Comparison of Two Interactive Search Refinement Techniques

Olga Vechtomova
Department of Management Sciences
University of Waterloo
200 University Avenue West, Waterloo,
Canada
ovechtom@engmail.uwaterloo.ca

Murat Karamuftuoglu
Department of Computer Engineering
Bilkent University
06800 Bilkent Ankara,
Turkey
hmk@cs.bilkent.edu.tr

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. One method consists of asking the user to select a number of sentences that may represent relevant documents, and then using the documents, whose sentences were selected for query expansion. The second method consists of 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 phrases selected by the user.

1 Introduction

Query expansion following relevance feedback is a well established technique in information retrieval, which aims at improving user search performance. It combines user and system effort towards selecting and adding extra terms to the original query. The traditional model of query expansion following relevance feedback is as follows: the user reads a representation of a retrieved document, typically its full-text or abstract, and provides the system with a binary relevance judgement. After that the system extracts query expansion terms from the document, which are then added to the query either manually by the searcher – interactive query expansion, or automatically – automatic query expansion. Intuitively interactive query expansion should produce better results than automatic, however this is not consistently so (Beaulieu 1997, Koenemann and Belkin 1996, Ruthven 2003).

In this paper we present two new approaches to automatic and interactive query expansion, which we developed and tested within the framework of the High Accuracy Retrieval from Documents (HARD) track of TREC (Text Retrieval Conference).

1.1 HARD track

The main goal of the new HARD track in TREC-12 is to explore what techniques could be used to improve search results by using two types of information:

1. Extra-linguistic contextual information about the user and the information need, which was provided by track organisers in the form of metadata. 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.

2. Relevance feedback given by the user in response to topic clarification questions. This information was elicited by each site by means of a (manually or automatically) composed set of clarification forms per

topic. The forms are filled in by the users (annotators), and provide additional search criteria.

In more detail the HARD track evaluation scenario consists of the following steps:

1) The track organisers invite annotators (users), each of whom formulates one or more topics. An example of a typical HARD topic is given below:

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

2) Participants receive Title, Description and Narrative sections of the topics, and use any information from them to produce one or more baseline runs.

3) Participants produce zero or more clarification forms with the purpose of obtaining feedback from the annotators. Only two forms were guaranteed to be filled out.

4) All clarification forms for one topic are filled out by the annotator, who has composed that topic.

5) Participants receive the topic metadata and the annotators' responses to clarification forms, and use any data from them to produce one or more final runs.

6) Two runs per site (baseline and final) are judged by the annotators. Top 75 documents, retrieved for each topic in each of these runs, are assigned binary relevance judgement by the annotator – author of the topic.

7) The annotators' relevance judgements are then used to calculate the performance metrics (see section 4).

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 1.7Gb.

The primary goal of our participation in the track was to investigate how to achieve high retrieval accuracy through relevance feedback. The secondary goal was to study ways of reducing the amount of time and effort the user spends on making a relevance judgement, and at the same time assisting the user to make a correct judgement.

We evaluated the effectiveness of two different approaches to eliciting information from the users. The first approach is to represent each top-ranked retrieved document by means of one sentence containing the highest proportion of query terms, and ask the user to select those sentences, which possibly represent relevant documents. The second method extracts noun phrases from top-ranked retrieved documents and asks the user to select those, which might be useful in retrieving relevant documents. 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 represented in the form of the titles and/or the first few sentences of each document. Based on this information the user has to make initial implicit relevance judgements: whether to refer to the full text document or not. Explicit relevance feedback is typically requested by IR systems after the user has seen the full text document, an example of such IR system is Okapi (Robertson et al. 2000, Beaulieu 1997). Reference to full text documents is obviously time-consuming, therefore it is important to represent documents in the hitlist in such 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 correct relevance judgement. Query-biased summaries, usually constructed through the extraction of 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 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.

The remainder of this paper is organised as follows: sections 2 and 3 present the two document representation and query expansion methods we developed, section 4 discusses their evaluation, and section 5 concludes the paper and outlines future research directions.

2 Query expansion method 1

According to the HARD track specifications, a

clarification form for each topic must fit into a screen with 1152 x 900 pixels resolution, and the user may spend no more than 3 minutes filling out each form.

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 representation of a document. The questions explored here were: What is the error rate in selecting relevant documents on the basis of one sentence representation of its content? How does sentence-level relevance feedback affect retrieval performance?

2.1 Sentence selection

The sentence selection algorithm consists of the following steps:

We take N top-ranked documents, retrieved in response to query terms from the topic title. 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. 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* of all query terms present in the sentence.

$$S1 = \sum idf_q \qquad (1)$$

Sentence selection score 2 (S2):

$$S2 = \frac{\sum W_i}{f_s} \qquad (2)$$

Where: W_i – Weight of the term *i*, see (3);
f_s – length normalisation factor for sentence *s*, see (4).

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

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

Where: idf_i – inverse document frequency of term *i* in the corpus; tf_i – frequency of term *i* in the document; *tmax* – *tf* of the term with the highest frequency in the document.

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

$$f_s = \frac{s\max}{slen_s} \qquad (4)$$

Where: *smax* – 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.

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. However, we recognise that it is more difficult, if not impossible, for users to reliably determine the relevance of the document on the basis of one sentence, especially in cases where the relevance of the document to the query is due to more subtle aspects of the topic.

2.2 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 et al. (2003) have demonstrated that expansion with long-span collocates of query terms obtained from 5 known relevant documents showed 72-74% improvement over the use of Title-only query terms on the Financial Times (TREC volume 4) 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:

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

Where: $f_r(x,y)$ – frequency of *x* and *y* occurring in the same windows in the known relevant document set (see (6); $f_c(y)$ – frequency of *y* in the corpus; $f_r(x)$ – frequency of *x* in the relevant documents; $v_x(R)$ – average size of windows around *x* in the known relevant document set (R); *N* – the total number of non-stopword occurrences in the corpus.

The frequency of *x* and *y* occurring in the same windows in the relevant set – $f_r(x,y)$ – is calculated as follows:

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

Where: *m* – number of windows in the relevant 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 et al. 1991). The remaining collocates are sorted by their Z score. The above Z score formula is described in more detail in (Vechtomova et al. 2003).

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. For each collocate the collocate score (C1) is calculated:

$$C1 = \sum n_i W_i \qquad (7)$$

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*.

The reason why we use the rank of the collocate in the above formula instead of its Z score is because Z scores of collocates of different terms are not comparable.

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.

We tested the algorithm on past TREC data (Financial Times and Los Angeles Times newswire corpora, topics 301-450) with blind feedback using Okapi BM25 search function (Sparck Jones et al. 2000). 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 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+Desc. runs; (2) larger span sizes also tend to degrade performance in both Title and Title+Desc runs.

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

We used Okapi BM25 document retrieval function for topics with granularity *Document*, and Okapi passage retrieval function BM250 (Sparck Jones et al. 2000) 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 in section 2.1 above.

3 Query expansion method 2

The second user feedback mechanism that we evaluated 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.

The research question explored here is whether noun phrases provide sufficient context for the user to select potentially useful terms for query expansion.

We take top 25 documents from the baseline run, and select 2 sentences per document using the algorithm described above. 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 and Marcus 1995) to extract noun phrases from these sentences. The phrases are then parsed in Okapi to obtain their term weights, removing all stopwords and phrases consisting entirely of the original query terms. The remaining phrases are 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 used to expand the original user query. The expanded query was then searched against the HARD track database in the same way as in the query expansion method 1 described in the previous section.

4 Evaluation

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

1. SOFT-DOC – document-level evaluation, where only the traditional TREC topic formulations (title, description, narrative) are used as relevance criteria.
2. HARD-DOC – the same as the above, plus 'purpose', 'genre' and 'familiarity' metadata are used as additional relevance criteria.
3. HARD-PSG – passage-level evaluation, which in addition to all criteria in HARD-DOC also requires that retrieved items satisfy the granularity metadata (Allan 2004).

Document-level evaluation was done by the traditional IR metrics of mean average precision and precision at various document cutoff points. In this paper we focus on document-level evaluation. Passage-level evaluation is discussed elsewhere (Vechtomova et al. 2004).

4.1 Document-level evaluation

For all of our runs we used Okapi BSS (Basic Search System). For the baseline run we used keywords from the title field only, as these proved to be most effective in our preliminary experiments described in section 2.2. Topic titles were parsed in Okapi, weighted and searched using BM25 function against the HARD track corpus.

Document-level results of the three submitted runs are given in table 1. UWAThard1 is the baseline run using original query terms from the topic titles. UWAThard2 is a final run using query expansion method 1, outlined earlier, plus the granularity and known relevant documents metadata. UWAThard3 is a final run using query expansion method 2 plus the granularity metadata.

Run	Run description	SOFT-DOC evaluation		HARD-DOC evaluation	
		Precision @ 10	Average Precision	Precision @ 10	Average Precision
UWAThard1*	Original title-only query terms; BM25 used for all topics	0.4875	0.3134	0.3875	0.2638
UWAThard2*	Query expansion method 1; granularity and related text metadata	0.5479	0.3150	0.4354	0.2978
UWAThard3*	Query expansion method 2; granularity metadata	0.5958	0.3719	0.4854	0.3335
UWAThard4	As UWAThard1, but BM250 is used for topics requiring passages	0.4729	0.2937	0.3667	0.2450
UWAThard5	As UWAThard2, but related text metadata is not used	0.5229	0.3016	0.4062	0.2828

Table 1. Document-level evaluation results (* runs submitted to TREC)

The fact that the query expansion method 1 (UWAThard2) produced no improvement over the baseline (UWAThard1) was a surprise, and did not correspond to our training runs with the Financial Times and Los Angeles Times collections, which showed 21% improvement over the original title-only query run. We evaluated the user selection of the sentence using average precision, calculated as the number of relevant sentences selected by the user out of the total number of sentences selected, and average recall – the number of relevant sentences selected by the user out of the total number of relevant sentences shown in the clarification form. Average precision of TREC sentence selections made by TREC annotators is 0.73, recall – 0.69,

what is slightly better than our selections during training runs (precision: 0.70, recall: 0.64). On average 7.14 relevant sentences were included in the forms. The annotators on average selected 4.9 relevant and 1.8 non-relevant sentences.

Figure 1 shows the number of relevant/non-relevant selected sentences by topic. It is not clear why query expansion method 1 performed worse in the official UWAThard2 run compared to the training run, given very similar numbers of relevant sentences selected. Corpus differences could be one reason for that – HARD corpus contains a large proportion of governmental documents, and we have only evaluated our algorithm on newswire corpora. More experiments need to be done to determine the effect of the governmental documents on our query expansion algorithm.

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 2.2. 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.

We thought that one possible reason for the poor performance of UWAThard2 compared to the baseline run UWAThard1 was the fact that we used document retrieval search function BM25 for all topics in the UWAThard1, whereas for UWAThard2 we used BM25 for topics requiring document retrieval and BM250 for the topics requiring passage retrieval. The two functions produce somewhat different document rankings. In UWAThard4 we used BM250 for the topics requiring passages, and got only a slightly lower average precision of 0.2937 (SOFT-DOC evaluation) and 0.2450 (HARD-DOC evaluation).

Our second query expansion method on the contrary did not perform very well in the training runs, achieving only 10% improvement over the original title-only query run. The official run UWAThard3, however resulted in 18% increase in average precision (SOFT-DOC evaluation) and 26.4% increase in average precision (HARD-DOC evaluation). Both improvements are statistically significant (using t-test at .05 significance level).

TREC annotators selected on average 19 phrases, whereas we selected on average 7 phrases in our tests. This suggests that selecting more phrases leads to a notably better performance. The reason why we selected fewer phrases than the TREC annotators could be due to the fact that on many occasions we were not sufficiently familiar with the topic, and could not determine how an out-of-context phrase is related or not related to the topic. TREC annotators are, presumably, more familiar with the topics they have formulated.

In total 88 runs were submitted by participants to the HARD track. All our submitted runs are above the median in all evaluation measures shown in table 1. The only participating site, whose expansion runs performed better than our UWAThard3 run, was the Queen's college group (Kwok et al. 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.

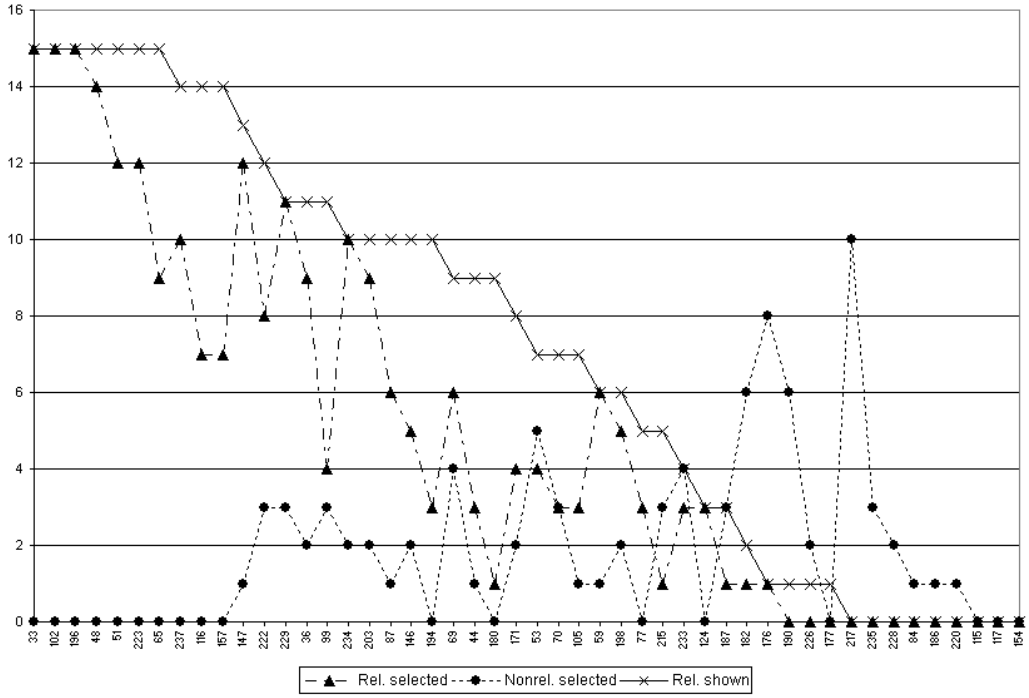

Figure 1. Sentences selected by TREC annotators from the clarification form 1.

4.2 The effect of different numbers of relevant and non-relevant documents on performance following user feedback

Query expansion based on relevance feedback is typically more effective than based on blind feedback, however as discussed in the previous section, only 73% of the sentences selected by users from the clarification form 1 were actually relevant. 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?

With this goal, we conducted a series of runs on 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

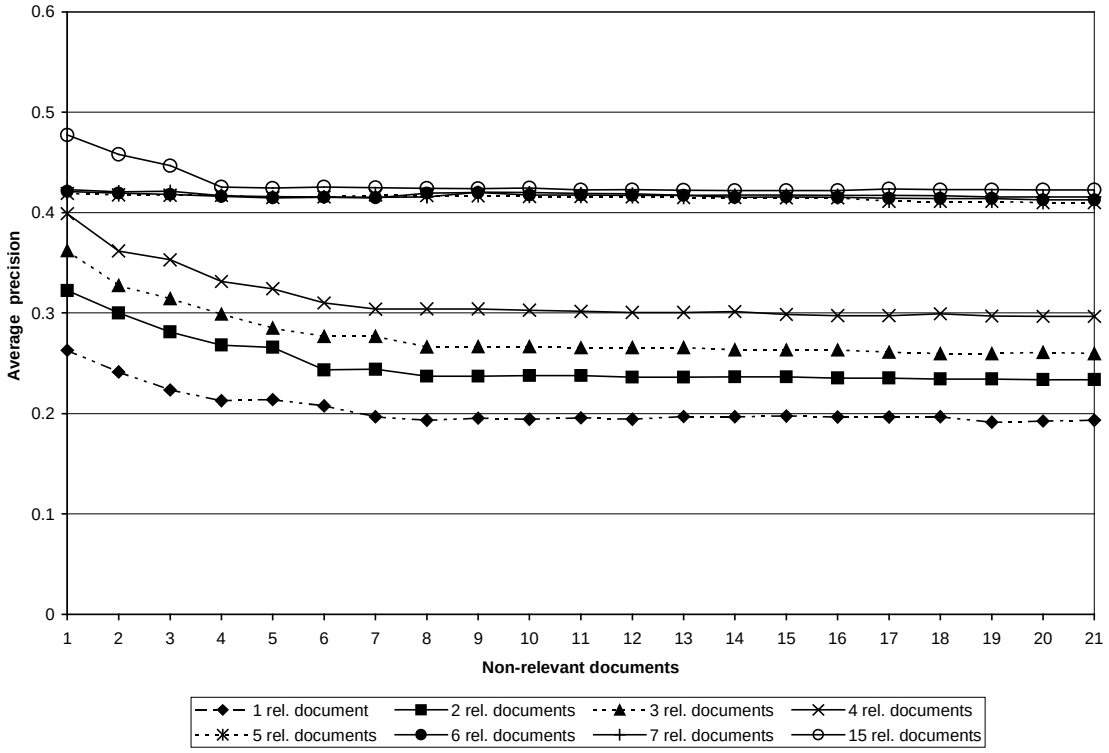

Figure 2: Effect of relevant and non-relevant documents on query expansion from user feedback

number of relevant/non-relevant documents is reached.

The graph in figure 2 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 (Figure 2). As long as there are more than three relevant documents that are used, a plateau is hit at around 4-5 non-relevant documents.

We can conclude from this experiment that as a general rule, the more relevant documents are used for query expansion, the better is the average precision. Even though use of 5 or more relevant documents does not increase the 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.

5 Conclusions and future work

In this paper we presented two user-assisted search refinement techniques:
(1) inviting the user to select from the clarification form a number of sentences that may represent relevant documents, and then using the documents whose sentences were selected for query expansion.
(2) 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 evaluation results suggest that the second expansion method overall is more promising than the first, demonstrating statistically significant performance improvement over the baseline run. More analysis needs to be done to determine the key factors influencing the performance of both methods.

The focus of our experiments in the HARD track of TREC-12 was on developing effective methods of gathering and utilising the user's relevance feedback. 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. We plan to address these issues in the next year's entry.

Acknowledgements

This material is based on work supported in part by Natural Sciences and Engineering Research Council of Canada.

References

Allan, J. 2004. HARD Track Overview. Proceedings of the Twelfth Text Retrieval Conference, November 18-21, 2003, Gaithersburg, MD.

Beaulieu, M. 1997. Experiments with interfaces to support Query Expansion. Journal of Documentation, 53(1), pp. 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), pp. 543-565.

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

Koenemann J. and 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, Zurich, pp. 205-215.

Kwok L. et al. 2004. TREC2003 Robust, HARD and QA track experiments using PIRCS. Proceedings of the Twelfth Text Retrieval Conference, November 18-21, 2003, Gaithersburg, MD.

Ramshaw L. and Marcus M. 1995. Text Chunking Using Transformation-Based Learning. Proceedings of the Third ACL Workshop on Very Large Corpora, MIT.

Robertson S.E., Walker S. and Beaulieu M. 2000. Experimentation as a way of life: Okapi at TREC. Information Processing and Management, 36, pp. 95-108.

Ruthven I. 2003. Re-examining the potential effectiveness of interactive query expansion. Proceedings of the 26th ACM-SIGIR conference, Toronto, Canada, pp. 213-220.

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

Tombros A., Sanderson M. 1998. Advantages of Query Biased Summaries in Information Retrieval. 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. Proceedings of the Twelfth Text Retrieval Conference, November 18-21, 2003, Gaithersburg, MD.

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