-332. Philadelphia. 1997.

[13] Robertson, S. E. On term selection for query expansion. Journal of Documentation. **46**. 4. pp 359-364. 1990.

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

[15] Spink, A. and Saracevic, T. Interaction in information retrieval: Selection and effectiveness of search terms. Journal of the American Society for Information Science. 48**.** 8. pp 741-761. 1997.

[16] Voorhees, E. H., and Harman, D. Overview of the sixth text retrieval conference (TREC-6). Information Processing and Management. 36. 1. pp 3 - 35. 2000.