Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 849–865
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Query expansion with terms selected using lexical cohesion analysis of documents

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

[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 9 March 2006; received in revised form 7 September 2006; accepted 9 September 2006
Available online 30 October 2006

Abstract

We present new methods of query expansion using terms that form lexical cohesive links between the contexts of distinct query terms in documents (i.e., words surrounding the query terms in text). The link-forming terms (link-terms) and short snippets of text surrounding them are evaluated in both interactive and automatic query expansion (QE). We explore the effectiveness of snippets in providing context in interactive query expansion, compare query expansion from snippets vs. whole documents, and query expansion following snippet selection vs. full document relevance judgements. The evaluation, conducted on the HARD track data of TREC 2005, suggests that there are considerable advantages in using link-terms and their surrounding short text snippets in QE compared to terms selected from full-texts of documents.

Keywords: Interactive information retrieval; Query expansion; Relevance feedback; Lexical cohesion

1. Introduction

Lexical cohesion is a characteristic of text, which is achieved through semantic connectedness between words, and expresses continuity between the parts of text (Halliday & Hasan, 1976). Segments of text which are about the same or similar subjects have higher lexical cohesion, i.e., share a larger number of semantically related or repeating words, than unrelated segments. The strength of lexical cohesion between two query terms can be useful in determining whether query terms are used in related contexts.

An earlier study by Vechtomova, Karamuftuoglu, and Robertson (2005, 2006) compared sets of relevant and non-relevant documents, demonstrating that on average distinct query terms tend to be more lexically cohesive in relevant documents than non-relevant documents. Lexical cohesion between two segments of text, such as sentences or fixed-size windows around query term occurrences, can be estimated based on how many identical or semantically related words they have in common. Arguably, if distinct query terms in a document occur in contexts that have few semantically related words in common, it is likely that they refer to something

* Corresponding author. Tel.: +1 519 888 4567x32675; fax: +1 519 746 7252.
E-mail addresses: ovechtom@uwaterloo.ca (O. Vechtomova), hmk@cs.bilkent.edu.tr (M. Karamuftuoglu).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.09.004

other than the query topic. By contrast, distinct query terms, whose contexts have many semantically related words in common, are more likely to be related to the user's information need, even though they may be separated by long distances in the text.

In this study we investigate whether words that form lexical (cohesive) links between the contexts of query term instances in a document are also good query expansion (QE) terms. In the present study a *lexical link* is defined as a relationship between two instances of the same word in the contexts (defined as fixed-size windows) of distinct query terms in a document. The link-forming words (link-terms) and surrounding terms (snippets) are then evaluated for their usefulness in both interactive and automatic QE.

The goal of this study is twofold. The first is to evaluate link-terms and terms in close proximity to them as QE terms by comparing their effectiveness against a well-known benchmark QE method in both interactive and automatic query expansion scenarios. The second is to evaluate in an interactive setting the effect of showing the query expansion terms with and without the surrounding context to the user. We also investigate whether extracting QE terms from one-line text snippets selected by the users is better than extracting QE terms from the whole documents that they represent. Lastly, we look at how query expansion following snippet-level relevance judgements compares to full document-level relevance judgements.

The paper is organised as follows: in Section 2, we give an overview of previous work on query expansion following relevance feedback, and briefly present the method of estimating lexical cohesion between query terms in documents, which is described in detail in Vechtomova et al. (2006); Section 3 describes the experiments and evaluation results; Section 4 concludes the paper.

2. Related work

2.1. Query expansion following relevance and blind feedback

Query expansion following relevance feedback has received significant attention in IR research in the past 30 years. Relevance feedback (RF) is a mechanism by which the system, having retrieved some information items in response to the user's query, asks the user to assess their relevance to his/her information need. Such information items are typically documents, which are shown to the user in some surrogate form, for example, as document titles, abstracts, snippets of text, query-biased or general summaries, keywords and key-phrases. The user may also have an option to see the whole document before making the relevance judgement. After the user has selected some documents as relevant, query expansion terms are extracted from them and either added to the query automatically, or shown to the user for further selection. The former process is known as automatic query expansion (AQE), while the latter as interactive query expansion (IQE). The expanded query is then used to retrieve a new ranked set of documents.

While relevance feedback has been demonstrated to yield substantial and consistent gains in performance in experimental settings (Spärck Jones, Walker, & Robertson, 2000), its uptake in real-world applications, in particular Web search, has been limited. This is due in part to technological limitations, e.g., the need to maintain session continuity, and partly to the fact that users seem to be used to manually reformulating queries. A related approach to RF is to use a number of top-ranked documents in the initially retrieved set for query expansion, without asking the user to assess their relevance. This approach, known as pseudo-relevance or blind feedback (BF), has been demonstrated to be less robust in performance than relevance feedback, especially on large collections. A study by Billerbeck and Zobel (2004) reports that on TREC-9 10Gb Web track corpus their implementation of Okapi blind feedback method improves less than a third of topics. They also conclude that the best values for such BF parameters as the number of documents and terms used for QE vary widely among topics, which prompted a new line of research towards query-specific setting of such parameters.

While there is clearly a scope for improvement in blind feedback, the process of relevance feedback also needs further attention to make it more attractive and less time-consuming for the user. Indeed, most of the benefits of RF reported in the literature have been observed in the conditions where the user judges relevance of a document, having read its entire text or human-written abstracts (e.g., Beaulieu & Jones, 1998). Although this is the most reliable way for the user to determine a document's relevance, it is a time-consuming process that few users are willing to pursue. Surrogate document representations shown in the

retrieved list, such as titles and URLs, are meant to help user quickly determine which documents are worth reading. Arguably, such representations are not very informative, and some research has focused on how to build representations that more accurately reflect the document contents and maximize user's chances to visit relevant documents.

For example, Tombros and Sanderson (1998) evaluated the effectiveness of query-biased summaries as document representations. White, Jose, and Ruthven (2005) describe an approach, whereby the most salient and query-related sentences are extracted from the retrieved document, then ranked and presented to the user independently of the documents they originate from. They compare sentence-based representations with traditional representations of the title and URL, and find that users interact more with the sentence-based representations. A study by Vechtomova and Karamuftuoglu (2006) evaluated one-sentence document representations in terms of how much they help users select relevant documents. The experiment, conducted as part of the HARD (High Accuracy Retrieval from Documents) track evaluation in TREC 2003, consisted of showing users one-sentence representations of documents, and asking them to select those that they find relevant. No other document-specific information (such as title) was shown. The same users, who selected sentences, later judged the whole documents for relevance. Sentences shown to the users were selected on the basis of the number of query terms they contain and the informativeness (calculated using *tf.idf*) of other terms they have. The results showed that users selected relevant documents based on their one-sentence representations with the average precision of 73% and average recall of 69%.

As well as acting as document surrogates, such representations can also be a better source of query expansion terms than complete documents. A long and multi-topic document, only a part of which is relevant to the query, can potentially yield some unrelated QE terms that can hurt performance. A study by Lam-Adesina and Jones (2001) evaluated the usefulness of both query-biased and general document summaries in query expansion following blind feedback, reporting substantial improvements over the document-based QE using RSV, a well-known QE selection method.

In this study we use the concept of lexical cohesion to develop new methods of selecting query expansion terms. We evaluate whether words that form lexical cohesion between the contexts of query terms in a document are good QE terms, and whether segments of text surrounding such words are effective in the interactive setting for eliciting relevance feedback from the user.

2.2. Lexical cohesion between query terms in a document

The query expansion methods presented in this paper rely on the method of calculating lexical cohesion between query terms' contexts in a document introduced in Vechtomova et al. (2006). Below we briefly describe the method.

The main goal of this method is to rank documents by taking into consideration how cohesive the contextual environments of distinct query terms are in each document. The assumption is that if there is a high degree of lexical cohesion between the contexts of distinct query terms in a document, they are likely to be topically related, and there is a greater chance that this document is relevant to the user's query. The ranking method combines the BM25 document matching score (Robertson, Walker, Jones, Hankock-Beaulieu, & Gatford, 1995) with the lexical cohesion based score, as described in more detail later in this section.

Lexical cohesion between query terms' contexts in a document is calculated by counting the number of lexical links between them. The *context* of a query term in a document is defined as a set of stemmed non-stop terms extracted from fixed-sized windows around each occurrence of the query term in the document. In Vechtomova et al. (2006) different window sizes were tested, and window size of 20 (10 words on either side of the query term instance) was found to be optimal, and is used in the present study. A *lexical link* is a relationship between two instances of the same lexeme (simple lexical repetition), its morphological derivatives (complex lexical repetition) or semantically related words (such as hyponyms, synonyms, meronyms, etc.). In Vechtomova et al. (2006) simple lexical repetition alone performed as well as the use of repetition plus semantically related words (determined using WordNet), therefore in this work we only use simple lexical repetition for identifying lexical links.

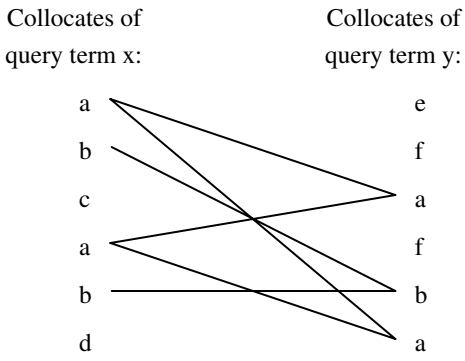

Fig. 1. Links between instances of common collocates in merged windows of query terms x and y.

Fig. 1 demonstrates lexical links between the contexts of two distinct query terms x and y in a document. The left column contains all non-stop terms in the fixed-size windows surrounding all occurrences of the query term x in the document, and the right column contains all non-stop terms in the fixed-size windows surrounding all occurrences of the query term y. The lines between instances of the same term in the figure represent lexical links. In this example there are altogether 6 links. If there are more than two distinct query terms in a document, a comparison of each pair is done. The number of links are recorded for each pair, and summed up to find the total number of lexical links in the document. A more detailed account of this method is given in Vechtomova et al. (2006).

For each document a *lexical cohesion score* (LCS) is calculated as follows:

$$\mathrm{LCS} = \frac{L}{V} \tag{1}$$

where L is the total number of lexical links between all pairs of distinct query terms in a document; V – the total number of non-stop terms in the contexts of all query terms in a document.

For document ranking, the following simple linear function which combines the BM25 document matching score (Robertson et al., 1995) and the lexical cohesion score is proposed (Vechtomova et al., 2006), where x is a tuning constant and MS is the BM25 matching score:

$$\mathrm{COMB-LCS} = \mathrm{MS} + x * \mathrm{LCS} \tag{2}$$

In Vechtomova et al. (2006) the following values of x were tested: 0.25, 0.5, 0.75, 1, 1.5, 3, 4, 5, 6, 7, 8, 10 and 30, with $x = 8$ showing the best performance, and statistically significant improvement over BM25 in P@10 (Precision at 10 documents).

In the following section we describe the experiments conducted to evaluate the use of link-forming words (link-terms) and words in the snippets surrounding them (snippet-terms) in both interactive and automatic query expansion scenarios.

3. Experiments

3.1. Dataset and evaluation methodology

The High Accuracy Retrieval from Documents (HARD) track of TREC 2005 (Allan, 2005) was used as the evaluation platform in this study. Unlike the traditional TREC ad hoc search scenario, the HARD track evaluation scenario includes a one-time interaction with the user.

In HARD 2005, 50 topics from the previous years' "Robust" track of TREC were used. Topics were defined in the standard TREC fashion, consisting of the Title (few words describing the information need), Description (a one-sentence brief description of the information need) and Narrative (a one-paragraph

detailed description of the information need, and, sometimes, the motivation for the need and the relevance criteria) sections. Topics used in the Robust track are known to be more complex and difficult to satisfy than topics in the ad hoc track, and the reason for their use in HARD was to evaluate whether the retrieval performance of such complex queries can benefit from limited interaction with the user.

Six assessors (users) were invited by NIST, and each was allocated 8–9 of the 50 topics. The participants performed baseline runs, for which they were free to use any section(s) of the topics and any IR methods and systems. The participants were also given an opportunity to engage in a single asynchronous interaction with the users by means of clarification forms. Having received the topics (queries), the participants were free to use clarification forms in any way they liked to elicit useful information from the users in order to improve the search results. The clarification forms had to be viewed by the user via a Web browser, but were not limited in size and the number of pages, and were free to contain any components (e.g., html forms, images, scripts, etc.). The only limitations were that the user had to spend no more than 3 min on each form, and the form could not use any resources or files outside the user's machine. The most common use of clarification forms by HARD participants was to elicit relevance feedback, whereby users were asked to select terms, phrases and text segments. Some also asked users to enter additional keywords or to answer a fixed set of questions. Each participant could submit no more than 3 clarification forms per topic. The order in which clarification forms from different participants were shown to the users was rotated.

When the participants received the users' responses to the clarification forms, they were free to use them in any manner to improve their final runs. As in the traditional ad hoc track scenario, participants were required to submit 1000 top-ranked documents for each run. Document pooling and judgement were also done according to the standard TREC procedure: the top 55 documents from one baseline (pre-interaction) run and one final (post-interaction) run submitted by each participant were pooled and given to the users for evaluation. For each topic, the user who judged the documents was the same person who provided responses to the clarification forms. There was a time gap of several weeks between the stages where the users responded to clarification forms and judged documents. Unlike the ad hoc track, where document judgement was binary (1 – relevant, 0 – non-relevant), in HARD 2005, documents were judged as follows: 2 – highly relevant, 1 – relevant and 0 – non-relevant. Since it is not clear what criteria were used for judging a document as either 1, or 2, in the evaluations reported in this paper we do not differentiate between them, i.e., consider all documents judged 1 or 2 as relevant.

The method against which the developed QE methods are compared in this paper is the Offer Weight (OW) – a well-established method of query expansion term selection (Robertson, 1990; Spärck Jones et al., 2000). According to this method, all non-stop terms are extracted from the full-texts of the known relevant documents and ranked by the Offer Weight, calculated as follows:

$$\mathrm{OW} = r\mathrm{RW} \tag{3}$$

where r is the number of relevant documents containing the candidate QE term; RW is the term relevance weight calculated as follows:

$$\mathrm{RW} = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \tag{4}$$

where N is the total number of documents in the corpus; n – number of documents containing the term; R – number of known relevant documents in the corpus; r – number of relevant documents containing the term. After the candidate QE terms are ranked by OW, a fixed number of top-ranked terms is added to the original query and used in the search.

In the following section we describe in detail the proposed query expansion (QE) methods, which rely on the lexical link structure in documents. The methods were evaluated in the interactive search scenario of HARD track, using the feedback provided by the users by means of clarification forms (Section 3.3). In Sections 3.4 and 3.5 we also describe batch-mode evaluations conducted following TREC 2005, in which we used HARD 2005 relevance judgements.

3.2. Query expansion term selection and clarification forms construction

To construct clarification forms, a baseline run (henceforth referred to as the "*Unexpanded*" run) was performed. The top 1000 documents per topic retrieved[1] by terms from the Title section of the HARD track topics, using Okapi BM25 (Robertson et al., 1995), were re-ranked using Eq. (2) above with the tuning constant $x = 8$, since that showed the best performance on other corpora (Vechtomova et al., 2006).[2]

All link-terms (words forming links by simple lexical repetition between the contexts of any two distinct query terms) were extracted from each of the top 25 documents retrieved in the "Unexpanded" run for each topic. They were ranked by their inverse document frequency (*idf*), and the two top-ranked link-terms per document were selected for inclusion in the HARD clarification form. Next to each term was checkbox for the user to click if she selected that term for query expansion. The user was given simple instructions in each clarification form: "Select the terms that you consider relevant to the topic." This form is henceforth referred to as "terms out of context" clarification form.

To evaluate the role of context, a second clarification form ("terms in context") was generated which contained the same link-terms as in the previous form, but this time in the context of short one-line snippets. A snippet was defined as consisting of 3 non-stopwords on either side of the link-term. The above number of words was chosen because we were interested in evaluating the effectiveness of very short contexts on QE. The detailed snippet-selection algorithm is as follows:

```
For each document A
   Find all link-terms in the document
   Rank them by idf
   For each of the two top-ranked link-terms in the document A
      Find all their instances
      For each instance
         Identify a snippet around the instance that contains 3 non-stopwords before and after it in text.
      End For
      Rank all snippets for each link-term by the average idf of their constituent non-stopwords.
      Include the top-ranked snippet into the clarification form.
   End For
End For
```

The link-terms shown in the clarification form in the context of surrounding terms (snippet-terms) were presented in bold font, and a checkbox was placed next to each snippet. Users were given instructions that by clicking on the checkbox, they selected only the highlighted link-term, but if they thought other terms from the snippets might be useful they could copy and paste them into the textbox at the bottom of the clarification form.[3] The text of the instructions was: "Terms suggested for query expansion are shown in bold in the context of surrounding terms (snippets). Select as many of the suggested terms as you like by clicking on the checkbox next to each snippet. You may also copy & paste any other word from the snippets into the textbox provided at the bottom of the window".

[1] Note that due to an indexing error a portion (about 25%) of the HARD 2005 corpus was not used in the construction of the clarification forms. In the expanded runs, we used the whole corpus; similarly, the baseline run (*Unexpanded*) results reported here are also based on the whole corpus. Although absolute values would possibly be different if the whole collection was used, the conclusions derived from the comparisons of the results of different QE methods remain the same as all runs discussed in this section make use of the same data taken from clarification forms.

[2] However, in post-TREC experiments on HARD 2005 corpus better performance is achieved when x is set to 1.5 and the window size to 10. This run gave an improvement of 8.5% in P@10 over the BM25 based run without hurting the MAP and R-Prec measures, while $x = 8$ gave similar results to BM25.

[3] These terms were not used in the study reported in this paper.

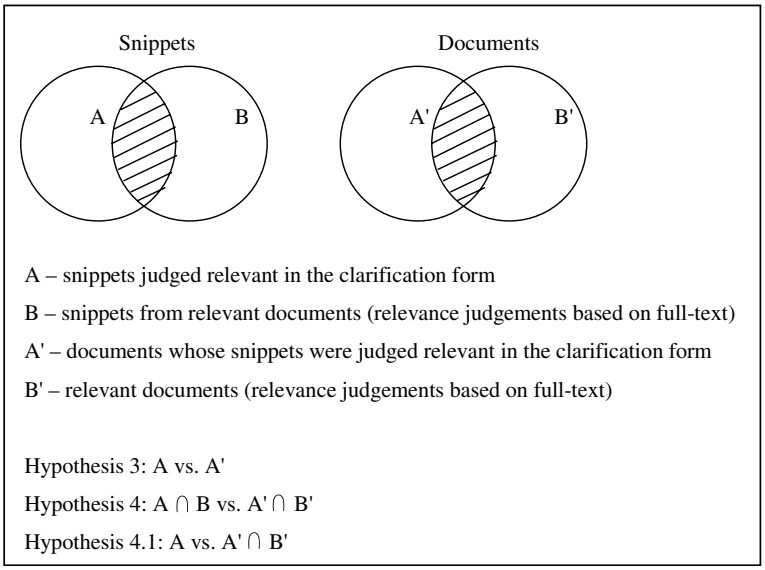

Fig. 2. Summary of Hypotheses 3, 4 and 4.1.

In total, from 49 "terms in context" clarification forms[4] users selected 281 (on average 5.7 per topic) QE terms, and 419 (on average 8.6 per topic) from the same number of "terms out of context" forms. There was an overlap of 164 terms selected from both "terms in context" and "terms out of context" forms.

3.3. Query expansion based on clarification forms

In this section we investigate the usefulness of link-terms and short contexts surrounding them (snippets) in interactive query expansion. The following hypotheses were studied:

Hypothesis 1. Link-terms selected by users for query expansion from clarification forms with or without context lead to significant performance improvements over the baseline run.

Hypothesis 2. Showing QE terms to the user in the context of short text snippets helps them select better terms than showing terms without context.

Hypothesis 3. If the user assesses the usefulness of short snippets, better performance can be achieved when all terms are extracted from the user-selected snippets and used in QE, compared to the method whereby QE terms are selected using the Offer Weight from the complete documents whose snippets were selected by the user (Fig. 2).

Hypothesis 4. If the user assesses the snippet as useful, and then judges the whole document as relevant, better performance can be achieved by using only the snippet in QE (Fig. 2).

A stronger sub-hypothesis of Hypothesis 4 is

Hypothesis 4.1. Query expansion using all terms from the user-selected snippets, regardless of whether the corresponding documents are judged relevant or not, can lead to better performance than QE from the full-texts of documents judged relevant (Fig. 2).

[4] One topic had only one query term in the Title, therefore no link-terms were offered in the clarification forms (link-terms can only be extracted from documents which have at least 2 distinct query terms).

Table 1
The baseline and query expansion runs

Run name	MAP	P@10	R-Precision
Unexpanded	0.1691	0.3580	0.2382
Link-terms-Context [$R = 25$]	0.2323	0.4780	0.2834
Link-terms-Context [$R =$ SelectedDocs]	0.2191	0.4920	0.2693
Link-terms-noContext [$R = 25$]	0.2372	0.4220	0.2806
Snippets [$R = 25$]	0.2588	0.5600	0.3003
Snippets [$R =$ SelectedDocs]	0.2390	0.5360	0.2859
OW-nQEterms [$R = 25$]	0.2317	0.4460	0.2709
OW-25QEterms [$R = 25$]	0.2268	0.4300	0.2674
OW-nQEterms [$R =$ SelectedDocs]	0.2286	0.5000	0.2691
OW-25QEterms [$R =$ SelectedDocs]	0.2271	0.4880	0.2672

All terms selected by the user from the "terms out of context" clarification form were added to the original query terms to create the expanded run *Link-terms-noContext*. Several runs were conducted based on the "terms in context" clarification form. In these runs either the link-terms (runs *Link-terms-Context*), or all snippet-terms (runs *Snippets*) are added to the original query to create the expanded query. BM25[5] with different values for the parameter R (number of known relevant documents) was used for the runs. We experimented with

- using all top 25 documents for QE term weighting (runs marked with "$R = 25$") and
- using only those documents for weighting, whose snippets were selected by the users (runs marked with "$R =$ SelectedDocs").

The motivation for evaluating these two different weighting approaches comes from a study by Lam-Adesina and Jones (2001), who demonstrated that extracting QE terms from a smaller set of top-ranked documents, and weighting them using a larger set of documents could lead to better results.

For comparison, several runs (prefixed with OW in Table 1) were conducted using the Offer Weight method (see Eqs. (3) and (4) above) of query expansion term selection (Robertson, 1990). In these runs all non-stopwords were extracted from the documents whose snippets were selected by users, ranked by OW and the top n ranked terms were added to the original query. In runs marked with "*nQEterms*" in Table 1, the same number of QE terms was used as in *Snippets* runs; in runs marked with "*25QEterms*" top-ranked 25 terms were used in QE.

User-selected link-terms from "terms in context" and "terms out of context" clarification forms led to statistically significant (t-test, $P < 0.05$) performance improvements over the baseline unexpanded run in all measures (except P@10 of *Link-terms-noContext* run) providing strong support for Hypothesis 1.

The use of terms selected from the "terms in context" clarification form (run *Link-terms-Context* [$R = 25$]) and terms selected from "terms out of context" clarification form (run *Link-terms-noContext* [$R = 25$]) led to very similar improvements in Mean Average Precision (MAP) and R-Precision compared to the unexpanded baseline run (*Unexpanded*). However, there is a substantial difference of 13% between their P@10. This suggests that showing terms to the user in the context of snippets leads to selection of better query expansion terms than showing them without context, providing some support for Hypothesis 2.

The use of all terms in the user-selected snippets (run *Snippets* [$R = 25$]) led to noticeably better performance in all measures compared to the use of only link-terms in the user-selected snippets (run *Link-terms-Context* [$R = 25$]). In particular, the P@10 of *Snippets* [$R = 25$] is 17.2% higher than that of *Link-terms-Context* [$R = 25$]. Note also that *Snippets* [$R = 25$] performs better than all other runs in all measures.

Another important result is that selection of QE terms using the OW method from the whole documents, whose snippets were selected by users (run *OW-nQEterms* [$R = 25$]) has lower performance in all measures than the use of all terms from user-selected snippets (*Snippets* [$R = 25$]). In particular, the P@10 of *OW-nQEterms* is

[5] Tuning constant k_1 (controlling the effect of within-document term frequency) was set to 1.2 and b (controlling document length normalisation) was set to 0.75 (Spärck Jones et al., 2000).

Table 2
Runs using only the documents whose snippets were selected and which were judged relevant

Run name	MAP	P@10	R-Prec
Rel-Snippets [$R = 25$]	0.2718	0.5980	0.3063
Rel-Snippets [$R =$ SelectedDocs]	0.2475	0.5720	0.2844
Rel-OW-nQEterms [$R = 25$]	0.2249	0.4200	0.2677
Rel-OW-25QEterms [$R = 25$]	0.2464	0.5140	0.2857
Rel-OW-nQEterms [$R =$ SelectedDocs]	0.2446	0.5700	0.2778
Rel-OW-25QEterms [$R =$ SelectedDocs]	0.2487	0.5940	0.2859

25.6% lower than that of *Snippets*. It also has lower performance in P@10 (by 7.2%) and R-Precision (by 4.6%) than the use of only link-terms taken from the user-selected snippets (*Link-terms-Context* [$R = 25$]). These results support Hypothesis 3, suggesting that in an interactive search scenario where the user indicates the relevance of the snippets extracted from documents, it is better to use only these representations for QE, rather than the whole documents.

In all of the runs discussed so far, QE terms were weighted using all 25 top-ranked documents (runs marked with [$R = 25$] in Table 1). The weighting of query expansion terms using only the documents whose snippets were selected (runs marked with [$R = SelectedDocs$]) leads to somewhat inconsistent results: on the one hand, it shows improvement in P@10 of *OW-nQEterms* [$R = SelectedDocs$] over *OW-nQEterms* [$R = 25$] (12.1%), and of *Link-terms-Context* [$R = SelectedDocs$] over *Link-terms-Context* [$R = 25$] (2.9%). On the other hand, it shows worse or similar performance in MAP and R-Precision of all runs, and in P@10 of *Snippets* [$R = SelectedDocs$] over *Snippets* [$R = 25$] (4.5%).

Interestingly, almost half of all user-selected snippets come from documents which were later judged by the same users as non-relevant. Altogether, users selected snippets from 224 documents, out of which 119 were later judged relevant, 102 were non-relevant and 3 were unjudged. It is interesting to see how query expansion following snippet selection compares to query expansion following full document relevance judgements. We replicated some of the runs in Table 1, but this time using only those documents, whose snippets were selected by the user and whose full-texts were later judged relevant (Table 2).

Comparison of the runs *Snippets* [$R = SelectedDocs$] (Table 1) and *Rel-Snippets* [$R = SelectedDocs$] (Table 2), indicates that using only the snippets from relevant documents improves P@10 by 6.7%, MAP by 3.5%, and slightly deteriorates R-Precision (by 0.5%). This suggests that user-selected snippets from non-relevant documents do not hurt performance much. On the other hand, runs using OW-selected terms from the whole documents show quite a different picture: compare *OW-25QEterms* [$R = SelectedDocs$] (Table 1) and *Rel-OW-25QEterms* [$R = SelectedDocs$] (Table 2). Extraction of the QE terms from only the relevant documents boosts performance on all measures: MAP improves by 9.5%, P@10 improves by 21.7% and R-Precision improves by 6.9%.

In the traditional IR scenario, the relevance feedback process consists of two steps: in the first step the user sees the document surrogate and assesses its relevance, in the second step, she assesses the relevance of the corresponding complete document. Query expansion or reweighting is normally based on the full-texts of the seen documents. Given the above scenario, let us suppose that the user clicks on the snippets she finds useful to see the full-texts of the corresponding documents and judge their relevance. We hypothesise (Hypothesis 4) that if the user assesses the snippet as useful, and then judges the whole document as relevant, better performance can be achieved by using only the snippet rather than the whole document in QE. Comparison of the best runs in Table 2 *Rel-Snippets* [$R = 25$] and *Rel-OW-25QEterms* [$R = SelectedDocs$] supports this hypothesis.[6] The MAP obtained by *Rel-Snippets* is 9.3% better than *Rel-OW-25QEterms*. Similarly, the former is

[6] Note that the users had a time gap of a few weeks between the stage in which they judged snippets for relevance, and the stage in which they judged pooled documents. As a result, some relevance criteria might have changed in this time period. Also, in a real-time interactive search scenario, knowing that they can see full documents, users may choose to click on a larger number of snippets to view the full documents, as opposed to the HARD track scenario, which may lead to somewhat different results. So, further experimentation in an interactive setting is needed to support the above result.

better by 7.1% compared to the latter in R-Precision, however P@10 is almost the same in both runs. This is an interesting and somewhat surprising result. The improvement may be largely due to the users' skill in identifying good terms. To test whether the improvement can be attributed to term selection skills of users, or whether there is benefit in restricting the term selection to snippets, we have conducted Automatic Query Expansion experiments, which are reported in the next section.

To test Hypothesis 4.1 we compared the best run using only the selected snippets (*Snippets* [$R = 25$] in Table 1) and the best run using the full-text of the documents judged relevant (*Rel-OW-25QEterms* [$R =$ *SelectedDocs*] in Table 2). The results indicate that better MAP and R-Precision are obtained when only snippets assessments are used in QE, providing some support for Hypothesis 4.1. P@10, however, benefits from selecting QE terms from full-documents by 6%.

To sum up the above, the following conclusions can be made:

- Showing QE terms to the users in the context of short one-line snippets helps them select better terms than showing terms without context.
- In the interactive QE scenario, whereby the user sees and evaluates for usefulness short document representations (snippets), better performance is obtained when all QE terms are taken from these representations, rather than selected from the complete documents they represent.
- If the user is given an option to evaluate the relevance of the full document, whose snippet she found useful, its use in QE may not lead to higher performance. In our experiments QE terms selected from whole documents judged relevant led to poorer R-Precision and MAP, but higher P@10, than QE terms taken from the user-selected snippets.

The next two sections describe experiments conducted to evaluate the effectiveness of lexical links-based QE methods in batch mode using the HARD 2005 test collection. Two sets of experiments were conducted: (1) relevance feedback, where all known relevant documents in the top 25 retrieved documents were used for QE (Section 3.4), and (2) blind feedback, where all top 25 retrieved documents were used for QE (Section 3.5).

3.4. Query expansion using relevance feedback data

In this section, we describe experiments conducted in order to evaluate the effectiveness of link-terms and snippet-terms in automatic QE. The following hypothesis was formulated:

Hypothesis 5. Link-terms and/or snippet-terms extracted from known relevant documents lead to comparable or better performance in QE than terms extracted from whole documents.

In the experiments various combinations of the following parameters are used:

- number of link-terms per document used in selecting snippets;
- number of snippets per document used for extracting QE terms;
- size of the snippets;
- number of QE terms added to the original query.

All the conducted runs in this section were evaluated using the residual rank evaluation method. In this method those documents which are used for QE are removed from the ranked set of retrieved documents before evaluation to avoid the ranking effect, whereby documents from which QE terms are extracted are promoted to the top of the ranked list, thus artificially boosting precision. Residual rank results, therefore, show the effectiveness of the QE methods in retrieving *new* documents. We also report the results of retrospective evaluation that uses the complete document collection. All conducted retrospective and residual runs are reported in Table A3 in Appendix, and selected runs, which are discussed here, are presented in Table 3. The benchmark, against which the effectiveness of the links-based QE methods is compared, is the OW QE term selection method from the full-texts of documents (Robertson, 1990).

Table 3
Selected relevance feedback runs

Runs	Residual rank evaluation			Retrospective evaluation		
	MAP	P@10	R-Prec	MAP	P@10	R-Prec
OW_25QEterms	0.2348	0.5260	0.2866	0.3250	0.7160	0.3491
OW_50QEterms	0.2428	0.5560	0.2911	0.3324	0.7340	0.3522
Snippet-2Col-allSnip-25QEterms	0.2409	0.5780	0.2899	0.3303	0.7360	0.3514
Snippet-2Col-allSnip-50QEterms	0.2439	0.5620	0.2965	0.3342	0.7340	0.3615

For the snippet-based runs query expansion terms are selected as follows:

Step 1: All link-terms in each document in the relevant set are ranked by *idf*; x top-ranked link-terms are then selected per document.
Step 2: Snippets of size $2w + 1$ words (where w is the number of words on either side of the link-term) are identified for each link-term selected in Step 1, ranked by average *idf* of the terms in the snippet, and y top-ranked snippets are selected.
Step 3: All non-stop terms are extracted from the snippets selected in Step 2, ranked by OW, and top z used for QE.

Tables A1 and A2 in Appendix contain detailed descriptions of the runs. The results of the residual rank experiments (Table A3 in Appendix) show that the best P@10 is achieved by the run *Snippet-2Col-allSnip-25QEterms* (size = 7), where terms from all snippets of the two top-ranked link-terms per document are extracted and ranked by OW, and then 25 top-ranked terms are used in QE. The P@10 of this run is about 4% higher than the best performing baseline run (*OW_50QEterms*, where 50 OW-ranked terms are extracted from the full-text of relevant documents). P@10 of this run is also higher by about 10% than the baseline OW run with the same number (25) of query terms (*OW_25QEterms*). The performance of the baseline and runs that use only link-terms (prefixed as *LinkCols* in Table A3) are also very similar in terms of MAP and R-Precision, but slightly lower in P@10 (Table A3).

The above results indicate that there is some benefit in selecting QE terms from short text snippets (extracted from known relevant documents), rather than the full-texts of the relevant documents, providing some support for Hypothesis 5. The results in Table A3 demonstrate that link-terms and snippet-terms are useful QE terms, which indicates they are highly relevant to the query topic and capable of retrieving other relevant documents. The results suggest that practically there is no need to extract terms from parts of the relevant documents other than surrounding the link-terms; in fact there appears to be some benefit in selecting terms from snippets. The presented results show that effectiveness of link-based QE methods is comparable to and in some cases better than full-text based QE methods in terms of three standard measures of retrieval performance (P@10, MAP, and R-Precision), which is an indication of the relevance discrimination value of link- and snippet-terms. The results of the blind feedback experiments reported in the next section give further evidence that supports the conclusions drawn above.

3.5. Query expansion using blind feedback data

Experiments with query expansion following blind feedback were conducted in order to evaluate how the proposed QE techniques perform in the absence of relevance feedback information, which is the case with most real-world IR applications, specifically the Web search engines. As in the relevance feedback experiments reported in the previous section, the runs based on selecting QE terms from snippets were compared against the runs using QE terms from complete documents. The following hypothesis was formulated:

Hypothesis 6. Link-terms and/or snippet-terms extracted from top retrieved documents *assumed* to be relevant lead to comparable or better performance in QE than terms extracted from whole documents.

Table 4
Selected blind feedback runs

Runs	Residual rank evaluation			Retrospective evaluation		
	MAP	P@10	R-Prec	MAP	P@10	R-Prec
OW_25QEterms	0.1738	0.3340	0.2226	0.2110	0.3800	0.2555
LinkCols_25QEterms	0.1836	0.3980	0.2272	0.2213	0.4020	0.2584

For all the runs, top 25 documents retrieved by the baseline run *Unexpanded* were used. For blind feedback we used both retrospective and residual ranking evaluation methods. Blind feedback (BF) is perhaps a better approach for testing retrieval effectiveness in operational settings, where relevance feedback mechanism is usually not available. BF can be performed immediately following the baseline retrieval, so that the user sees only the result of the blind feedback run. For this reason, there is no need to control the ranking effect as in relevance feedback. We also report the results of the residual ranking evaluation for consistency with the analysis in the previous section on relevance feedback, and to have an understanding of how well the methods used in the experiments rank relevant documents other than in the top 25, which were used for QE. In the residual ranking evaluation the top 25 documents, which were used for BF, were removed from the results before evaluation.

Retrospective evaluation results (Table A4, Appendix) suggest that some gains in performance can be achieved by extracting QE terms from snippets and/or link-terms as opposed to whole documents. In particular, the P@10 of the run *LinkCols_25QEterms*, which uses link-terms only in QE, is about 5.8% higher than P@10 of the run *OW_25QEterms*. The latter uses QE terms extracted from the whole documents. Smaller gains are obtained in other measures when QE terms are extracted from snippets or link-terms, instead of the whole documents. Larger gains are observed in residual rank results (Table A4, Appendix). For instance, 19.2% improvement is achieved in P@10 when QE terms are selected from link-terms (*LinkCols_25QEterms*), instead of whole documents (*OW_25QEterms*). The results of these two runs are presented in Table 4.

As the results in Table A4 demonstrate, the methods of selecting QE terms from link-terms or snippet-terms led to performance comparable to and in some runs higher than QE term selection from whole documents. The observed gains, although not statistically significant, provide some support for Hypothesis 6. The conclusion that can be drawn from these findings is that link-terms and snippet-terms in a document have strong semantic relatedness to the query, supporting the conclusion reached in the preceding section based on the relevance feedback experiments.

4. Conclusions

In this paper we have introduced new query expansion methods: instead of selecting query expansion terms from entire documents, we propose to select QE terms from only those terms that form lexical links between the contexts of distinct query terms (link-terms) or terms that surround the link-terms (snippet-terms). The proposed QE methods were evaluated in interactive and automatic query expansion scenarios in comparison to the selection of query expansion terms from the full-texts of documents.

The interactive query expansion experiments described investigated several aspects of relevance feedback: presentation of QE terms to the user in and out of the context of short text snippets; the use of only the terms from user-selected snippets vs. the use of all terms from their documents as candidates for QE; the value of eliciting document-level relevance judgements vs. only snippet-level relevance judgements.

The results indicate that link-terms selected by users from ''terms in context'' and ''terms out of context'' clarification forms lead to considerable performance improvements over the unexpanded run in all measures, providing strong support for Hypothesis 1. Showing terms in context leads to a substantially higher P@10 (but not MAP or R-Precision) than without context, providing some support for Hypothesis 2.

Selection of QE terms using the OW method from the whole documents, whose snippets were selected by users has lower performance in all measures than the use of all terms from user-selected snippets, supporting Hypothesis 3. This means that in the interactive IR scenario, where the user evaluates the usefulness of short document representations (snippets), better performance is obtained when QE terms are extracted from only these representations, rather than the complete documents they represent.

Extracting QE terms from user-selected snippets, the whole documents of which were judged relevant, yields better performance (in MAP and R-Precision) in comparison to selecting from the full-texts of the relevant documents, providing support for Hypothesis 4. Comparison of the run using all selected snippets with the run using the full-texts of documents, whose snippets were selected, and which were judged relevant, demonstrates that better MAP and R-Precision are obtained when only snippet-level judgements are used, providing some support for Hypothesis 4.1.

We also evaluated the developed QE methods in automatic query expansion in relevance and blind feedback scenarios, comparing them to query expansion from the full-texts of documents. Overall, performance of link-terms and snippet-terms is very close to that of the QE terms extracted from the full-texts in the relevance feedback experiments. Some runs using snippet-terms show small gains in P@10, suggesting that there is some benefit in selecting QE terms from short text snippets extracted from known relevant documents rather than the full-texts of the relevant documents, thereby providing some support for Hypothesis 5. In blind feedback experiments, the use of link-terms and snippet-terms also leads to comparable performance to the QE terms extracted from the full-texts. Performance gains are observed for link-term based runs in P@10, providing some support for Hypothesis 6.

A study by Vechtomova et al. (2006) has discovered that there exists a significant difference in the number of lexical links between distinct query terms in relevant and non-relevant document sets. Based on this finding, it was hypothesised that query terms in relevant documents tend to occur in related contexts more (i.e., have on average more lexical links) than non-relevant documents. Ranking experiments using link counts between distinct query terms' contexts reported in the above paper showed some improvements over the standard BM25 ranking (without relevance information). The study described here gives further evidence that link-forming terms are highly related to query topics and capable of retrieving other relevant documents, as demonstrated by their good performance in QE experiments. This provides indirect support to the hypothesis put forward in Vechtomova et al. (2006) that link-terms are good indicators of whether query terms occur in related or unrelated contexts, and consequently whether the document is likely to be relevant or not.

To evaluate further the usefulness of link- and snippet-terms in QE, it is necessary to compare them to other types of document representations (e.g., sentences or summaries). It will also be useful to compare their effectiveness against other QE term selection methods.

Acknowledgements

We would like to thank Susan Jones (City University, London) and anonymous referees for their valuable comments and suggestions.

Appendix

See Tables A1–A4.

Table A1
Descriptions of snippet-based query expansion runs

Run name	Number of link-terms (x)	Number of snippets (y)	Number of OW-ranked QE terms (z)
Snippet-2Col-1Snip-allQEterms	2	1	All
Snippet-2Col-1Snip-25QEterms	2	1	25
Snippet-2Col-allSnip-25QEterms	2	All	25
Snippet-2Col-allSnip-50QEterms	2	All	50
Snippet-2Col-2Snip-25QEterms	2	2	25
Snippet-2Col-2Snip-50QEterms	2	2	50
Snippet-allCol-1Snip-25QEterms	All	1	25
Snippet-allCol-1Snip-50QEterms	All	1	50
Snippet-allCol-allSnip-25QEterms	All	All	25
Snippet-allCol-allSnip-50QEterms	All	All	50
Snippet-allCol-2Snip-25QEterms	All	2	25
Snippet-allCol-allSnip-50QEterms	All	2	50

Table A2
Descriptions of query expansion runs

Run name	Number of QE terms	QE terms are selected from	QE term selection method
OW_nQEterms	The same as in Snippet-2Col-1Snip-allQEterms run	All non-stop terms in the (pseudo-) relevance set	OW
OW_25QEterms	25	All non-stop terms in the (pseudo-) relevance set	OW
OW_50QEterms	50	All non-stop terms in the (pseudo-) relevance set	OW
LinkCols_25QEterms	25	All link-terms in the (pseudo-) relevance set	OW
LinkCols_50QEterms	50	All link-terms in the (pseudo-) relevance set	OW
LinkCols_AllQEterms	All	All link-terms in the (pseudo-) relevance set	–
LinkCols_AllQEterms_top2perDoc	All	All link-terms in the (pseudo-) relevance set	Top 2 idf-ranked link-terms per document
LinkCols_25QEterms_freq	25	All link-terms in the (pseudo-) relevance set	Frequency in the windows around user's query terms
LinkCols_25QEterms_freq	50	All link-terms in the (pseudo-) relevance set	Frequency in the windows around user's query terms
LinkCols_AllQEterms_freq_top2perDoc	All	All link-terms in the (pseudo-) relevance set	Frequency in the windows around user's query terms. Top 2 ranked terms are selected per document

Table A3
Relevance feedback run results

Runs	Residual rank evaluation			Retrospective evaluation		
	MAP	P@10	R-Prec	MAP	P@10	R-Prec
Unexpanded	0.1011	0.0220	0.1650	0.1691	0.3580	0.2382
OW_nQEterms	0.2399	0.5380	0.2894	0.3316	0.7220	0.3537
OW_25QEterms	0.2348	0.5260	0.2866	0.3250	0.7160	0.3491
OW_50QEterms	0.2428	0.5560	0.2911	0.3324	0.7340	0.3522
LinkCols_25QEterms	0.2429	0.5180	0.2928	0.3309	0.7000	0.3594
LinkCols_50QEterms	0.2388	0.5180	0.2889	0.3287	0.7000	0.3553
LinkCols_AllQEterms	0.2340	0.5100	0.2828	0.3245	0.7060	0.3516
LinkCols_AllQEterms_top2perDoc	0.1955	0.4200	0.2459	0.2871	0.6440	0.3136
LinkCols_25QEterms_freq	0.2220	0.4860	0.2705	0.3089	0.6620	0.3354
LinkCols_25QEterms_freq	0.2289	0.5000	0.2764	0.3178	0.6960	0.3445
LinkCols_AllQEterms_freq_top2perDoc	0.1979	0.4480	0.2533	0.2816	0.6180	0.3146
Snippets size = 7						
Snippet-2Col-1Snip-allQEterms	0.2257	0.5400	0.2758	0.3227	0.7200	0.3457
Snippet-2Col-1Snip-25QEterms	0.2297	0.5120	0.2806	0.3201	0.6940	0.3452
Snippet-2Col-allSnip-25QEterms	0.2409	0.5780	0.2899	0.3303	0.7360	0.3514
Snippet-2Col-allSnip-50QEterms	0.2439	0.5620	0.2965	0.3342	0.7340	0.3615
Snippet-2Col-2Snip-25QEterms	0.2378	0.5380	0.2888	0.3272	0.7300	0.3485
Snippet-2Col-2Snip-50QEterms	0.2336	0.5420	0.2837	0.3270	0.7400	0.3474
Snippet-allCol-1Snip-25QEterms	0.2435	0.5480	0.2970	0.3331	0.7160	0.3611
Snippet-allCol-1Snip-50QEterms	0.2481	0.5520	0.3005	0.3379	0.7180	0.3633
Snippet-allCol-allSnip-25QEterms	0.2397	0.5380	0.2852	0.3292	0.7160	0.3497
Snippet-allCol-allSnip-50QEterms	0.2489	0.5480	0.2993	0.3379	0.7420	0.3614
Snippet-allCol-2Snip-25QEterms	0.2432	0.5420	0.2884	0.3328	0.7260	0.3514
Snippet-allCol-allSnip-50QEterms	0.2455	0.5420	0.2964	0.3350	0.7280	0.3586
Snippets size = 11						
Snippet-2Col-1Snip-allQEterms	0.2255	0.5320	0.2763	0.3236	0.7420	0.3465
Snippet-2Col-1Snip-25QEterms	0.2318	0.5180	0.2812	0.3201	0.7020	0.3416
Snippet-2Col-allSnip-25QEterms	0.2365	0.5400	0.2866	0.3270	0.7160	0.3488
Snippet-2Col-allSnip-50QEterms	0.2448	0.5480	0.2933	0.3350	0.7340	0.3587
Snippet-2Col-2Snip-25QEterms	0.2375	0.5460	0.2879	0.3270	0.7280	0.3513
Snippet-2Col-2Snip-50QEterms	0.2376	0.5580	0.2888	0.3287	0.7400	0.3528
Snippet-allCol-1Snip-25QEterms	0.2361	0.5260	0.2838	0.3254	0.7140	0.3464
Snippet-allCol-1Snip-50QEterms	0.2433	0.5480	0.2927	0.3325	0.7180	0.3559
Snippet-allCol-allSnip-25QEterms	0.2363	0.5300	0.2835	0.3265	0.7180	0.3463
Snippet-allCol-allSnip-50QEterms	0.2473	0.5540	0.2942	0.3361	0.7320	0.3568
Snippet-allCol-2Snip-25QEterms	0.2389	0.5360	0.2850	0.3282	0.7180	0.3461
Snippet-allCol-allSnip-50QEterms	0.2470	0.5560	0.2953	0.3358	0.7320	0.3564

Table A4
Blind feedback run results

Runs	Residual rank evaluation			Retrospective evaluation		
	MAP	P@10	R-Prec	MAP	P@10	R-Prec
Unexpanded	0.1251	0.2420	0.1901	0.1691	0.3580	0.2382
OW_nQEterms	0.1580	0.3360	0.2049	0.1952	0.3940	0.2395
OW_25QEterms	0.1738	0.3340	0.2226	0.2110	0.3800	0.2555
OW_50QEterms	0.1594	0.3340	0.2107	0.1960	0.3580	0.2318
LinkCols_25QEterms	0.1836	0.3980	0.2272	0.2213	0.4020	0.2584
LinkCols_50QEterms	0.1781	0.3760	0.2209	0.2138	0.3980	0.2518
LinkCols_AllQEterms	0.1733	0.3620	0.2192	0.2106	0.3920	0.2506
LinkCols_AllQEterms_top2perDoc	0.1674	0.3360	0.2060	0.2053	0.3920	0.2452
LinkCols_25QEterms_freq	0.1619	0.3340	0.2060	0.1921	0.3400	0.2324
LinkCols_25QEterms_freq	0.1711	0.3440	0.2156	0.2037	0.3540	0.2448
LinkCols_AllQEterms_freq_top2perDoc	0.1682	0.3640	0.2109	0.2037	0.3840	0.2398
Snippets size = 7						
Snippet-2Col-1Snip-allQEterms	0.1713	0.3660	0.2124	0.2062	0.3900	0.2532
Snippet-2Col-1Snip-25QEterms	0.1711	0.3360	0.2135	0.2100	0.3980	0.2541
Snippet-2Col-allSnip-25QEterms	0.1709	0.3560	0.2169	0.2074	0.3880	0.2445
Snippet-2Col-allSnip-50QEterms	0.1463	0.2360	0.2003	0.1993	0.3660	0.2456
Snippet-2Col-2Snip-25QEterms	0.1701	0.3460	0.2158	0.2086	0.3720	0.2455
Snippet-2Col-2Snip-50QEterms	0.1688	0.3420	0.2173	0.2051	0.3760	0.2475
Snippet-allCol-1Snip-25QEterms	0.1719	0.3580	0.2221	0.2114	0.3940	0.2473
Snippet-allCol-1Snip-50QEterms	0.1686	0.3500	0.2152	0.2031	0.3960	0.2407
Snippet-allCol-allSnip-25QEterms	0.1666	0.3380	0.2113	0.2056	0.3860	0.2472
Snippet-allCol-allSnip-50QEterms	0.1598	0.3300	0.2064	0.1963	0.3740	0.2308
Snippet-allCol-2Snip-25QEterms	0.1552	0.2700	0.2036	0.2098	0.3920	0.2516
Snippet-allCol-allSnip-50QEterms	0.1468	0.2400	0.1988	0.2000	0.3860	0.2435
Snippets size = 11						
Snippet-2Col-1Snip-allQEterms	0.1668	0.3600	0.2148	0.2004	0.3820	0.2468
Snippet-2Col-1Snip-25QEterms	0.1700	0.3500	0.2147	0.2072	0.3800	0.2466
Snippet-2Col-allSnip-25QEterms	0.1700	0.3380	0.2180	0.2064	0.3860	0.2494
Snippet-2Col-allSnip-50QEterms	0.1467	0.2320	0.1960	0.1997	0.3620	0.2411
Snippet-2Col-2Snip-25QEterms	0.1717	0.3540	0.2173	0.2085	0.3800	0.2477
Snippet-2Col-2Snip-50QEterms	0.1683	0.3340	0.2161	0.2030	0.3660	0.2460
Snippet-allCol-1Snip-25QEterms	0.1698	0.3520	0.2176	0.2111	0.3920	0.2487
Snippet-allCol-1Snip-50QEterms	0.1669	0.3460	0.2109	0.2031	0.4000	0.2417
Snippet-allCol-allSnip-25QEterms	0.1702	0.3420	0.2166	0.2030	0.3660	0.2460
Snippet-allCol-allSnip-50QEterms	0.1606	0.3240	0.2084	0.2077	0.3880	0.2507
Snippet-allCol-2Snip-25QEterms	0.1539	0.2780	0.2027	0.2087	0.3840	0.2483
Snippet-allCol-allSnip-50QEterms	0.1454	0.2480	0.1944	0.1986	0.3760	0.2385

References

Allan, J. (2005). HARD, Track Overview in TREC 2005. (Notebook). High Accuracy Retrieval from Documents. In E. Voorhees, & L. Buckland (Eds.), TREC 2005 notebook proceedings. Gaithesburg, MD, United States.

Beaulieu, M., & Jones, S. (1998). Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting with Computers, 10*, 237–248.

Billerbeck, B., & Zobel, J. (2004). Questioning query expansion: an examination of behaviour and parameters. In *Proceedings of the 15th Australasian Database Conference, Dunedin, New Zealand* (pp. 69–76).

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Lam-Adesina, A., & Jones, G. (2001). Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States* (pp. 1–9).

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

Robertson, S. E., Walker, S., Jones, S., Hankock-Beaulieu, M., & Gatford, M. (1995). Okapi at TREC-3. In D. Harman (Ed.), *Proceedings of the third text retrieval conference* (pp. 109–126). Gaithersburg, US: NIST.

Spärck 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).

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., & Robertson, S. E. (2005). A study of document relevance and lexical cohesion between query terms. In *Proceedings of the workshop on methodologies and evaluation of lexical cohesion techniques in real-world applications (ELECTRA 2005), the 28th ACM SIGIR conference, Salvador, Brazil* (pp. 18–25).

Vechtomova, O., & Karamuftuoglu, M. (2006). Elicitation and use of relevance feedback information. *Information Processing and Management, 42*(1), 191–206.

Vechtomova, O., Karamuftuoglu, M., & Robertson, S. E. (2006). On document relevance and lexical cohesion between query terms. *Information Processing and Management, 42*(5), 1230–1247.

White, R. W., Jose, J. M., & Ruthven, I. (2005). Using top-ranking sentences to facilitate effective information access. *Journal of the American Society for Information Science and Technology, 56*(10), 1113–1125.

Olga Vechtomova is an Assistant Professor in the Department of Management Sciences, University of Waterloo, Canada. She received Ph.D. in Information Science from City University, London in 2001. Her research interests are in Information Retrieval, Natural Language Processing applications for IR and user interaction with IR systems. Her works are published in Information Retrieval, Information Processing and Management and Journal of Information Science (http://ovecht2.uwaterloo.ca).

Murat Karamuftuoglu is an Assistant Professor in the Computer Engineering Department, Bilkent University, Turkey. He received Ph.D. in Information Science from City University, London in 1998. His research interests are in Interactive Information Retrieval, Knowledge Management and Computer Mediated Communication. His works are published in Journal of the American Society for Information Science and Technology, Journal of Information Science, Information Processing and Management.