Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 42 (2006) 191–206

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

# Elicitation and use of relevance feedback information

Olga Vechtomova [a,*], Murat Karamuftuoglu [b,1]

[a] Department of Management Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ont., Canada N2L 3GE
[b] Department of Computer Engineering, Bilkent University, Bilkent 06800 Ankara, Turkey

Received 5 April 2004; accepted 21 October 2004
Available online 8 December 2004

**Abstract**

The paper presents two approaches to interactively refining user search formulations and their evaluation in the new High Accuracy Retrieval from Documents (HARD) track of TREC-12. The first method consists of asking the user to select a number of sentences that represent documents. The second method consists of showing to the user a list of noun phrases extracted from the initial document set. Both methods then expand the query based on the user feedback. The TREC results show that one of the methods is an effective means of interactive query expansion and yields significant performance improvements. The paper presents a comparison of the methods and detailed analysis of the evaluation results.

*Keywords:* Information retrieval; Query expansion; Natural language processing; Interactive retrieval; Relevance feedback

## 1. Introduction

The traditional models of query expansion based on relevance feedback (e.g., Beaulieu, 1997; Rocchio, 1971) consist of the following steps: the user reads representations of retrieved documents, typically their full-text or abstracts, and judges them as relevant or non-relevant. After that the system extracts query expansion terms from the relevant documents, and either adds them to the original query automatically (automatic query expansion), or asks the searcher to select terms to be added to the query (interactive query expansion). In this paper we present two approaches to automatic and interactive query expansion based on limited amount of information elicited from the user. The approaches were evaluated in the newly

[*] Corresponding author. Tel.: +1 519 888 4567x2675; fax: +1 519 746 7252.
*E-mail addresses:* ovechtom@engmail.uwaterloo.ca (O. Vechtomova), hmk@cs.bilkent.edu.tr (M. Karamuftuoglu).
[1] Tel.: +90 312 290 2577; fax: +90 312 266 4047.

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.10.006

formed High Accuracy Retrieval from Documents (HARD) track (Allan, 2004) of TREC (Text Retrieval Conference) 2003. One of the approaches proved to be quite successful within the HARD track evaluation framework. The paper presents the details of both approaches, analysis of the HARD TREC results as well as comparison of the best performing systems.

The first approach consists in representing each top-ranked retrieved document by means of one sentence containing the highest proportion of query terms. The documents, whose one-sentence representations were selected by the user, are then used to extract query expansion terms automatically. We developed a new method of query expansion using collocates—words significantly co-occurring in the same contexts with the query terms. A number of automatically selected collocates are then used for query expansion. The second approach consists in presenting to the user a list of noun phrases extracted from the most representative sentences taken from top-ranked documents. The terms from user-selected noun phrases are then used for query expansion. Both approaches aim to minimise the amount of text the user has to read, and to focus the user's attention on the key information clues from the documents.

Traditionally in bibliographical and library IR systems the hitlist of retrieved documents is presented in the form of the titles and/or the first few lead-sentences of each document. Reference to full text of documents is obviously time-consuming, therefore it is important to represent documents in the hitlist in a form that would enable the users to reliably judge their relevance without referring to the full text. Arguably, the title and the first few sentences of the document are frequently not sufficient to make the correct relevance judgement. Query-biased summaries constructed by extracting sentences that contain higher proportion of query terms than the rest of the text may contain more relevance clues than generic document representations. Tombros and Sanderson (1998) compared query-biased summaries with the titles plus the first few lead-sentences of the documents by how many times the users have to request full-text documents to verify their relevance/non-relevance. They discovered that subjects using query-biased summaries refer to the full text of only 1.32% documents, while subjects using titles and first few sentences refer to 23.7% of documents. This suggests that query-biased representations are likely to contain more relevance clues than generic document representations. White, Jose, and Ruthven (2003) compared one-sentence representation of documents in the retrieved set to the representations used by the Google search engine. Sentences were selected on the basis of a number of parameters, including position of the sentence in document, the presence of any emphasized words and the proportion of query terms they contain. They conducted interactive experiments with users in live settings, measuring the search task completion time, user satisfaction with the representations and user perception of task success. The results indicate that both experienced and inexperienced users found one-sentence representations significantly more useful and effective. Dziadosz and Chandrasekar (2002) also investigated the effectiveness of displaying thumbnail screenshots of the retrieved webpages along with the short text summaries of their content. They also did the evaluation with users in live settings. Their findings suggest that the use of thumbnails along with text summaries helps users in predicting the document relevance with higher degree of accuracy than using only the summaries.

The above studies of document representations were focused mainly on measuring the user-related characteristics of the search process, such as user satisfaction with the document representations, their perception of the search task completion and task completion time. However, they did not measure the search effectiveness using the traditional IR metrics of recall and precision. It is difficult to apply these measures in interactive IR experiments, due to the necessity of obtaining a large number of relevance judgements from the users. We contribute to this research area by evaluating the developed document representation and query expansion techniques by the traditional measures of recall and precision using the HARD track evaluation framework.

The rest of the paper is organised as follows: In the next section we introduce the HARD track of TREC 2003. In Sections 3 and 4 the two methods evaluated in TREC are described in detail. Section 5 presents a detailed analysis and comparison of the results obtained in the HARD track by the two methods. A brief description of alternative approaches used by other participating sites and the comparison of their results in relation to our results are presented in Section 6. The final section summarises the main points of the paper and draws conclusions about possible improvements to the approaches presented.

## 2. HARD Track

The primary goal of our participation in the HARD track was to investigate how to improve retrieval precision through limited amount of interaction with the user. The new HARD track in TREC-12 facilities the exploration of the above question by means of a two-pass retrieval process. In the first pass each site was required to submit one or more baseline runs—runs using only the data from traditional TREC topic fields (title, description and narrative). In the second pass the participating sites may submit one or more clarification forms per topic with some restrictions: Each clarification form must fit into a screen with 1152 × 900 pixels resolution, and the user (annotator) may spend no more than 3 min filling out each form.

Each site then submits one or more final runs, which would make use of the user's feedback to clarification forms, and/or make use of any of the metadata that comes with each topic. The metadata in HARD track 2003 consisted of extra-linguistic contextual information about the user and the information need, which was provided by the user who formulated the topic. It specifies the following:

- *Genre*—the type of documents that the searcher is looking for. It has the following values:
  - Overview (general news related to the topic);
  - Reaction (news commentary on the topic);
  - I-Reaction (as above, but about non-US commentary);
  - Any.
- *Purpose* of the user's search, which has one of the following values:
  - Background (the searcher is interested in the background information for the topic);
  - Details (the searcher is interested in the details of the topic);
  - Answer (the searcher wants to know the answer to a specific question);
  - Any.
- *Familiarity* of the user with the topic on a five-point scale.
- *Granularity*—the amount of text the user is expecting in response to the query. It has the following values: Document, Passage, Sentence, Phrase, Any.
- *Related text*—sample relevant text found by the users from any source, except the evaluation corpus.

An example of a HARD track topic is shown in Table 1.

The evaluation corpus used in the HARD track consists of 372,219 documents, and includes three newswire corpora (New York Times, Associated Press Worldstream and Xinghua English) and two governmental corpora (The Congressional Record and Federal Register). The overall size of the corpus is 1.7 Gb.

Table 1
Example of a HARD track topic

| Title | Red Cross activities |
|---|---|
| *Description* | What has been the Red Cross's international role in the last year? |
| *Narrative* | Articles concerning the Red Cross's activities around the globe are on topic. Has the RC's role changed? Information restricted to international relief efforts that do not include the RC are off-topic |
| *Purpose* | Details |
| *Genre* | Overview |
| *Granularity* | Sentence |
| *Familiarity* | 2 |

The users (assessors) invited by the track organizers formulated altogether 50 topics. The same assessor who formulated the topic filled out the clarification forms corresponding to the topic and did the document relevance judgements. Two runs per site (one baseline and one final run) were judged by the assessors as follows: top 75 documents, retrieved for each topic in each of these runs were pooled together, and allocated to the assessor who formulated the topic. The assessor then assigned binary relevance judgements to the documents.

Our main aim in HARD track 2003 was to study the ways of improving retrieval performance through limited amount of information elicited by means of the clarification forms. We did not make extensive use of the metadata available other than ‘‘granularity’’ and ‘‘related text’’ metadata categories.

## 3. Query expansion method 1

The method consists in building document representations consisting of one sentence, selected on the basis of the query terms it contains; showing them to the user in the clarification form; asking the user to select sentences which possibly represent relevant documents; and finally, using these documents to automatically select query expansion terms. The goal that we aim to achieve with the aid of the clarification form is to have the users judge as many relevant documents as possible on the basis of one sentence per document. The main questions that we explore in this set of experiments are: ‘What is the error rate in selecting relevant documents on the basis of one sentence representation of its content? If it is less than 100%, what is the effect of different numbers of relevant and non-relevant documents in the relevance feedback document set on the performance of query expansion?’

### 3.1. Sentence selection

The sentence selection algorithm consists of the following steps:

We take $N$ top-ranked documents, retrieved using Okapi BM25 (Sparck Jones, 2000) search function in response to query terms from the topic titles. Given the screen space restrictions, we can only display 15 three-line sentences, hence $N = 15$. The full-text of each of the documents is then split into sentences. [2] For every sentence that contains one or more query terms, i.e. any term from the title field of the topic, two scores are calculated: S1 and S2.

Sentence selection score 1 (S1) is the sum of *idf*—inverse document frequency (Sparck Jones, 1972) of all query terms present in the sentence.

$$\mathrm{S1} = \sum idf_q \tag{1}$$

Sentence selection score 2 (S2):

$$\mathrm{S2} = \frac{\sum W_i}{f_s} \tag{2}$$

where $W_i$—weight of the term $i$, see (3); $f_s$—length factor for sentence $s$, see (4).

The weight of each term in the sentence, except stopwords, is calculated as follows:

$$W_i = idf_i\left(0.5 + \left(0.5 * \frac{tf_i}{t\max}\right)\right) \tag{3}$$

where $idf_i$—inverse document frequency of term $i$ in the corpus; $tf_i$—frequency of term $i$ in the document; $t$max—$tf$ of the term with the highest frequency in the document.

[2] We used the sentence splitter provided for the Document Understanding Conference (DUC) 2002 evaluation framework.

To normalise the length of the sentence we introduced the sentence length factor *f*:

$$f_s = \frac{s\max}{slen_s} \tag{4}$$

where *s*max—the length of the longest sentence in the document, measured as a number of terms, excluding stopwords; *slen*—the length of the current sentence.

All sentences in the document were ranked by S1 as the primary score and S2 as the secondary score. Thus, we first select the sentences that contain more query terms, and therefore are more likely to be related to the user's query, and secondarily, from this pool of sentences select the one which is more content-bearing, i.e. containing a higher proportion of terms with high *tf*idf* weights.

Because we are restricted by the screen space, we reject sentences that exceed 250 characters, i.e. three lines. In addition, to avoid displaying very short, and hence insufficiently informative sentences, we reject sentences with less than 6 non-stopwords. If the top-scoring sentence does not satisfy the length criteria, the next sentence in the ranked list is considered to represent the document. Also, since there are a number of almost identical documents in the corpus, we remove the representations of the duplicate documents from the clarification form using pattern matching, and process the necessary number of additional documents from the baseline run sets. Each clarification form, therefore, displays 15 sentences, i.e. one sentence per document. Document titles or any other information about the document was not displayed.

By selecting the sentence with the query terms and the highest proportion of high-weighted terms in the document, we are showing query term instances in their typical context in this document. Typically a term is only used in one sense in the same document. Also, in many cases it is sufficient to establish the linguistic sense of a word by looking at its immediate neighbours in the same sentence or a clause. Based on this, we hypothesise that users will be able to reject those sentences, where the query terms are used in an unrelated linguistic sense.

The TREC assessors were asked to select all sentences which possibly represent relevant documents. The relevance of the full-text documents was determined by the same assessor later at the document judgement stage. We were interested in finding how accurately the users can determine the relevance of the document based on a one-sentence representation of its contents. To answer this question we calculated precision and recall of sentence selection as follows:

$$\text{Precision} = \frac{\text{Relevant selected}}{\text{Selected sentences}}$$

$$\text{Recall} = \frac{\text{Relevant selected}}{\text{Relevant shown}}$$

where *Relevant selected*—the number of sentences, which were selected by the user from the clarification form, and which represent documents judged later relevant by the same user; *Selected sentences*—the number of sentences selected by the user from the clarification form; *Relevant shown*—the number of sentences shown to the user in the clarification form, which represent documents judged later relevant by the same user.

The results show that users selected relevant documents with average precision of 73% and average recall of 69%. Out of 7.14 relevant documents represented on average in the clarification forms, users selected 4.9 relevant documents. And out of 7.86 non-relevant documents represented on average in the clarification forms, users selected 1.8 non-relevant documents. Fig. 1 shows the number of relevant/non-relevant documents by topic. Experiments investigating the effect of different numbers of relevant and non-relevant documents in the relevance feedback document set on the performance of query expansion are described in Section 6.

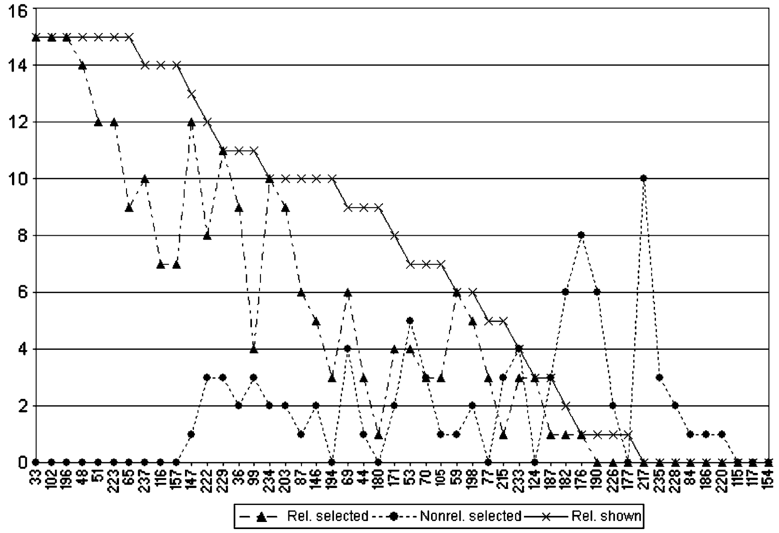

Fig. 1. Sentences selected by users from clarification forms.

### 3.2. Selection of query expansion terms

The user's feedback to the clarification form is used for obtaining query expansion terms for the final run. For query expansion we use collocates of query terms–words co-occurring within a limited span with query terms. Vechtomova, Robertson, and Jones (2003) have demonstrated that expansion with long-span collocates of query terms obtained from 5 known relevant documents showed significant improvement over the use of title-only query terms on the Financial Times corpus with TREC-5 ad hoc topics.

We extract collocates from windows surrounding query term occurrences. The span of the window is measured as the number of sentences to the left and right of the sentence containing the instance of the query term. For example, span 0 means that only terms from the same sentence as the query term are considered as collocates, span 1 means that terms from 1 preceding and 1 following sentences are also considered as collocates.

In more detail the collocate extraction and ranking algorithm is as follows: For each query term we extract all sentences containing its instance, plus $s$ sentences to the left and right of these sentences, where $s$ is the span size. Each sentence is only extracted once. After all required sentences are selected we extract stems from them, discarding stopwords. For each unique stem we calculate the $Z$ score to measure the significance of its co-occurrence with the query term as follows (Vechtomova et al., 2003):

$$Z = \frac{f_r(x,y) - \frac{f_c(y)}{N} f_r(x) v_x(R)}{\sqrt{\frac{f_c(y)}{N} f_r(x) v_x(R)}} \tag{5}$$

where $f_r(x,y)$—frequency of $x$ and $y$ occurring in the same windows in the document set $R$, [3] see (6); $f_c(y)$—frequency of $y$ in the corpus; $f_r(x)$—frequency of $x$ in the document set $R$; $v_x(R)$—average size of windows around $x$ in the document set $R$; $N$—the total number of non-stopword occurrences in the corpus.

[3] Here $R$ is the set of documents, the representative sentences of which were selected by the user from the clarification form.

The frequency of $x$ and $y$ occurring in the same windows in the document set $R$—$f_r(x,y)$—is calculated as follows:

$$f_r(x,y) = \sum_{w=1}^{m} f_w(x) f_w(y) \tag{6}$$

where $m$—number of windows in the set $R$; $f_w(x)$—frequency of $x$ in the window $w$; $f_w(y)$—frequency of $y$ in the window $w$.

All collocates with an insignificant degree of association: $Z < 1.65$ are discarded (see Church, Gale, Hanks, & Hindle, 1991). The remaining collocates are sorted by their $Z$ score.

After we obtain sorted lists of collocates of each query term, we select those collocates for query expansion, which co-occur significantly with two or more query terms. First, for each collocate the collocate score (C1) is calculated:

$$\mathrm{C1} = \sum n_i W_i \tag{7}$$

where $n_i$—rank of the collocate in the z-sorted collocation list for the query term $i$; $W_i$—weight of the query term $i$.

Finally, collocates are ranked by two parameters: the primary parameter is the number of query terms they co-occur with, and the secondary—C1 score. Then, $k$ top-ranked query expansion terms are added to the original query terms extracted from the Title section of the TREC topics, and searched in the TREC database using Okapi BM25 search function (Sparck Jones, Walker, & Robertson, 2000).

The parameters for the above algorithm were experimentally selected using the past TREC data: Financial Times and Los Angeles Times newswire corpora, [4] topics 301–450 [5] with pseudo-relevance (blind) feedback using Okapi BM25 search function. The goal was to determine the optimal values for $R$—the size of the pseudo-relevant set, $s$—the span size, and $k$—the number of query expansion terms. The following values were tried: $R = [5,10,20,30,50]$, $s = [0,1,2,3,4,5]$, $k = [10,20,30,40]$. The results indicate that variations of these parameters have an insignificant effect on precision. However, some tendencies were observed, namely: (1) larger $R$ values tend to lead to poorer performance in both Title-only and Title + Description runs; (2) larger span sizes also tend to degrade performance in both Title and Title + Description runs.

Average precision (AveP) of the Title-only unexpanded run (0.2620) was 10% better than Title + Description (0.2357). Expansion of Title + Description queries resulted in relatively poorer performance than expansion of Title-only queries. For example, AveP (0.1814) of the worst Title + Description expansion run ($R = 50$, $s = 4$, $k = 40$) is 23% worse than the baseline, and AveP (0.2563) of the best run ($R = 5$, $s = 1$, $k = 10$) is 8% better than the baseline. AveP (0.2502) of the worst Title-only run ($R = 50$, $s = 5$, $k = 20$) is 4.5% worse than the baseline, and AveP (0.2940) of the best Title-only run ($R = 5$, $s = 1$, $k = 40$) is 10.9% better than the baseline.

Based on this data we decided to use Title-only terms for the official TREC run '*UWAThard2*', and, given that values $k = 40$ and $s = 1$ contributed to a somewhat better performance, we used these values in all of our official expansion runs. The question of $R$ value is obviously irrelevant here, as we used all documents selected by users in the clarification form.

### 3.3. Use of query expansion terms in searching

The weights of terms in the expanded query were calculated using relevance data according to the BM25 term weighting scheme in the probabilistic model (Sparck Jones et al., 2000), i.e. $R$—the number of

[4] From TREC collection volumes 4 and 5.

[5] From ad hoc tracks of TRECs 6 through 8.

documents, the representative sentences of which were selected by the user from the clarification form, and $r_i$—the number of documents out of $R$, which contain the term $i$. Original query terms were not given any special treatment compared to the query expansion terms, but they were always kept in the expanded query.

Another question is whether documents whose sentences were not selected by the user should be used for query expansion to provide negative evidence against relevance. We did not do any experimentation with negative weighting of terms. Some experiments with negative term weighting were documented in (AbduJaleel et al., 2004; Robertson, Zaragoza, & Taylor, 2004).

We used Okapi BM25 document retrieval function for topics with granularity *Document*, and Okapi BM250 passage retrieval function for topics with other granularity values. For topics with granularity *Sentence* the best sentences were selected from the passages, returned by BM250, using the algorithm described above. The results of evaluation of this method are presented in Sections 5.1 and 5.2.

## 4. Query expansion method 2

The second user feedback mechanism consists of automatically selecting noun phrases from the top-ranked documents retrieved in the baseline run, and asking the users to select all phrases that contain possibly useful query expansion terms.

We take top 25 documents from the baseline run, and select 2 sentences per document using the algorithm described above in Section 3.1. We have not experimented with alternative values for these two parameters. We then apply Brill's rule-based tagger (Brill, 1995) and BaseNP noun phrase chunker (Ramshaw & Marcus, 1995) to extract noun phrases from these sentences. Following the stemming [6] and the removal of the stopwords and phrases consisting entirely of the original query terms, the *idf* value of each term in each phrase is calculated. The phrases are then ranked by the sum of weights of their constituent terms. Top 78 phrases are then included in the clarification form for the user to select. This is the maximum number of phrases that could fit into the clarification form.

All user-selected phrases were split into single terms, which were then added to the original query terms from the topic title. Terms in the expanded query were weighted and used in search in the same way as described in Section 3.3 above. We only used phrases selected by the user, and did not experiment with negative weighting of non-selected phrases. On average assessors selected 19 phrases from clarification forms. The average query size after query expansion (i.e. original terms plus phrase-terms with duplicates eliminated) is 32 words.

An alternative to splitting user-selected phrases and using their components as single terms would be the use of phrases as complete units in search. Some preliminary experiments did not show any improvement. Currently we are working on a new method for phrase search.

## 5. Evaluation

Every run submitted to the HARD track was evaluated in three different ways. The first two evaluations were done at the document level only, whereas the last one took into account the granularity metadata.

1. SOFT-DOC—document-level evaluation, where only the traditional TREC topic formulations (title, description, narrative) were used as relevance criteria.

[6] Porter stemming algorithm was used (Porter, 1980).

2. HARD-DOC—the same as the above, plus ‘purpose’, ‘genre’ and ‘familiarity’ metadata were used as additional relevance criteria.
3. HARD-PSG—passage-level evaluation, which in addition to all criteria in HARD-DOC also required that retrieved items satisfied the granularity metadata.

Document-level evaluation was done by the traditional IR metrics of mean average precision and precision at various document cut-off points. Passage-level evaluation was done using modified passage recall, precision, $F$ score and $R$-precision (Allan, 2004).

### 5.1. Document-level evaluation

The document-level results of the three submitted runs are given in Table 2. UWAThard1 is the baseline run using original query terms from the topic titles. UWAThard2 is an experimental run using query expansion method 1 plus the granularity and known relevant documents metadata. UWAThard3 is an experimental run using query expansion method 2 plus the granularity metadata (Vechtomova, Karamuftuoglu, & Lam, 2004).

UWAThard2 did not achieve statistically significant improvement over the baseline. In addition to clarification forms, we used the ‘*related text*’ metadata for UWAThard2, from which we extracted query expansion terms using the method described in Section 3.2. To determine the effect of this metadata on performance, we conducted a run without it (UWAThard5), which showed only a slight drop in performance. This suggests that additional relevant documents from other sources do not affect performance of this query expansion method significantly.

Another possible reason why there was not a big difference between UWAThard2 and the baseline could be due to the fact that the baseline run used BM25 (document retrieval function) for all topics, whereas UWAThard2 used BM25 for topics with granularity = document, and BM250 (passage retrieval function) for topics with granularity = passage/sentence. Two functions produce different document rankings. An additional run UWAThard4 was conducted as an unofficial baseline run, using BM250 for topics with granularity = passage/sentence. It resulted, however, in only a slightly lower average precision of 0.2937 (SOFT-DOC evaluation) and 0.2450 (HARD-DOC evaluation).

Table 2
Document-level evaluation results[a]

| Run | Run description | SOFT-DOC evaluation | | | HARD-DOC evaluation | | |
|---|---|---|---|---|---|---|---|
| | | P@10 | *R*-Prec. | AveP | P@10 | *R*-Prec. | AveP |
| Baseline, BM25 (UWAThard1) | Original title-only query terms; BM25 used for all topics | 0.4875 | 0.3336 | 0.3134 | 0.3875 | 0.2893 | 0.2638 |
| Baseline, BM25/BM250 (UWAThard4) | As UWAThard1, but BM250 is used for topics requiring passages | 0.4729 | 0.3126 | 0.2937 | 0.3667 | 0.2703 | 0.2450 |
| Sentence expansion, BM25/BM250, Related text (UWAThard2) | Query expansion method 1; granularity and related text metadata | 0.5479 | 0.3417 | 0.3150 | 0.4354 | 0.3263 | 0.2978 |
| Sentence Expansion, BM25/BM250 (UWAThard5) | As UWAThard2, but related text metadata is not used | 0.5229 | 0.3286 | 0.3016 | 0.4062 | 0.3132 | 0.2828 |
| Noun phrase expansion, BM25/BM250 (UWAThard3) | Query expansion method 2; granularity metadata | 0.5958 | 0.3780 | 0.3719 | 0.4854 | 0.3466 | 0.3335 |

[a] UWAThard1, UWAThard2 and UWAThard3 were submitted to TREC. Top 75 documents from UWAThard1 and UWAThard2 were included in the pool of documents judged by assessors.

We compared our sentence-based query expansion method used in UWAThard2 with the standard query expansion technique used in Okapi, where query expansion term candidates are extracted from the entire document (Robertson, 1990), using the same number of query expansion terms (40). The results were very similar (AveP = 0.3037), suggesting that regardless of the specific query expansion method used, automatic query expansion on this collection gives poor results.

Our second experimental run (UWAThard3) performed very well, gaining an 18% improvement over the baseline in average precision in soft-doc evaluation and 26.4% in hard-doc evaluation, both of which are statistically significant (using *t*-test at 0.05 significance level). On average 19 phrases were selected by users per topic.

Comparison with other HARD submissions (88 in total) shows that all our submitted runs are above the median in all evaluation measures shown in Table 3.

### 5.2. Passage-level evaluation

According to passage-level evaluation, a document should satisfy all metadata criteria, including "granularity" as well as be relevant to the topic. Passage-level evaluation results of the runs submitted to TREC are given in Tables 4 and 5. UWAThard3 showed 27% improvement in *R*-precision over UWAThard1, while UWAThard2-23%. Such big difference between the expansion runs and the baseline was expected, since we only did document-level retrieval for the baseline run. All our runs were above the median in all passage-level measures.

Table 3
Statistics of document-level evaluation computed over 88 runs submitted to HARD track by participating sites

| | P@10 | | | *R*-precision | | | AveP | | |
|---|---|---|---|---|---|---|---|---|---|
| | Best | Median | Worst | Best | Median | Worst | Best | Median | Worst |
| SOFT-DOC evaluation | 0.65 | 0.4729 | 0.0417 | 0.4250 | 0.2994 | 0.0038 | 0.4069 | 0.2841 | 0.0026 |
| HARD-DOC evaluation | 0.5271 | 0.3792 | 0.0312 | 0.3875 | 0.2673 | 0.0038 | 0.3604 | 0.2490 | 0.0024 |

Table 4
Number of topics with average precision at, above and below median

| | UWAThard1 | | UWAThard2 | | UWAThard3 | |
|---|---|---|---|---|---|---|
| | HARD-DOC evaluation | SOFT-DOC evaluation | HARD-DOC evaluation | SOFT-DOC evaluation | HARD-DOC evaluation | SOFT-DOC evaluation |
| Best | 0 | 0 | 4 | 2 | 0 | 2 |
| Above median | 32 | 33 | 27 | 27 | 37 | 37 |
| At median | 2 | 1 | 1 | 1 | 2 | 1 |
| Below median | 14 | 14 | 20 | 20 | 9 | 10 |
| Worst | 0 | 0 | 1 | 1 | 0 | 0 |

Table 5
Passage-level evaluation results

| Run | Passage P@10 | *R*-Precision | *F*(30) |
|---|---|---|---|
| **UWAThard1** | 0.2668 | 0.1908 | 0.1255 |
| **UWAThard2** | 0.3305 | 0.2359 | 0.1454 |
| **UWAThard3** | 0.3617 | 0.2426 | 0.1559 |

### 5.3. Analysis of performance by topic

As the second query expansion method (UWAThard3) is more promising, we have conducted a topic-by-topic analysis of its performance in comparison with the baseline. Fig. 2 shows the average precision (SOFT-DOC) of these two runs by topic.

It is not surprising, that performance of query expansion following blind feedback tends to depend on performance of the original query. The correlation between the AveP values of the baseline and UWAThard3 is very strong ($r = 0.9$). This tendency is evident from Fig. 2.

We have analysed three groups of topics: (1) topics, which yielded substantially worse results in runs with the expanded query (UWAThard3) than runs with the original query terms (baseline); (2) topics, which had low performance both with the original and the expanded queries; and finally (3) topics which performed better with the expanded query (UWAThard3) than the original query.

Some examples of topic titles in the first group are: "Corporate mergers" (topic 222), "Sports scandals" (223), "Oscars" (53) and "IPO activity" (196). One factor that all of these topics have in common is that query expansion phrases selected by the users from the candidate phrases shown to them contain a large number of proper names. We evaluated the contribution of each term in these expanded queries by: (a) conducting the search with all expansion terms and (b) conducting the search with all expansion terms except the term being evaluated, and calculating the difference between the average precision values of these runs. Among the terms that affected performance most were many proper names. For example: the 5 expansion terms that most negatively affected performance of the topic "IPO activity" were: 'ABC', 'Disney', 'investment', 'CBS' and 'Viacom'. In the topic "Sports scandals" such terms were: 'ethics', 'Lake', '2002', 'Salt', 'SLOC'. One of the problems could be that in our current model we break user-selected multi-term phrases into their constituent terms and use them in the search process. Therefore terms like 'Salt' and 'Lake' could match unrelated concepts and therefore cause topic drift. Intuitively using complete phrases in search should lead to better performance, however so far our experimentation with phrase search gave inferior results. We continue to work in this direction.

Examples of topic titles in the second group are: "National leadership transitions" (187), "School development" (182), "Virtual defense" (115), "Rewriting Indian history" (177) and "Restricting the Internet"

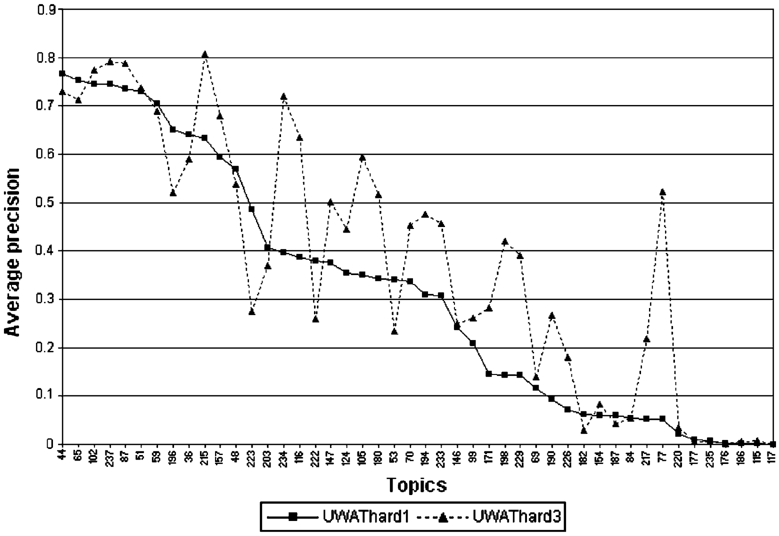

Fig. 2. Results (SOFT-DOC AveP) by topic of the baseline (UWAThard1) and the second query expansion method (UWAThard3).

(186). The majority of terms in these queries have very high number of postings, what suggests that they are either topic-neutral words (e.g., restrict, rewrite, transition), or they represent ideas or entities that were popular in newswire and governmental publications at the time (e.g., Internet, Indian). Moreover, these queries do not represent fixed phrases, i.e., that co-occur frequently in English language. Compare queries in this group to the query "Mad cow disease" (65), which performed very well. Although, the number of postings of individual terms is very high, the query represents a fixed expression, which occurs as a phrase in 213 documents.

Another reason of failure, which applies to both groups above, is over-stemming. We used Porter's stemmer with the strong stemming function in our searches. This function reduces various derivatives of the same lexeme to a common stem. For example, topic "Product customization" failed, because stems 'product' and 'custom' matched such words as 'production', 'productivity', 'customer', 'customs'. Strong stemming is seen as a recall-enhancing technique. Weak stemming is likely to be more appropriate to the HARD task, as we are more interested in achieving high precision, rather than recall. Weak stemming keeps suffixes, and removes only endings, such as plural forms of nouns and past tense forms of verbs.

Another common reason for failure is that, some topic titles simply have insufficient information, for example in topic 186 ("Restricting the Internet") the Description and Narrative sections narrow down the relevance criteria to the documents related to governmental restrictions of the Internet use in China.

The above discussion suggests that a method for evaluating the topic titles and user-selected phrases may be useful in deciding in advance whether or not to use them in the search. For instance, if the terms in a topic title do not constitute a well-formed phrase, then we could select additional terms from the Description and Narrative fields of topics. To this end, we are planning to experiment with co-occurrence statistics and part of speech categories of terms with the aim of developing a better method of query term selection.

Finally, some examples of topics which performed substantially better (38% and more) with the expanded queries than the original are "Insect-borne illnesses" (77), "Genetic Modification technology" (116), "Wartime Propaganda" (198), "The history of nanotechnology" (229) and "Iraq Disarmament" (217). The highest improvement of 90% was achieved for the topic "Insect-borne illnesses". The topic is rather broad, requesting all items which discuss insect-borne disease warnings and how they have affected the life-style of people during the summer months. The success of the query expansion is due to the presence of few very specific highly-weighted relevant terms among those selected by the assessor, for example "Lyme disease" and "St. Louis encephalitis". We evaluated the contribution of each term using the same method as discussed above. For example, with the removal of the term 'Lyme' from the expanded query AveP drops by 14%, and with the removal of the term "Encephalitis"—by 5%. In the topic 229 ("The history of nanotechnology") among the expansion terms which contributed most to performance are: molecule, atom and microscope. Their removal causes average precision to drop by 32%, 19% and 9% respectively.

On the other hand, some highly specific terms for the topic 198 ("Wartime propaganda") that intuitively seemed relevant to the user, such as "kosovar", "Milosevich", "Slobodan" and "Yugoslavia", had a strongly negative effect on performance, most likely because they frequently appeared in related but non-relevant topics.

## 6. The effect of relevant and non-relevant documents on query expansion following user feedback

Query expansion based on relevance feedback is typically more effective than based on blind feedback, however as discussed earlier in Section 3.1 only 73% of the sentences, selected by the users from the clarification form, were actually from relevant documents. In other words, the evaluators who selected sentences were only 73% of the time right in identifying the relevant documents from the one-sentence representations (the evaluators who selected the sentences from clarification forms and those who judged the relevance of documents were the same). This has prompted us to explore the following question: How does the presence of different numbers of relevant and non-relevant documents in the feedback affect average precision?

Previous studies have looked into the effect on performance of the numbers of documents selected in the process of pseudo-relevance (blind) feedback and the correspondence between the performance of the initial run and the expanded run following blind feedback (Carpineto, De Mori, Romano, & Bigi, 2001; Xu & Croft, 1996). The goal of our study was to determine how different numbers of relevant and non-relevant documents in the subset used for query expansion affect average precision.

We conducted a series of runs on the Financial Times and Los Angeles Times corpora and TREC topics 301–450. For each run we composed a set, consisting of the required number of relevant and non-relevant documents. To minimize the difference between relevant and non-relevant documents, we selected non-relevant documents ranked closely to relevant documents in the ranked document set.

The process of document selection is as follows: First all documents in the ranked set are marked as relevant/non-relevant using TREC relevance judgements. Then, each time a relevant document is found, it is recorded together with the nearest non-relevant document, until the necessary number of relevant/non-relevant documents is reached.

The graph in Fig. 3 shows that as the number of relevant documents increases, Average Precision (AveP) after feedback increases considerably for each extra relevant document used up to the point when we have 4 relevant documents. The increment in AveP slows down when more relevant documents are added.

Adding few non-relevant documents to relevant ones causes a considerable drop in the AveP. However, the precision does not deteriorate further when more non-relevant documents are added. As long as more than 3 relevant documents are used, a plateau is hit at around 4–5 non-relevant documents.

The results suggest that the more relevant documents are used for query expansion, the better is the average precision. Even though the use of 5 or more relevant documents does not increase precision considerably, it still does cause an improvement compared to 4 and fewer relevant documents. Another finding is that non-relevant documents do not affect average precision considerably, as long as there are a sufficient number of relevant documents.

To verify these findings and to confirm that the poor performance of the sentence-based query expansion technique at TREC (method 1, TREC run 'UWAThard2') was not due to the presence of non-relevant

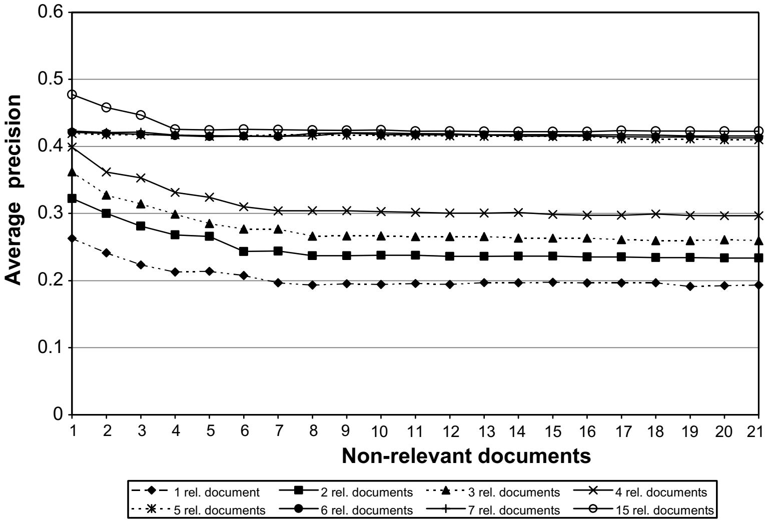

Fig. 3. Effect of relevant and non-relevant documents on query expansion from user feedback.

documents among those selected by the user, we conducted a run only using the documents which were later judged as relevant. The results were very similar (AveP = 0.3188). As discussed earlier in Section 5.1 experiments with a different query expansion method gave similar results.

## 7. Comparison with other systems

The only participating site in HARD track, whose experimental runs performed better than our UWAT-hard3 run, was Queen's college group (Grunfeld, Kwok, Dinstl, & Deng, 2004). Their best baseline system achieved 32.7% AveP (hard-doc) and their best result after clarification forms was 36%, which gives 10% increase over the baseline. We have achieved 26% improvement over the baseline (hard-doc), which is the highest increase over baseline among the top 50% highest-scoring baseline runs.

Queen's college group used clarification forms to present three types of information items to the users for selection: WordNet synonyms of query terms, terms extracted from documents following blind feedback, and titles or first sentences of documents. In addition they provided a free keyword input section for the users to enter any extra terms they deemed useful.

Among HARD track participants, the most common types of information presented to the users in clarification forms, were excerpts from documents and terms/phrases. Several sites used various clustering algorithms for baseline retrieval, representing clusters somewhat differently: University of Illinois (Shen & Zhai, 2004) showed centroid passages (68.8 words on average) of the top clusters; University of Massachusetts (AbduJaleel et al., 2004) represented each cluster with the title of the centroid document and ten top-ranked query expansion terms. University of Maryland (He & Demner-Fushman, 2004) experimented with using either the most informative headline among all documents in each cluster, or merging the headlines of documents in a cluster using a multi-document headline generation tool.

The reason why we chose to use noun phrases as units for interactive query expansion, was to provide more contextual information to help the user. Two other sites showed terms in context: Microsoft Research Cambridge (MSRC) (Robertson et al., 2004) asked the user to do a selection among 15 statistically selected phrases, consisting of two adjacent words, while UMass showed 30 single terms, each with a short sample context from the retrieved documents. UMass used only the actual selected terms for query expansion, not their context, MSRC used each phrase as a whole, whereas we broke down each selected phrase into single terms, which we used in searching. We experimented using whole phrases, but the results were much worse than the baseline.

In addition to asking the user to mark good terms, both MSRC and UMass provided the user with a choice of explicitly marking bad terms. The rationale is to use such terms in downweighting documents containing them.

Three sites (UMass, Queen's and University of Maryland) provided the user with an option of entering any extra terms they considered relevant to their query, a feature that assessors found useful. University of Maryland also asked users to indicate which document sub-collections in the HARD corpus they preferred, and which response format is most useful in satisfying their information need. The latter feature seems to overlap with the granularity metadata provided for each topic.

In addition to the usefulness of clarification forms for relevance feedback purposes, it is important to take into account what was considered helpful by the users. Feedback received from track annotators includes the features they considered helpful in making more confident choices:

- a free text input box,
- a combination of several types of information in one form (e.g. terms and document titles),
- words/phrases in context,
- document titles plus lists of terms from documents.

Further interactive experiments are needed to evaluate how helpful are noun phrases to users in selecting potential expansion terms. Nevertheless, feedback from track annotators, and high results obtained by using noun phrases in our experiments suggest that they facilitate selection of useful query expansion terms.

## 8. Conclusions and future work

The focus of the work reported in this paper is on developing effective methods of gathering and utilising the user's relevance feedback. We have tested two approaches to user-assisted search refinement that aim to minimise the amount of text the user has to read in providing feedback. The first method involved inviting the user to select from the clarification form a number of sentences that may represent relevant documents, and then using those documents whose sentences were selected for query expansion. Although the approach did not produce statistically significant improvement over the baseline in the official TREC runs, the results showed that users were able to identify the relevant documents based on the best sentences shown in the clarification forms with average precision of 0.73 and average recall of 0.69 which suggests that further research in this direction may yield better results in the future.

The second method involved showing to the user a list of noun phrases, extracted from the initial document set, and then expanding the query with the terms from the user-selected phrases. The HARD TREC evaluation results showed that this method yields significant performance improvements. We hypothesize that phrases provide a context for users to judge the usefulness of the terms, in contrast to single terms, which do not provide a context, makes them more appropriate for interactive query expansion, however this point needs to be investigated in future research. Another research question we would like to investigate in the future is whether those terms that do not contribute highly to the overall weight of phrases, nevertheless, contribute significantly to the retrieval performance.

The evaluation results suggest that the second expansion method overall is more promising than the first, and could yield substantial performance improvements, however more analysis needs to be done to determine the key factors influencing the performance of both methods.

Another major goal of the HARD track, which we did not address this time, is to promote research into how contextual and extra-linguistic information about the user and the user's search task could be harnessed to achieve high accuracy retrieval. To effectively use information such as user's familiarity with the topic, the purpose of the user's search or the user's genre preferences we need more complex linguistic and stylistic analysis techniques.

## References

AbduJaleel, N., Corrada-Emmanuel, A., Li, Q., Liu, X., Wade, C., & Allan, J. (2004). UMass at TREC 2003: HARD and QA. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the twelfth text retrieval conference*, November 18–21, 2003, NIST, Gaithersburg, MD, pp. 715–725.

Allan, J. (2004). HARD track overview in TREC 2003 high accuracy retrieval from documents. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the twelfth text retrieval conference*, November 18–21, 2003, NIST, Gaithersburg, MD, pp. 24–37.

Beaulieu, M. (1997). Experiments with interfaces to support Query Expansion. *Journal of Documentation, 53*(1), 8–19.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics, 21*(4), 543–565.

Carpineto, C., De Mori, R., Romano, G., & Bigi, B. (2001). An information theoretic approach to automatic query expansion. *ACM Transactions on Information Systems, 19*(1), 1–27.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon* (pp. 115–164). Lawrence Elbraum Associates: Englewood Cliffs NJ.

Dziadosz, S., & Chandrasekar, R. (2002). Do thumbnail previews help users make better relevance decisions about web search results? *Proceedings of the 25th ACM-SIGIR conference*, Tampere, Finland.

Grunfeld, L., Kwok, K. L., Dinstl, N., & Deng, P. (2004). TREC 2003 Robust, HARD and QA track experiments using PIRCS. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the twelfth text retrieval conference*, November 18–21, 2003, NIST, Gaithersburg, MD, pp. 510–521.

He, D., & Demner-Fushman, D. (2004). HARD experiment at Maryland: from need negotiation to automated HARD process. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the twelfth text retrieval conference*, November 18–21, 2003, NIST, Gaithersburg, MD, pp. 707–714.

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

Ramshaw, L., & Marcus, M. (1995). Text Chunking Using Transformation-Based Learning. *Proceedings of the third ACL workshop on very large corpora, MIT.*

Robertson, S. E., Zaragoza, H., & Taylor, M. (2004). Microsoft Cambridge at TREC-12: HARD track. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the twelfth text retrieval conference*, November 18–21, 2003, NIST, Gaithersburg, MD, pp. 418–425.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation, 46*(4), 359–364.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART Retrieval System* (pp. 312–323). Prentice Hall.

Shen, X., & Zhai, C. (2004). Active feedback—UIUC TREC-2003 HARD Experiments. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the twelfth text retrieval conference*, November 18–21, 2003, NIST, Gaithersburg, MD, pp. 662–666.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing and Management, 36*(6), 779–808 (Part 1); 809–840 (Part 2).

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st ACM SIGIR conference*, Melbourne, Australia, pp. 2–10.

Vechtomova, O., Karamuftuoglu, M., & Lam, E. (2004). Interactive search refinement techniques for HARD tasks. In E. Voorhees & L. Buckland (Eds.), *Proceedings of the twelfth text retrieval conference*, November 18–21, 2003, NIST, Gaithersburg, MD, pp. 820–827.

Vechtomova, O., Robertson, S. E., & Jones, S. (2003). Query expansion with long-span collocates. *Information Retrieval, 6*(2), 251–273.

White, R. W., Jose, J. M., & Ruthven, I. (2003). A granular approach to Web search result presentation. *Proceedings of the 9th international conference on human computer interaction*, September 1–5, Zurich, Switzerland.

Xu, J., & Croft, B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (SIGIR'96). Zurich, Switzerland, pp. 4–11.