# Where do good query terms come from?


## Gheorghe Muresan

School of Communications, Information and Library Studies Rutgers University New Brunswick, NJ 08901 muresan@scils.rutgers.edu

## Dmitri Roussinov

W.P. Carey School of Business Arizona State University Tempe, AZ 85287 dmitri.roussinov@asu.edu



**This paper describes a framework for investigating the quality of different query expansion approaches, and applies it in the HARD TREC experimental setting. The intuition behind our approach is that each topic has an optimal term-based representation, i.e. a set of terms that best describe it, and that the effectiveness of any other representation is correlated with the overlap that it has with the optimal representation. Indeed, we find that, for a wide number of candidate topic representations, obtained through various query-expansion approaches, there is a high correlation between standard effectiveness measures (R-P, P@10, MAP) and term overlap with what is estimated to be the optimal representation.**

**An important conclusion of comparing different query expansion approaches is that machines are better than humans at doing statistical calculations and at estimating which query terms are more likely to discriminate documents relevant for a given topic. This explains why, in the HARD track of TREC 2005, the overall conclusion was that interaction with the searcher and elicitation of additional information could not over-perform automatic procedures for query improvement. However, the best results are obtained from hybrid approaches, in which human relevance judgments are used by algorithms for deriving terms representations. This result suggest that the best approach in improving retrieval performance is probably to focus on implicit relevance feedback and novel interaction models based on ostention or mediation, which have shown great potential.**


### Introduction

A key issue in current information retrieval (IR) research is how to take account of the searcher's profile and context in order to personalize the list of documents estimated by the system to be relevant (Ingwersen and Jarvelin, 2004). The research work reported here was triggered by our interest in investigating how implicit or explicit forms of relevance feedback can generate extra information, and how such information can be employed to alter the search results and boost search performance.

The High Accuracy Retrieval from Documents (HARD) track of the Text Retrieval Conference (TREC), organized by the National Institute for Standards and Technology (NIST), was introduced with the aim of exploring methods for improving the accuracy of document retrieval systems based on various forms of personalization (Allen 2004, 2005), so it provides an appropriate framework for conducting such work. The 2003-2004 runs of HARD TREC supported a simulation of implicit relevance feedback by including information which could be, in principle, obtained through logs or observations of previous user interactions with a retrieval system. This was implemented via metadata which specified the searcher's familiarity with a topic, as well as her preference with regard to document genre, geographic coverage or granularity. A simulation of explicit relevance feedback was also supported via clarification forms, through which the retrieval systems could get extra information via a brief interaction with the human searcher.

In 2005, clarification forms were the only means to obtain personalization information: for each topic submitted by an information seeker, simulated by a NIST assessor, participating research groups were allowed to generate a clarification form and to try to leverage the additional information elicited, in order to improve retrieval effectiveness (compared to a baseline obtained without such information). Typical questions submitted by HARD TREC participants in these forms were aimed at reducing query ambiguity (they asked whether some document titles seemed relevant, which of a number of cluster labels seemed a better match for the topic, or which of a number of candidate terms were more related to the topic) or at query expansion (they asked for additional descriptions of the information need).

Like most other participants in HARD TREC 2005, our group used a range of query expansion approaches, some automatic (based on query clarity, or on mining the web), some interactive (by asking the searcher, in the clarification forms, to provide extra information) and some mixed (we asked the searcher to filter expansion terms coming from automatic methods). The relevance judgments were made available only at the end of the TREC experiment, so it was not possible for participants to analyze the results and estimate which approaches worked better before submitting the personalized search runs to NIST. We submitted the runs that our informed guess estimated to be significant

improvements to our baseline. When the relevance judgments were made available and we were able to compute effectiveness measures, we were unpleasantly surprised to realize that, while our sophisticated approaches did better than the baseline (simple search based on the standard topic representation), they were not better than pseudo relevance feedback (PRF). PRF is a simple automatic procedure implemented as standard functionality in most IR toolkits, which assumes that the top ranking documents returned from a search are relevant, extracts the most representative terms from these documents and uses them for expanding the original query, and re-runs the search with the expanded query. At the TREC conference it became rather clear than other participants in the HARD TREC had a similar experience: sophisticated expansion techniques and simulations of interactions with the searcher (which are expensive in terms of time spent and cognitive effort) did not show a significant improvement over the standard PRF.

Those results triggered the work described here. We are systematically investigating sets of expansion terms coming from different sources of evidence and to assess their quality and potential to improve retrieval effectiveness. We must clarify that, while we are employing the HARD TREC experimental setting (document collection, topics, relevance judgments, clarification forms), the kind of investigation described here was not possible during the TREC experiment, when relevance judgments were not available. Based on these judgments, we can now establish upper-bounds of performance and compare them with a number of approaches to query improvement.

The objective of this work is two-fold. Firstly, and more importantly, we are interested in comparing methods for evaluating the quality of expansion term sets. Specifically, we investigate whether the level of term overlap with the estimated optimal set correlates with, and can predict, higher retrieval effectiveness. Secondly, we are comparing sets of expansion terms from different sources in order to estimate the quality of these sources and to inform future design of IR systems.

The rest of the paper is structured as follows: after presenting the experimental setting, we describe the methodology for our investigation, build optimal upper-bounds, analyze the quality of the original topic representations, analyze a number of query expansion term sets, from various sources, and discuss the results.

## Experimental Setting

The corpus used in HARD TREC 2005 and in our experiment is the AQUAINT Corpus, produced by the Linguistic Data Consortium (LDC), catalog number LDC2002T31 and ISBN 1-58563-240-6. It consists of newswire text data in English, drawn from three

sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service. The corpus is roughly 3GB of text and includes 1,033,461 documents (about 375 million words of text). All documents in the collection were used for the HARD evaluation.

The 50 topics were selected from among existing TREC topics on which automatic approaches had produced low retrieval effectiveness in previous years; the intention was to verify if the simulated interaction with the human searcher could improve performance. Because those old topics were to be judged on a new corpus they were manually vetted to ensure that at least three relevant documents existed in the AQUAINT corpus for each of them. As the original authors of the adopted topics were not available to perform relevance judgments, at least some degree of consistency in judging was insured by having the same NIST assessor answer clarification forms and judge the relevance of the submitted documents for each topic. No attempt was made to ensure that the assessor's notion of relevance matched that of the original topic author.

Documents were judged as either *not relevant*, *somewhat relevant*, or *highly relevant*. For consistency, we adopted the same interpretation for the judgments as used officially in HARD TREC: for evaluation purposes of this experiment, judgments of somewhat relevant and highly relevant were both treated as relevant.

While R-precision (*R-P*) was the official effectiveness measure of the track, the participants were also encouraged to report precision at 10 retrieved documents (*P@10*) and mean average precision (*MAP*), in order to provide a better picture of the effect of the techniques employed (Buckley and Voorhees, 2005). For computing these effectiveness measures, we employed the standard trec_eval tool provided by NIST. It is worth noting that this was a precision-oriented experiment, with particular focus on top-ranking documents.

In terms of software tools, we used the Lemur open source IR toolkit, widely used in TREC, which provides functionality for indexing and searching, as well as additional functions for computing query clarity, clustering, etc. Based on preliminary tests to compare the effect of various parameters, we chose a combination of indexing parameters and pre-processing tools that tend to be effective, including the Krovetz stemmer (Krovetz, 1993) and the SMART stopword list. As retrieval model we adopted the popular TfIdf model. It tends not to yield best effectiveness, but it is the only model implemented in Lemur's modules for retrieval based on both flat, unweighted queries (RetEval, RelFBEval) and structured, weighted queries (StructQueryEval). Therefore, we were able to consistently use the same underlying retrieval model throughout our experiments, and thus avoid the potential effect of the model compounding the results

We obtained our baseline run by running Lemur on queries that comprised the topic titles and descriptions (see the Appendix). For comparing the effectiveness of different sets of results (called "runs" in TREC), and thus the quality of the expansion terms used to produce them, we used matched-pairs Wilcoxon tests. This is justified on grounds that the difference in effectiveness scores may be larger between individual topics than the difference that we want to observe, between methods or sources of query expansion terms. Also, this non-parametric test is appropriate because many of the effectiveness scores are not normally distributed.

## Methodology

First, we use relevance judgments to build optimal expansion term sets; details are given in the next section. We then build candidate expansion term sets, based on a variety of sources and with a variety of methods and parameters. By evaluating the quality of these sets, we will be able to conclude which sources and which methods of query expansion work better.

We propose two ways to estimate the quality of the expansion sets or, more generally, of queries as representations of topics. The first is simply to compare them with the optimal sets by applying set operations and looking at the overlap. Because the sizes of the expansion sets vary, measuring the overlap is justified only if some form of normalization is applied. Considering that some of our sets were generated by the human searchers during the search interaction, and thus relatively short, we measure the overlap at cut-offs of 10, 20 and 30 terms. Note that virtually all automatic approaches to generating candidate terms for query expansion assign a weight to each term. We use the weights to rank the terms before applying the cutoff and computing the overlap. In the future we plan to actually consider correlation of weights rather than overlap, hoping for more accuracy in estimating the quality of topic representations.

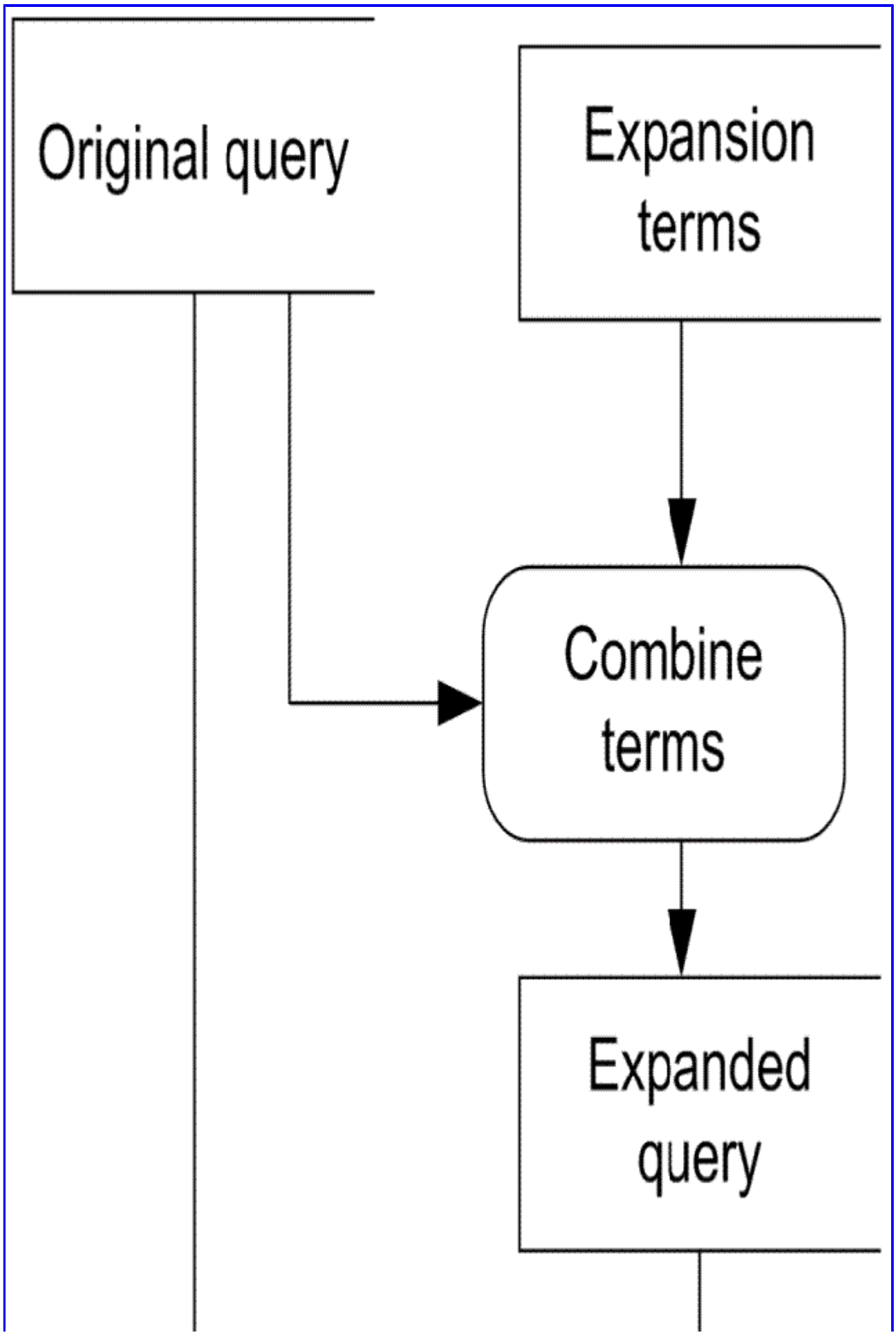
Original query
Expansion terms
Combine terms
Expanded query

Figure 1. Standard query expansion procedure

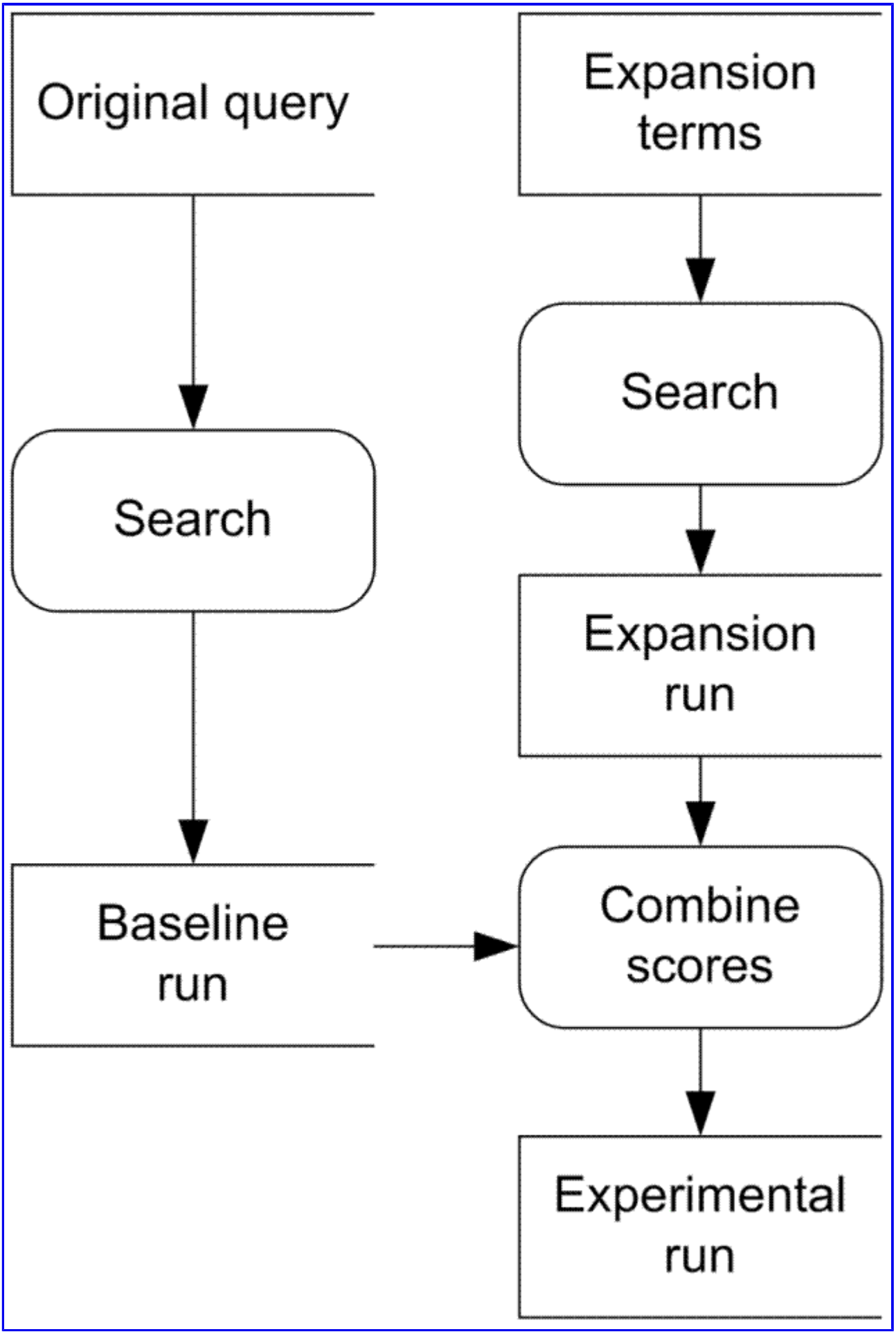
Original query
Expansion terms
Search
Search
Expansion run
Baseline run
Combine scores
Experimental run

Figure 2. Expansion based on score combination

The second approach looks at the actual effectiveness improvement effected by the expansion terms. There are two approaches to using expansion terms. The standard approach, depicted in Figure 1, is to combine the original query terms with the expansion terms either by a simple concatenation, or by applying a weighting scheme such as the Rocchio formula (Rocchio, 1971). In practice negative feedback is rarely used, so the formula becomes:

expandedQuery = originalQuery + *w* * expansionTerms

where expandedQuery, originalQuery and expansionTerms are vectors of term weights, and *w* is a weighting coefficient that controls the contribution of the new expansion terms, typically based on the confidence assigned to a certain source of evidence. In practice it is common that *w = 1*, i.e. the expansion terms are simply appended to the original query.

The alternative approach, depicted in Figure 2, uses the expansion terms as a query and it generates an "expansion run", a list of documents that match the "expansion query". If a retrieval model with a linear weighting formula is employed, then the same effect as the standard approach can be obtained by combining the scores in the baselines with those in the expansion run, and using the same weighting coefficient:

experimentalRun = baselineRun + w * expansionRun

The standard approach is implemented in most experimental IR systems and is therefore very easy to use. In Lemur, for example, the researcher can specify a file with relevance judgments, together with the weighting coefficient for relevance feedback (the function that implements this is RelFBEval). Alternatively, the researcher can use the standard retrieval function (RetEval), but set the flag for the optional pseudo relevance feedback effect and specify the number of top ranked documents and the number of top weighting terms to be considered.

However, this approach is not necessarily convenient for IR experiments, as it confounds the quality of the expansion terms with the effect of the weighting scheme and with the quality of the original query. The alternative approach allows one to look not only at the effectiveness of the final experimental run and to compare it to the baseline, but also to isolate the effectiveness of the expansion run, which directly reflects the quality of the expansion terms.

Lemur does not provide functionality for combining runs, based on a linear combination of their scores, so we have to write scripts to that effect. In fact, combining the baselineRun

with the expansionRun is only necessary if we want to compare the two combinations of evidence approaches. For evaluating the quality of expansion term sets, it is sufficient to compute the expansion run scores.

## Optimal Upper-Bounds

### *Choosing Upper-Bounds*

An ideal query would be one which, for a certain information need, would retrieve all the relevant documents and no non-relevant documents. Although such an ideal query is probably impossible to create, we attempt to use the existing relevance judgments to build "optimal queries", which achieve upper-bounds of performance. The reasoning is as follows: if an IR system with relevance feedback (RF) capability is given a set of documents judged relevant, it performs a statistical analysis of the documents and it produces a weighted terms representation of those documents. The terms can be ranked based on their weights, and the top ranking terms can be used for query expansion. If all and only those documents that are known to be relevant for a topic are used in this relevance feedback process, the obtained representation is optimal in terms of retrieval effectiveness.

Therefore, we use Lemur's relevance feedback module, RelFBEval, to compute weighted terms representations for the sets of documents judged relevant for each of the 50 test topics, using as input empty queries and the set of all documents judged relevant by the NIST assessors. In NIST terminology, these judgments are called the "qrels"; therefore, expansion term sets obtained based on them will be labeled as "qrels" in some of the tables included in the paper. The execution of RelFBEval produces two outcomes: (i) the optimal term representation of each topic; and (ii) the results of searching the corpus based on the optimal representations, i.e. our "optimal run", which constitutes our upper-bound of performance.

The term "optimal" is relative, as the outcome of executing RelFBEval dependends on a set of parameters, the most important of which being: (1) feedbackDocCount, which specifies how many documents judged relevant should be considered; and (2) feedbackTermCount, which specifies how many query terms should be generated, and used for searching. Exploring the effect of those two parameters is made more difficult by the distribution of the number of documents judged relevant over the topics: this number varies between 9 (for topic 345) and 376 (for topic 354). Therefore, in choosing different values for feedbackDocCount, we considered different percentages of the number of relevant judgments for each topic: all 100%, 75%, 50%, 25% and 10%. For feedbackTermCount we took absolute values: 10, 20, 30, 50, 75, 100, 150, 200, 10000

(the intent for the last value was to cover all the terms that may be extracted from the documents used for feedback).

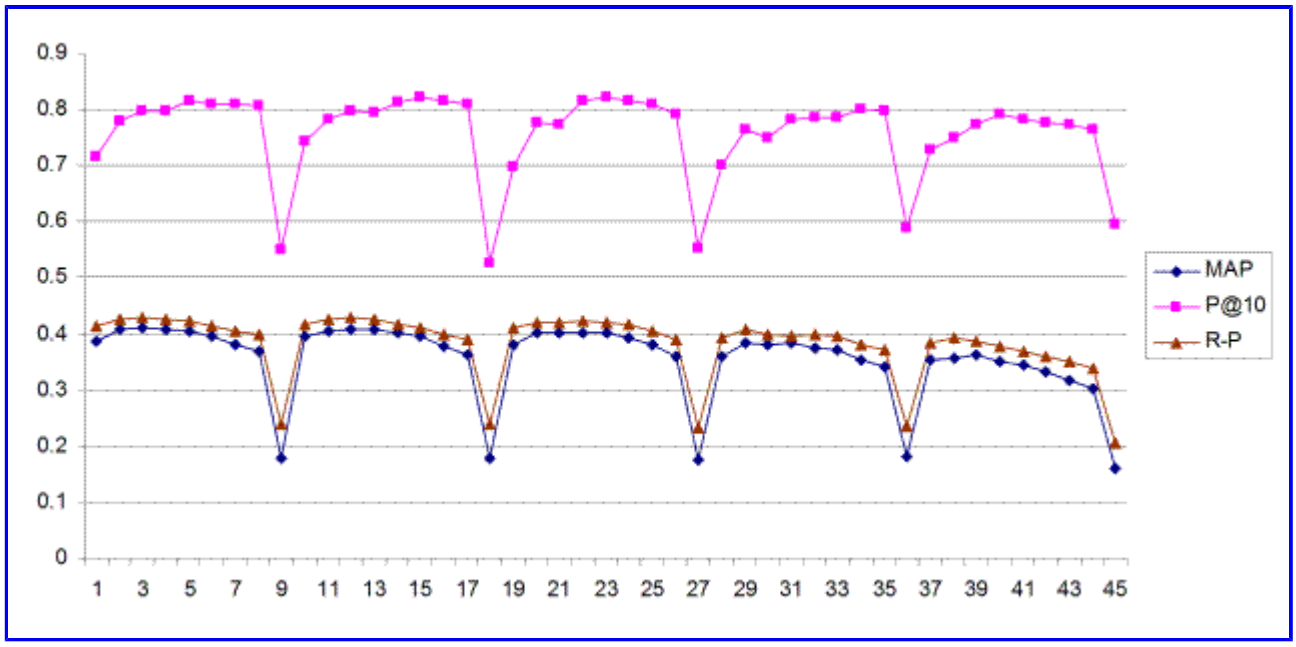


Figure 3. The effect of relevance judgment parameters on retrieval effectiveness

Figure 3 depicts the retrieval effectiveness, as measured by *MAP*, *P@10* and *R-P*, of different candidates to the title of optimal run. The five clearly visible groups of results correspond to the distinct values for feedbackDocCount, decreasing from 100% to 10%; within each group, feedbackTermCount increases from 10 to 10,000. First of all, as each group ends with a dramatic drop in performance, it is clear that representing each topic by all the terms that appear in a sample of relevant documents is rather disastrous; representing each topic by a "reasonable" number of terms gives much better performance. For P@10, it appears that performance peaks at 50-75 terms, while for *MAP* and *R-P* (which are highly correlated) it is best to consider just the best 10-30 terms.

While Figure 3 shows only summaries and not the variability of the data, it does suggest choosing as optimal run the one obtained based on all relevant judgments, and the most representative 30 terms. This result is encouraging for our experiment, because it makes it justified to compare manual representations with the optimal representation and to use term overlap at cutoffs of 10, 20 and 30 as measures of query quality: human searchers typically submit short queries and, even when prompted to give more details, they cannot be expected to generate more than 10-30 query terms.

Table 1 shows more details: for each of the 9 runs based on the full set of judgments, it shows the mean and standard deviation of the derived run, as well as the result of a Wilcoxon test that compares that run with the candidate optimal run, obtained with 30

terms.

| Expansion term count | R-P | P@10 | MAP |
|---|---|---|---|
| **10** | 0.415 (0.149) | 0.716 (0.266) | 0.387 (0.185) |
| | F = 492, p = 0.446 | F = 43, p = 0.000 | F = 462, p = 0.090 |
| **20** | 0.427 (0.137) | 0.778 (0.245) | 0.409 (0.172) |
| | F = 470, p = 0.971 | F = 51.5, p = 0.075 | F = 590, p = 0.647 |
| **30** | **0.428 (0.132)** | **0.798 (0.242)** | **0.411 (0.168)** |
| **50** | 0.427 (0.132) | 0.798 (0.239) | 0.409 (0.167) |
| | F = 450.5, p = 0.604 | F = 83.5, p = 0.642 | F = 591.5, p = 0.657 |
| **75** | 0.422 (0.135) | 0.816 (0.231) | 0.404 (0.168) |
| | F = 435, p = 0.484 | F = 169, p = 0.583 | F = 526.5, p = 0.284 |
| **100** | 0.415 (0.138) | 0.810 (0.235) | 0.395 (0.168) |
| | F = 413, p = 0.164 | F = 124.5, p = 0.753 | F = 474, p = 0.114 |
| **150** | 0.406 (0.141) | 0.810 (0.220) | 0.381 (0.172) |
| | F = 375, p = 0.045 | F = 192, p = 0.942 | F = 395, p = 0.019 |
| **200** | 0.398 (0.145) | 0.806 (0.227) | 0.369 (0.177) |
| | F = 355, p = 0.027 | F = 208, p = 0.909 | F = 345, p = 0.005 |
| **10000** | 0.238 (0.157) | 0.550 (0.298) | 0.180 (0.159) |
| | F = 34, p = 0.000 | F = 136.5, p = 0.000 | F = 24, p = 0.000 |

It is clear that our candidate run is the best in terms of MAP and R-P and close to the top (compared to the best run, the difference is not statistically significant) in terms of P@10. Therefore, we adopted this run as the optimal run, and refer to it as such in the rest of the paper.

Table 2. Comparison between the baseline and the optimal run

| | R-PM (sd) | P@10M (sd) | MAPM (sd) |
|---|---|---|---|
| **Baseline run** | 0.222 (0.158) | 0.338 (0.281) | 0.156 (0.150) |
| **Optimal run** | 0.428 (0.132) | 0.798 (0.242) | 0.411 (0.168) |
| | F = 35, p < 0.001 | F = 19, p < 0.001 | F = 14, p < 0.001 |

The optimal run is significantly better than our baseline run on all three measures of effectiveness used in HARD TREC, as depicted in Table 2. It is particularly relevant to observe the high precision of the optimal run, beneficial for interactive retrieval sessions, simulated in our experiment.

### *Optimal Run Versus Approximations*

Another interesting conclusion that can be drawn from Figure 3 is that, while effectiveness decreases with the size of the relevance judgment sample, the decrease is not substantial: even when only 10% of the relevant documents are used for feedback, we obtain close to optimal retrieval effectiveness. In other words, good performance can be obtained "on the cheap", by using just a relatively low number of positive judgments. A natural question that arises is whether it is possible to manage with no judgments at all, simply relying on the pool of top ranking documents known to be retrieved by a number of search engines. We can simulate that approach by considering NIST's file of relevance judgments, but ignoring the actual judgments and simply using all the documents in the file for relevance feedback.

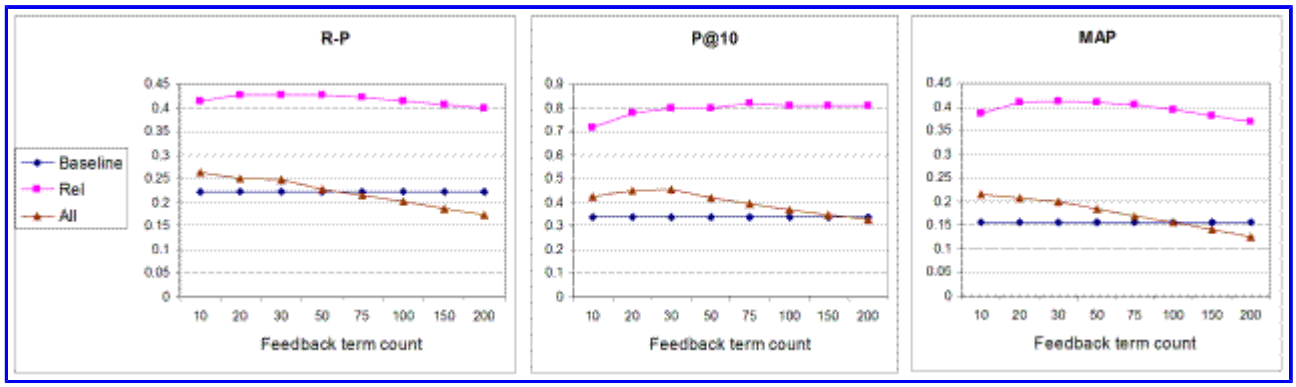


Figure 4. Comparing the baseline, the optimal run, and the cheap run

Figure 4 depicts the results, which are extremely interesting. The baseline represents the retrieval effectiveness based on the human searcher articulating an information need. It can be argued that this is a rather generous baseline: in practice searchers submit queries much shorter than the topic title and description generated by NIST assessors for the TREC experiment. The results labeled "All" simulate a cheap form of implicit relevance feedback in which the human user's searches are logged and all the top ranking documents are recorded and analyzed statistically, whether the searcher opens them and uses them or not. In other words, this is a very naïve form of implicit relevance feedback, in which no attempt is made to interpret user's actions such as opening, bookmarking, saving or printing documents; a document is viewed as possibly relevant based on the fact that it was ranked highly in a search performed by the user. The results labeled "Rel" simulates an explicit form of relevance feedback, in which the user has marked each document retrieved in the past as relevant or non-relevant, so that a more precise models of the topics of interest to the use can be built.

Not surprisingly, "Rel" performs much better than the baseline and than "All". What is very

interesting and extremely promising for research in personalization of IR, is the fact that "All" is better than the baseline. It means that an *intelligent agent* that logs the user's searches and records the retrieved documents, can generate better topic representations and obtain better results than the user submitting queries, after a reasonable amount of training (the TREC QRELS file contains several hundred documents for each topic). If the agent learns to interpret user actions that indicate document relevance, the potential for improving performance is enormous.

Table 3 gives a more refined view of the same result: it captures the means and standard deviations of the three measures of effectiveness (R-P, P@10 and MAP) for runs obtained based on queries consisting of the best 10, 20, 30 , ..., 200 terms representing the documents judged relevant (label "Rel"), respectively all the high-ranking documents retrieved, whether they are relevant or not (label "All"). The means of the effectiveness values and the output of the matched-pairs Wilcoxon tests indicate that the Rel runs constitute a substantial improvement, the All runs are better overall better than the baseline, but this is not a statistically significant conclusion.

Table 3. Comparing the baseline, the optimal run, and the cheap run

| Feedback term count | All | | | Rel | | |
|---|---|---|---|---|---|---|
| | R-P | P@10 | MAP | R-P | P@10 | MAP |
| 10 | 0.263 (0.171) | 0.422 (0.331) | 0.215 (0.189) | 0.415 (0.149) | 0.716 (0.266) | 0.387 (0.185) |
| | F = 775, p = 0.055 | F = 529, p = 0.052 | F = 913, p = 0.008 | F = 1219, p = 0.000 | F = 1010, p = 0.000 | F = 1251, p = 0.000 |
| 20 | 0.251 (0.167) | 0.450 (0.319) | 0.207 (0.184) | 0.427 (0.137) | 0.778 (0.245) | 0.409 (0.172) |
| | F = 701, p = 0.246 | F = 645, p = 0.038 | F = 867.5, p = 0.026 | F = 1234, p = 0.000 | F = 1026.5, p = 0.000 | F = 1261, p = 0.000 |
| 30 | 0.247 (0.166) | 0.454 (0.347) | 0.198 (0.186) | 0.428 (0.132) | 0.798 (0.242) | 0.411 (0.168) |
| | F = 639, p = 0.282 | F = 696, p = 0.019 | F = 837, p = 0.054 | F = 1190, p = 0.000 | F = 1157, p = 0.000 | F = 1261, p = 0.000 |
| 50 | 0.228 (0.168) | 0.420 (0.353) | 0.184 (0.189) | 0.427 (0.132) | 0.798 (0.239) | 0.409 (0.167) |
| | F = 572.5, p = 0.928 | F = 639, p = 0.093 | F = 756, p = 0.253 | F = 1182, p = 0.000 | F = 1022, p = 0.000 | F = 1266, p = 0.000 |
| 75 | 0.214 (0.163) | 0.394 (0.354) | 0.168 (0.182) | 0.422 (0.135) | 0.816 (0.231) | 0.404 (0.168) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | F = 513, p = 0.442 | F = 512, p = 0.291 | F = 664, p = 0.798 | F = 1215, p = 0.000 | F = 1023.5, p = 0.000 | F = 1263, p = 0.000 |
| **100** | 0.202 (0.164) | 0.370 (0.348) | 0.156 (0.175) | 0.415 (0.138) | 0.810 (0.235) | 0.395 (0.168) |
| | F = 430, p = 0.156 | F = 490.5, p = 0.625 | F = 587, p = 0.626 | F = 1157, p = 0.000 | F = 1063, p = 0.000 | F = 1256, p = 0.000 |
| **150** | 0.188 (0.159) | 0.346 (0.343) | 0.140 (0.168) | 0.406 (0.141) | 0.810 (0.220) | 0.381 (0.172) |
| | F = 355, p = 0.027 | F = 422, p = 0.872 | F = 462, p = 0.134 | F = 1135, p = 0.000 | F = 1024, p = 0.000 | F = 1241, p = 0.000 |
| **200** | 0.174 (0.156) | 0.326 (0.341) | 0.126 (0.159) | 0.398 (0.145) | 0.806 (0.227) | 0.369 (0.177) |
| | F = 296, p = 0.005 | F = 404.5, p = 0.736 | F = 398, p = 0.021 | F = 1154, p = 0.000 | F = 1017, p = 0.000 | F = 1219, p = 0.000 |
| **Baseline** | **0.222 (0.158)** | **0.338 (0.281)** | **0.156 (0.150)** | **0.222 (0.158)** | **0.338 (0.281)** | **0.156 (0.150)** |

### *Overlap With The Optimal Topic Representation*

The intuition is that, if an optimal term-based representation of a topic exists and can be built, then the quality of any other term-based representation can be estimated based on the overlap with the optimal set. Of course, one limitation of this method is the fact that term weights are ignored. We plan to extend the method in the future so that correlations of weights are also considered.

Table 4. Overlap between optimal term representation and sample representations

| | **Top 10 terms** | | | **Top 20 terms** | | | **Top 30 terms** | | |
|---|---|---|---|---|---|---|---|---|---|
| **100% vs. 75%** | 100% | 75% | Overlap | 100% | 75% | Overlap | 100% | 75% | Overlap |
| | 10 | 10 | 9.12 | 20 | 20 | 18.14 | 30 | 30 | 26.98 |
| **100% vs. 50%** | 100% | 50% | Overlap | 100% | 50% | Overlap | 100% | 50% | Overlap |
| | 10 | 10 | 8.68 | 20 | 20 | 17.12 | 30 | 30 | 25.06 |
| **100% vs. 25%** | 100% | 25% | Overlap | 100% | 25% | Overlap | 100% | 25% | Overlap |
| | 10 | 10 | 7.54 | 20 | 20 | 15.24 | 30 | 30 | 22.52 |
| **100% vs. 10%** | 100% | 10% | Overlap | 100% | 10% | Overlap | 100% | 10% | Overlap |
| | 10 | 9.8 | 6.02 | 20 | 19.6 | 11.86 | 30 | 29.4 | 17.54 |

Table 4 depicts the overlap, in terms of the most representative 10, 20 and 30 terms, between the topic representation obtained from all the documents judged relevant (the optimal representation, corresponding to the optimal run), and topic representations obtained from random samples of 75%, 50%, 25% and 10% of the documents judged relevant. The overlap values indicated are averaged over the 50 topics. For example, if only half (50%) of the documents judged relevant are considered, the topic representation still captures an average of 8.68 terms from the top 10 terms of the optimal representation, 17.12 terms from the top 20 and 26.06 terms from the top 30. Even if only 10% of the relevant judgments are used for building topic models, 6.02 of the top 10 optimal terms, 11.86 of the top 20 and 17.54 of the top 30 are captured. (Note that for this situation at least one topic is represented by fewer than 10 terms.) These high numbers correlate with high retrieval effectiveness values, and support the hypothesis that there are indeed sets (or rather ranked lists) of terms that are optimal in terms of representing each topic.

Let us also look at the topic representation obtained the “cheap” way, based on the pool of top ranking documents, with disregard to any relevance judgments. The results shown in the next table, also correlated with the effectiveness results, are very encouraging: the set of top ranking documents, even without relevance judgments, does a very good job of representing the searcher’s topic of interest.

Table 5. Overlap between topic representations obtained from all relevant documents vs. all judged documents

| Top 10 terms | | | Top 20 terms | | | Top 30 terms | | |
|---|---|---|---|---|---|---|---|---|
| Rel | All | Overlap | Rel | All | Overlap | Rel | All | Overlap |
| 10 | 10 | 6.3 | 20 | 20 | 11.94 | 30 | 30 | 17.82 |

The overlap is comparable to that obtained when a sample of 10% of the relevant documents is used for relevance feedback, which is encouraging. On the other hand, considering that the retrieval effectiveness is much lower, it appears that the non-relevant documents also yield terms that have no relation to the topic, which have a strong negative impact on search performance.

## Original TREC Topic Representation

In IR experiments researchers typically choose a reasonable baseline and then apply various effectiveness-improvement techniques to verify if these techniques yield

significantly better performance than the baseline. It is common in TREC experiments to select a combination of topic title and description to derive a query that will be submitted to the search engine; that is indeed the approach chosen for our baseline.

In the context of this paper it is suitable to ask ourselves whether other topic representations would produce significantly different results. Table 6 captures the comparison in terms of *R-P*, *P@10* and *MAP* between runs obtained from different combinations of topic representations. The identifier of the run suggests the source of the query terms (e.g. "title.title.description" indicates that title terms were included twice and the description terms once), and the "noDup" particle indicate that duplicate words were removed. Apart from mean and standard deviation of each effectiveness measure, we are reporting the result of matched-pairs Wilcoxon tests comparing the baseline with each other run. The results are interesting, and informative for choosing baselines in future experiments.

Using the title gives the best means for the performance measures. However, the variance is also the highest, so the improvement over the baseline is not statistically significant. This indicates that the title on its own is a risky choice, which can probably be attributed to the fact that a small number of words can be ambiguous and may not clearly convey a topic. The best choice may be the "title.title.description" combination, which is consistently and significantly better than the baseline, probably because the description provides some context, while the title terms, with double weight, indicate the focus of the topic. On its own, the description consistently yields significantly inferior performance but, as we have seen, it can add value to the title in a combination. This is consistent with results obtained in dissertation work by Muresan (2002) and Harper (1980).

Table 6. Comparison of various topic representations retrieval effectiveness

| | R-P | P@10 | MAP |
|---|---|---|---|
| **title.description** | 0.222 (158) | 0.338 (0.281) | 0.156 (0.150) |
| **title.description_noDup** | 0.176 (0.149) | 0.286 (0.272) | 0.123 (0.134) |
| | F = 58.5, p < 0.001 | F = 80.5, p = 0.015 | F = 189, p < 0.001 |
| **title** | 0.236 (0.163) | 0.402 (0.313) | 0.178 (0.162) |
| | F = 457, p = 0.73 | F = 418, p = 0.09 | F = 738, p = 0.33 |
| **title_noDup** | 0.236 (0.163) | 0.402 (0.313) | 0.178 (0.162) |
| | F = 457, p = 0.73 | F = 418, p = 0.09 | F = 738, p = 0.33 |
| **description** | 0.183 (0154) | 0.286 (0.277) | 0.129 (0.138) |
| | F = 54, p < 0.001 | F = 70, p = 0.012 | F = 173, p < 0.001 |

| | | | |
|---|---|---|---|
| **description_noDup** | 0.177 (0.149) | 0.272 (0.261) | 0.123 (0.133) |
| | F = 65, p < 0.001 | F = 49, P = 0.004 | F = 167, p < 0.001 |
| **title.title.description** | 0.228 (0.155) | 0.376 (0.295) | 0.1645 (0.151) |
| | F = 43, p = 0.02 | F = 126, p = 0.019 | F = 934, p = 0.004 |
| **title.title.description_noDup** | 0.176 (0.149) | 0.286 (0.272) | 0.123 (0.134) |
| | F = 58.5, p < 0.001 | F = 80.5, p = 0.015 | F = 189, p < 0.001 |

It is obvious that removing duplicate terms is consistently and significantly detrimental to performance. Therefore, it can be expected that searchers using natural language queries have the potential to get better results than by typing Google-type queries. To state the obvious, “title.title.description_noDup” and “title.description_noDup” are identical; it is also clear that topic titles tend not to have duplicate terms, unlike their descriptions.

Table 7. Overlap between original topic representation and qrels-based term sets

| Count of top terms | Original topic representation | Optimal representation | Overlap |
|---|---|---|---|
| 10 | title | qrels_10 | title & qrels_10 |
| | 2.5 | 10 | 1.08 |
| 20 | title | qrels_20 | title & qrels_20 |
| | 2.5 | 20 | 1.2 |
| 30 | title | qrels_30 | title & qrels_30 |
| | 2.5 | 30 | 1.28 |
| 10 | description | qrels_10 | description & qrels_10 |
| | 8.76 | 10 | 1.38 |
| 20 | description | qrels_20 | description & qrels_20 |
| | 8.76 | 20 | 1.72 |
| 30 | description | qrels_30 | description & qrels_30 |
| | 8.76 | 30 | 1.92 |
| 10 | title.description | qrels_10 | title.description & qrels_10 |
| | 9.06 | 10 | 1.46 |
| 20 | title.description | qrels_20 | title.description & qrels_20 |
| | 9.06 | 20 | 1.8 |
| 30 | title.description | qrels_30 | title.description & qrels_30 |
| | 9.06 | 30 | 2 |

It is interesting to look at the overlap (i.e. number of common terms) between the various topic representations and the optimal representations, in Table 7. The topic title, of average length 2.5, tends to specify, on average, just over one “optimal term”; the description is also of rather low quality, which explains the effectiveness results discussed above.

The topics were generated by NIST assessors, former intelligence analysts. If trained human analysts are not able to generate better expression of their topic of interest, then no better performance can be expected from the average user, especially one that does not have the patience to carefully consider which terms would best convey their information need. The fact that the qrels-based approaches give much better performance suggests that alternative ways to specify one’s information need should be considered. The ostensive model, proposed by Campbell (1996), or the mediated retrieval model,

proposed by Muresan and Harper (2004), could be viable alternatives especially for exploratory searches, when the user does not quite know what she wants, which makes the query formulation more difficult. These interactive models rely on the information seeker to explore a collection of documents and to point to documents that are interesting; the system then builds a topic model and retrieves "more like this" documents.

## Pseudo Relevance Feedback

We obtained pseudo relevance feedback (PRF) runs with Lemur's RetEval function. As Table 5 shows, we used a number of topic representations, by combining titles and descriptions, and employed two different sets of parameters, (5, 10) and (10, 20), where the first number, *feedbackDocCount* indicates how many top ranking documents should be assumed relevant and the second, *feedbackTermCount*, indicates how many terms should be used for topic representation and query expansion. A more systematic approach that focuses solely on pseudo relevance feedback would consider more combinations of parameters. Our limited set is still sufficient to provide better understanding of what is happening.

Table 8. Comparison of pseudo relevance feedback runs with the baseline

| | | R-P | P@10 | MAP |
|---|---|---|---|---|
| **Baseline** | **title.description** | 0.222 (0.158) | 0.338 (0.281) | 0.156 (0.150) |
| | **FB$10$20_description** | 0.211 (0.190) | 0.344 (0.344) | 0.168 (0.183) |
| | | F = 410.5, p = 0.227 | F = 210.5, p = 0.605 | F = 670, p = 0.754 |
| | **FB$10$20_title** | 0.249 (0.194) | 0.424 (0.357) | 0.209 (0.200) |
| | | F = 683, p = 0.208 | F = 443, p = 0.036 | F = 887, p = 0.016 |
| | **FB$10$20_title.description** | 0.238 (0.186) | 0.382 (0.337) | 0.192 (0.192) |
| | | F = 635, p = 0.050 | F = 194.5, p = 0.027 | F = 1012, p < 0.001 |
| | **FB$10$20_title.title.description** | 0.248 (0.182) | 0.408 (0.333) | 0.202 (0.192) |

| | | | | |
|---|---|---|---|---|
| 10feedbackTermCount = 20 | | F = 716.5, p = 0.003 | F = 324, p = 0.006 | F = 1074, p = 0.000 |
| feedbackDocCount = 5feedbackTermCount = 10 | FB$5$10_description | 0.198 (0.179) | 0.370 (0.364) | 0.160 (0.178) |
| | | F = 248, p = 0.011 | F = 345.5, p = 0.245 | F = 540, p = 0.347 |
| | FB$5$10_title | 0.252 (0.190) | 0.466 (0.361) | 0.213 (0.196) |
| | | F = 667, p = 0.276 | F = 514.5, p = 0.004 | F = 872, p = 0.024 |
| | FB$5$10_title.description | 0.234 (0.180) | 0.404 (0.356) | 0.191 (0.190) |
| | | F = 468, p = 0.079 | F = 296, p = 0.010 | F = 981, p = 0.001 |
| | FB$5$10_title.title.description | 0.242 (0.177) | 0.416 (0.347) | 0.197 (0.188) |
| | | F = 617.5, p = 0.015 | F = 296.5, p = 0.010 | F = 1061, p < 0.001 |

The results of the matched-pair Wilcoxon analysis (*F, p*) refer to comparisons between the "title.description" baseline and each other run. Strictly speaking, if we are interested in the effect of pseudo-relevance feedback, the test is only relevant for the PRF runs based on title and description; however, for other cases the results are still informative, even if not so rigorous. It is apparent that PRF is a reliable technique that is consistent in improving performance for most queries; when applying PRF to our baseline, the improvement is indeed statistically significant, even if only by a relatively small margin.

It is clear that for the precision-oriented measure, *P@10*, using a smaller number of documents and terms is beneficial. For *R-P* and *MAP*, using more relevance evidence is better, with the exception of the case when just the title terms are used from the original query. This corroborates with the results from the previous section, indicating that the title can be ambiguous, so documents that string-match the title may in fact not be relevant.

Table 9. Overlap between prf-based and qrels-based expansion term sets

| | | | | |
|---|---|---|---|---|
| | | 10 | 10 | 3.16 |
| | 20 | prf_title_20 | qrels_20 | prf_title_20 & qrels_20 |
| | | 20 | 20 | 5.96 |
| | 30 | prf_title_30 | qrels_30 | prf_title_30 & qrels_30 |
| | | 30 | 30 | 8.92 |
| description | 10 | prf_description_10 | qrels_10 | prf_description_10 & qrels_10 |
| | | 10 | 10 | 2.48 |
| | 20 | prf_description_20 | qrels_20 | prf_description_20 & qrels_20 |
| | | 20 | 20 | 4.34 |
| | 30 | prf_description_30 | qrels_30 | prf_description_30 & qrels_30 |
| | | 30 | 30 | 5.82 |
| title.description | 10 | prf_title.description_10 | qrels_10 | prf_title.description_10 & qrels_10 |
| | | 10 | 10 | 2.3 |
| | 20 | prf_title.description_20 | qrels_20 | prf_title.description_20 & qrels_20 |
| | | 20 | 20 | 4.12 |
| | 30 | prf_title.description_30 | qrels_30 | prf_title.description_30 & qrels_30 |
| | | 30 | 30 | 5.74 |

Table 9 depicts the overlap between PRF-based expansion terms generated by Lemur, and the optimal term sets, for different topic representations and at different size cutoffs. Overall, the results are relatively good, especially when compared to the other tables in this paper. The best term overlap between expansion term sets happens when only the title is used for PRF, which correlates with the effectiveness results from Table 7. Note that overlap is computed based on set intersection, so duplicate terms are removed. Therefore, the overlap for “title.description” actually corresponds to the effectiveness level for “title.description_noDup”, which is inferior to that for description-only. It is apparent that the rank correlation between effectiveness, on the one side, and term overlap with the optimal sets, on the other side, is remarkable.

### Interactive Elicitations

In our HARD TREC 2005 work we were inspired by Rutgers work in the Interactive tracks of TREC (Belkin, et al., 2002, 2003) and UNC work in HARD TREC 2004 (Kelly et al, 2004, 2005) to use a clarification form (CF) and present the searcher with three specific requests for additional terms:

1. "Describe what you already know about this topic"
2. "What sort of information would you like to have as a result of this search?"
3. "Please input any additional keywords that describe your topic."

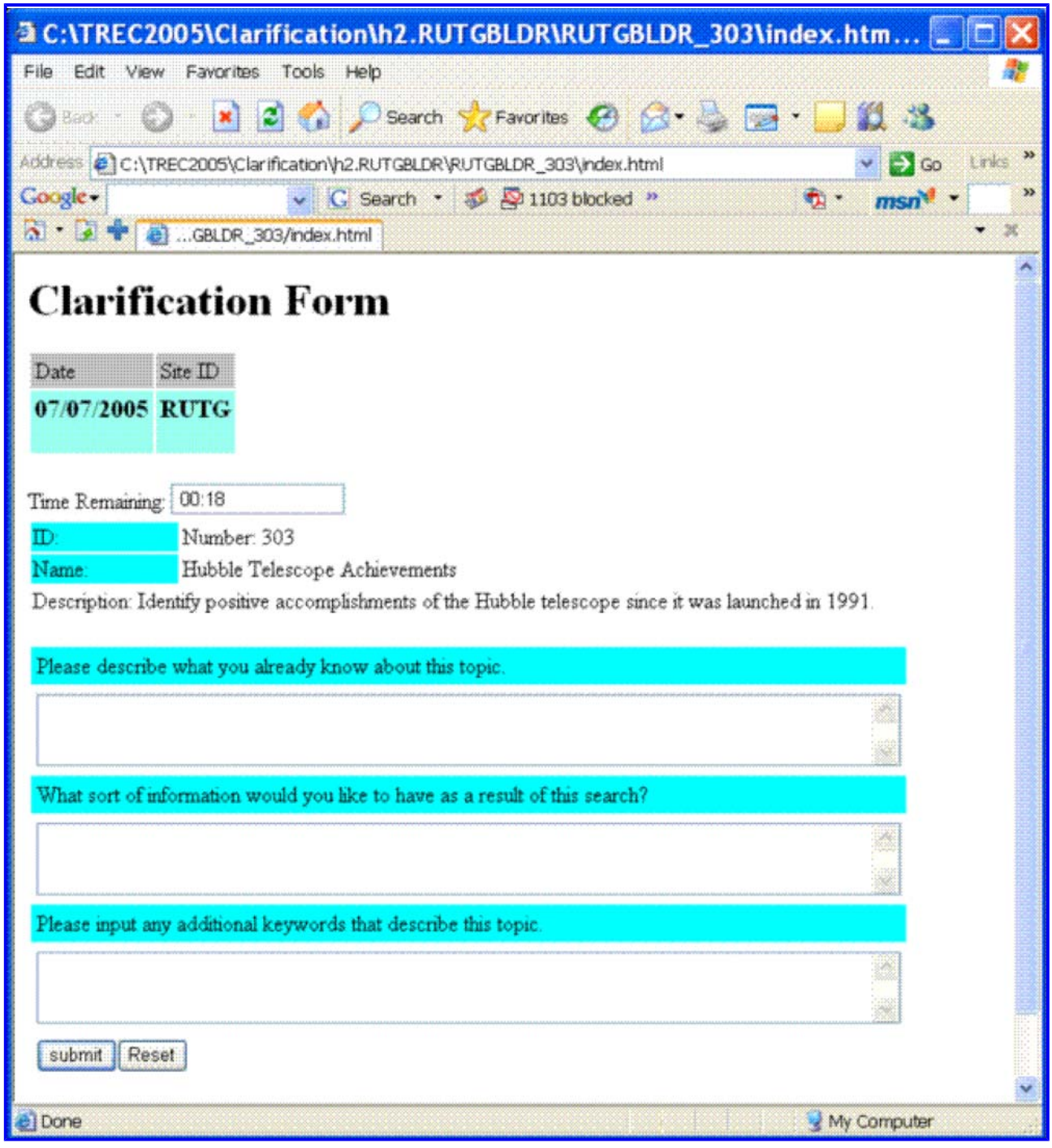

Figure 5. Clarification form for eliciting additional information

The first two questions are related to the ASK model (Belkin, et al, 1982), which states that, especially in assigned or exploratory searches, information seekers are better able to describe what they already know than to describe their information need. The third question is expected to be more useful when the searcher is relatively knowledgeable of the problem domain. We were curious to see which kind of question provides more useful expansion terms. On the one hand, the topics were assigned to the NIST assessors (they are topics created in previous years). On the other hand, due to the HARD TREC setting, the assessors answered clarification forms from multiple research groups. If they answered our questions on a certain topic after having dealt with other participant's forms, then one can expect them to have developed some familiarity with the topic, or at least with the terminology associated with the topic.

The first and third questions are identical to UNC's 2004 study, while the second is our replacement to their "Why do you want to know about this topic?". The difference is that in 2004 the NIST assessors generated their own topics; in 2005, that question was not appropriate. Our replacement was actually from Belkin's original ASK study.

In the UNC study, the first question produced more terms than the other questions (30.98 vs. 23.11 and 2.47 terms respectively) and was the best at improving the baseline. However, the authors raised the issue that the order of the questions may have affected that result. To address the issue, we rotated the three questions before submitting them in the clarification form.

Table 10. The effectiveness of query expansion based on elicited terms

| Source of query terms | R-P | P@10 | MAP |
|---|---|---|---|
| title.description | 0.222 (0.158) | 0.338 (0.281) | 0.156 (0.150) |
| CF_q123 | 0.095 (0.133) | 0.170 (0.290) | 0.062 (0.111) |
| | F = 110, p = 0.000 | F = 148.5, p = 0.000 | F = 169.5, p = 0.000 |
| title.description.CF_q123 | 0.166 (0.141) | 0.314 (0.330) | 0.121 (0.142) |
| | F = 253, p = 0.005 | F = 290.5, p = 0.504 | F = 391, p = 0.017 |
| title.description.CF_q123_noDup | 0.150 (0.142) | 0.252 (0.309) | 0.101 (0.132) |
| | F = 214, p = 0.001 | F = 222, p = 0.031 | F = 273, p = 0.000 |

It is not the focus of this paper to compare the effect of the different elicitation questions. Therefore, in this analysis we concatenated all the terms derived from answers to all three questions in the clarification forms (and thus the label "CF_q123"), with stopwords

removed.

Table 10 compares the effectiveness of the baseline (the run obtained based on the topics title and description) with three other runs: (i) CF_123, obtained by simply using the elicited terms as queries; (ii) title.description.CF_q123 obtained by concatenating the elicited terms to the baseline query; and (iii) title.description.CF_q123_noDup, obtained as the previous run, but with duplicate terms removed from the query. It is apparent that the outcome of the clarification forms is rather poor: when only the elicited terms are used as queries, the search effectiveness is abysmal. Even when the elicited terms are concatenated to the original queries, used in the baseline, the effectiveness drops significantly (and the drop is more pronounced when duplicate terms are removed). This suggests that the quality of the terms provided by the human searchers in order to clarify their information needs is rather poor.

Table 11. Overlap between query terms elicited via clarification forms and qrels-based terms

| cf | qrels_1000Rel_10 | cf & qrels_1000Rel_10 |
|---|---|---|
| 14.34 | 10 | 1.54 |
| cf | qrels_1000Rel_20 | cf & qrels_1000Rel_20 |
| 14.34 | 20 | 2.08 |
| cf | qrels_1000Rel_30 | cf & qrels_1000Rel_30 |
| 14.34 | 30 | 2.4 |

The overlap results shown in Table 11 provide a reasonable explanation: on average, the user's clarifications match just over two words from the optimal representation. The consequence is that, instead of helping, the extra information actually harms the baseline.

Looking at the actual answers reveals that some users did not expect their words to be used for query expansion; they tried instead to communicate with a presumed intelligent retrieval system, which was capable of formulating the questions. Arguably the best example: asked what she already knew about a certain topic, a searcher answered "Not much".

## Discusions And Conclusions

### *Summary Of Results*

This paper addresses a core question of information retrieval: what are good query terms, and what are good sources of query terms. The results discussed here corroborate with

our HARD TREC results and explain them: the system's interaction with a human information seeker is less likely to produce good query terms, and therefore less likely to achieve retrieval effectiveness superior to that obtained via fully automatic methods. This could be attributed to the human searcher's inability to grasp the statistics of a document collection and to generate good query terms, i.e. terms that are representative for the relevant documents, and also distinguish them from non-relevant documents.

Algorithms are obviously better than humans at doing statistical calculations and at estimating which terms best represent a group of relevant documents. One conclusion could be that the power of the algorithms, and especially of machine-learning procedures, should be harnessed even for highly interactive retrieval systems. This could be done by employing implicit relevance feedback, where the system "observes" behavioral cues that indicate interest in the document being examined, builds mathematical models of the topics of interest to the searcher, and retrieves more documents that match the topic model and the user profile, with the searcher's query just one source of evidence about what the user is interested in finding.

Such approaches dictate a re-evaluation of current interactive models, with more attention given to system based on ostention or on mediated retrieval, which have shown substantial potential. For today's less intelligent retrieval systems, a piece of advice for searchers comes from our results: use natural language queries; they have better chances of success than simple keywords.

Such approaches dictate a re-evaluation of current interactive models, with more attention given to system based on ostention or on mediated retrieval, which have shown substantial potential. For today's less intelligent retrieval systems, a piece of advice for searchers comes from our results: use natural language queries; they have better chances of success than simple keywords.

In terms of evaluating our experimental framework, the results are encouraging. For the limited cases that we have investigated, our hypothesis holds: there appears to be an optimal term-based representation for each topic, and queries that overlap with it tend to yield high retrieval effectiveness.

### *Limitations*

The purpose of our study was to propose and demonstrate a methodology for exploring the quality of query expansion terms. We were inevitably limited in the number of parameter combinations that we could try. For example, we chose the Tfldf retrieval model, which affected not only the search results yielded by Lemur, but also the expansion

terms produced when applying relevance feedback or pseudo relevance feedback. The study can be repeated for other retrieval models such as Okapi (Robertson et al, 1994) or Kullback-Leibler (Manning and Schutze, 1999), and for a variety of different parameters, in order to verify the consistency of our results.

We also limited to 5 and 10 the number of top-ranking documents examined in pseudo relevance feedback, and only looked at the sets of top 10, 20 or 30 query expansion terms (as sorted by their weights). The rational for these decisions is our interest in studying interactive information retrieval, with real human searchers, and in comparing such experiments with simulations of interactions, in order to identify ways in which simulations fail to reflect human behavior. Therefore, we limited ourselves to realistic situations: real searchers may be expected to judge 5-10 documents and to examine and accept or reject up to 30 terms proposed by the system, but no more than that. Nevertheless, it may be useful to repeat the experiments for larger numbers in order to verify if the retrieval effectiveness increases, and also to check the validity of our methodology.

Finally, we need to mention the inherent limitations of a laboratory-type experiment, with searchers simulated by NIST analysts, who provided answers to clarification forms and judged the relevance of the retrieved documents. Without a doubt, the behavior of real users attempting to resolve their own information needs would be rather different, and even the relevance judgments may not match those of the analysts.

### *Future Work*

An important issue to consider in future work is the weights of the terms representing topics, queries, or expansion terms. They have the potential to add a certain level of refinement and precision to topic representations, as they indicate the relative importance of different terms to the topic. When comparing manual and automatic relevance feedback methods, it is apparent that the manual methods are at a disadvantage: human searchers are typically asked to specify additional terms, or to vet terms suggested by the system, but not to weigh the terms. On the other hand, algorithms for generating expansion terms typically weigh those terms. One may therefore argue that the weighting makes an essential difference and that the automatic procedures have an advantage without having to generate better expansion terms. We intend to address this issue and distinguish between the quality of the terms themselves, and the contribution of their weights.

Another refinement of our work, planned for the future, is to distinguish between different

levels of relevance. This can be applied in two ways: (i) when generating topic representations based on relevance feedback, distinguish between highly relevant and somewhat relevant documents; it would be interesting to see if using only the former could improve representation quality and implicitly search performance; (ii) when computing retrieval effectiveness, look at the effectiveness of retrieving highly relevant documents.

Finally, our plan was rather ambitious and we have not quite completed it: we have not computed effectiveness results for all the terms sets generated in our HARD TREC work. We are still to investigate the quality of query expansion based on clarity (Cronen-Townsend et al, 2004) or on using the Web as a corpus for query disambiguation (Roussinov et al, 2005).

## Acknowledgments

We acknowledge the Interactive Information Retrieval research group at Rutgers University (N.J. Belkin, M. Cole, J. Gwizdka, Y.-L. Li, J.-J. Liu, G. Muresan, D. Roussinov, C. L. Smith, A. Taylor, X.-J. Yuan), whose work in HARD TREC 2005 constitutes the starting point of the work described in this paper.

## References

Allen, J. (2004) HARD Track Overview in TREC 2004 - High Accuracy Retrieval from Documents *Proceedings of TREC 2004, Gaithersburg, November 2004*

Allen, J. (2005) HARD Track Overview in TREC 2004 - High Accuracy Retrieval from Documents *Proceedings of TREC 2005, Gaithersburg, November 2005*

Belkin, N.J. (1980) Anomalous states of knowledge as a basis for information retrieval *Canadian Journal of Information Science* 5: 133-143

Belkin, N.J., Cool, C, Head, Jeng, J., Kelly, D., Lin, S. et al. (2000) Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience In: E.M. Voorhees & D.K. Harman (Eds.) *The Eighth Text REtrieval Conference (TREC 8) (pp. 565-576). Washington, D.C.*

Belkin, N. J., Cool, C., Kelly, D., Kim, G., Lee, H.-J., Muresan, G., et al. (2002) Rutgers interactive track at TREC 2002 *Paper presented at the Eleventh Text Retrieval Conference (TREC 2002), Washington, D.C.*

Belkin, N. J., Kelly, D., Lee, H.-J., Li, Y.-L., Muresan, G., Tang, M.-C., et al. (2003) Rutgers' HARD and web interactive track experiences at TREC 2003 *Proceedings of the Twelfth*

*Text Retrieval Conference (TREC 2003)*

Belkin, N.J., Cole, M., Li, Y.-L., Liu, L., Liu, Y.-H., Muresan, G. et al. (2004) Rutgers' HARD Track Experiments at TREC 2004 *Proceedings of TREC 2004, Gaithersburg, November 2004*

Belkin, N.J., Cole, M., Gwizdka, J., Li, Y.-L., Liu, J.-J., Muresan, G. et al. (2005) Rutgers Information Interaction Lab at TREC 2005: Trying HARD *Proceedings of TREC 2005, Gaithersburg, November 2005*

Belkin, N. J., Oddy, R. N., & Brooks, H. (1982) Ask for information retrieval part ii. Results of a design study *Journal of Documentation* 38(3), 145-164

Buckley, C. & Voorhees, E. M. (2005) Retrieval System Evaluation In Voorhees, E. M. and Harman, D. K. (eds) *TREC - Experiment and Evaluation in Information Retrieval* The MIT Press, Cambridge, MA

Campbell, I. (1996) The Ostensive Model of Developing Information Needs *Proceedings of COLIS-96, 2nd International Conference on Conceptions of Library Science*

Cronen-Townsend, S., Zhou, Y. & Croft, W.B. (2004) A Framework for Selective Query Expansion *Proceedings of CIKM 2004, Washington, DC, November 8-13, 2004*

Harper, D. J. (1980) *Relevance Feedback in Document Systems: An Evaluation of Probabilistic Strategies* PhD thesis, Jesus College, Cambridge, UK, February 1980

Ingwerwsen, P. & Järvelin, K. (2004) Information retrieval in contexts In Ingwersen, P., Van Rijsbergen, C. J., Belkin, and Nick, Larsen, B. (eds.). *Information in Context: IRiX:ACM-SIGIR Workshop 2004 Proceedings. Sheffield: Sheffield University, 2004* pp. 6-9 http://ir.dcs.gla.ac.uk/context/

Kelly, D., Dollu, V. D., & Fu, X. (2004) University of North Carolina's HARD track experiments at TREC 2004 *Thirteenth Text Retrieval Conference (TREC 2004)*

Kelly, D., Dollu, V. D., & Fu, X. (2005, August 15-19) The loquacious user: A document-independent source of terms for query expansion *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil*

Koenemann, J., & Belkin, N. J. (1996) A case for interaction: A study of interactive information retrieval behavior and effectiveness *Proceedings of the Human Factors in Computing Systems Conference (CHI'96)* ACM Press, New York, 1996

Krovetz, R. (1993) Viewing morphology as an inference process *Proc. 16th ACM SIGIR*

*Conference, Pittsburgh, June 27-July 1, 1993* pp. 191-202

Manning, C. D. & Schutze, H. (1999) *Foundations of Statistical Natural Language Processing* The MIT Press, Cambridge, MA

Muresan, G. (2002) *Using Document Clustering and Language Modelling in Mediated Information Retrieval* PhD thesis, Robert Gordon University, Aberdeen, UK, January 2002

Muresan, G. & Harper, D.J. (2004) Topic Modelling for Mediated Access to Very Large Document Collections *JASIST* 55 (10): 892 - 910

Robertson, S. E., Walker, S., Hancock-Beaulieu, M .M. & Gatford, M. (1994) Okapi at TREC-3, *Proceedings of the Third Text Retrieval Conference, November 1994*

Rocchio, J. J. (1971) Relevance Feedback in Information Retrieval In Salton, G., (ed.) *The SMART retrieval system* Prentice Hall, 1971

Roussinov, D., Zhao, L., & Fan, W. (2005) Mining Context Specific Similarity Relationships Using The World Wide Web *Proceedings of 2005 Conference on Human Language Technologies*

Salton, G. & Buckley, C. (1988) Term-weighting approaches in automatic text retrieval *Information Processing & Management* 24(5): 513-523

Wilkinson, R. (1997) Using combination of evidence for term expansion *Information Retrieval Research - Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, Aberdeen, Scotland, April 1997*

## Appendix

The purpose of this appendix is to provide samples of data used in our experiments, or generated during the experiments, so that the reader has a better understanding of our work. More specifically, we include various representation for topic 303, the first in the set of topics used in HARD TREC 2005.

### *Official Representation*

TREC topics are specified in text format with mark-up that distinguishes the topic number, for identification purposes, the title of the topic, a description of the topic, and a narrative that discusses in some detail what aspect of the topic is of interest, and what kind of documents would be accepted as relevant. Typically the narrative part is used by assessors when judging relevance of candidate documents, and it is the title and description that are used by automatic algorithms to formulate queries and to run

searches based on them. Exemplified below is topic 303, the first in the set of topics adopted for HARD TREC 2005.

<top>
<num> Number: 303

<title>HubbleTelescope Achievements
<desc> Description:
Identify positive accomplishments of the Hubble telescope since it was launched in 1991.
<narr> Narrative:
Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses.
Documents limited to the shortcomings of the telescope would be irrelevant.
Details of repairs or modifications to the telescope without reference to positive achievements would not be relevant.

### *Representations Based On Relevance Judgments*

The top 30 terms derived from the entire set of relevant documents (the optimal representations), and from random samples of it:

100%: hubble telescope astronomer galaxy universe space observation infrared planet light earth faint cosmology observatory star detect astronomic nasa distance gravitational cosmic scientist astronomy image science orbit instrument object dust bright

75%: hubble telescope astronomer galaxy universe space observation planet faint infrared detect earth light gravitational cosmology astronomic star observatory distance science nasa image theoretical astronomy bang scientist guzzardi orbit instrument cosmic

50%: hubble telescope astronomer galaxy universe space nasa observation planet infrared earth light observatory orbit detect scientist star astronomic faint image solar dust instrument distance cosmology cosmic science distant astronomy camera

25%: hubble telescope galaxy astronomer universe space observation astronomic detect observatory light star ngc cosmology infrared earth image orbit object gravitational distance milky radiate astronomy dust science bright study instrument supernova

10%: hubble telescope astronomer galaxy universe infrared space planet faint aeronautics scientist earth astronomic nebulae light star image disk observatory constellate researcher distance nebula solar padgett wavelength science dust invisible peletier

The top 30 terms derived from all judged documents:

hubble telescope astronomer galaxy space nasa universe observatory orbit shuttle earth observation discovery astronomic astronaut mission astronomy instrument light object faint scientist gyroscope infrared planet star science cosmic detect sky

### *Representations Based On Pseudo-Relevance Feedback*

When pseudo relevance feedback was applied on a search based on the baseline query, derived from the topic title and description, the lists of top ranking expansion terms were influenced by the PRF parameters:

- for feedbackDocCount = 5:

spacewalk gyro hubble gyroscope spacewalker telescope grunsfeld nicollier astronaut nasa sensor transmitter foale shuttle astronomer astronomic repair recorder discovery guidance install observation instrument solar observatory weiler mission smith space orbit

- for feedbackDocCount = 10:

hubble telescope gyroscope astronomer spacewalk nasa observatory nicollier spacewalker astronaut astronomic observation foale gyro shuttle universe grunsfeld discovery transmitter space instrument sensor repair weiler recorder mission solar galaxy data install

### *Representations Derived From Elicitations Via Clarification Forms*

Terms elicited from subjects via the three questions in the clarification forms:

Q1: discovery of galaxies nebulae

Q2: sightings deep space objects forming

Q3: infra red space dust