

# Generation of Query-biased Concepts Using Content and Structure for Query Reformulation

Youjin Chang<sup>1</sup>, Jun Wang<sup>1</sup>, Mounia Lalmas<sup>1</sup>

<sup>1</sup> Queen Mary, University of London,  
London, E1 4NS, UK  
{youjinchang, wangjun, mounia}@dcs.qmul.ac.uk

**Abstract.** This paper proposes an approach for query reformulation based on the generation of appropriate query-biased concepts. Query-biased concepts are generated from retrieved documents using their content and structure. In this paper, we focus on three aspects of the concept generation; the selection of query-biased concepts from retrieved documents, the effect of the structure, and the number of retrieved documents used for generating the concepts.

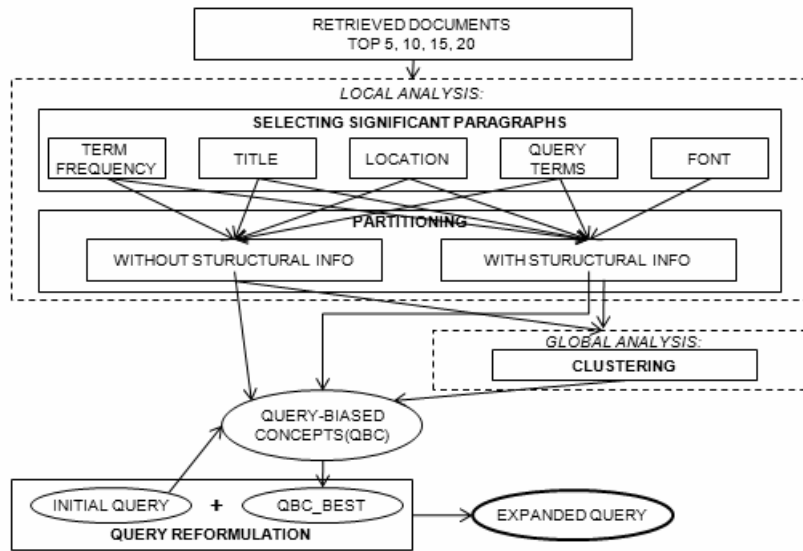
**Keywords:** query reformulation, feature extraction, concept generation, structure, relevance feedback

## 1 Motivation

A main issue in information retrieval (IR) is for users to define queries, i.e. the query terms, that properly express their information needs. If we assume that IR engines successfully find all the relevant documents using the terms contained in the initial query, the remaining problem is how to properly formulate the query. In IR, users often need to reformulate their initial queries more than once to obtain better results. There has been wide interest in the selection of terms to be reformulate the query [1,4,5]. There are three main approaches: approaches based on relevance feedback information from the user, approaches based on information derived from the set of documents initially retrieved, and approaches based on global information derived from the document collection. The first approach, query reformulation from relevance feedback, has been shown effective if appropriate feedback (i.e. explicit - this document is/is not relevant; or implicit - through click-through data) is given by the user. This paper is concerned with the first type of approach. In this paper, we propose a query reformulation process based on so-called query-biased concepts (QBC). This process is performed as one of the relevance feedback task. We try to enhance the initial query with query-biased concepts generated from the analysis of the content and structural information of documents retrieved by the initial query.

We assume that the retrieved documents have several topics or themes that can be expressed by a set of terms. For example, let us consider an article about 'speech recognition'. The article may discuss the definition of speech recognition, the history of speech recognition, a speech recognition case study, etc. It is necessary to select the themes of the article so that the article can be effectively represented. Furthermore, it

is also necessary to find “overall” concepts by joining those themes that are related. Since there may be similar documents or paragraphs about a ‘speech recognition case study’ in other documents, we need to integrate those themes across the documents globally. Through these local and global analyses, we aim to construct the concepts that identify the main themes of retrieved documents. The framework for constructing the query-biased concepts is illustrated in Figure 1.



**Fig. 1.** The procedure of experiments to construct the query-biased concepts

In the local analysis, we select significant characteristics from each (retrieved and relevant) document and name them ‘features’. This is done by selecting the significant paragraphs and partitioning those paragraphs. We use the following criteria for scoring each paragraph according to its significance: 1) the ratio of significant terms in a paragraph: the terms that frequently occur in a document are arranged in a significance term list; 2) the location of paragraph; 3) the presence of a title of the document within the paragraph; 4) the presence of query terms within the paragraph; 5) the presence of bold or italic term within the paragraph. The top ranked  $k$  paragraphs are chosen as the significant paragraphs. We then partition the selected paragraphs. Through partitioning, the features of each document are generated. It is important to make the selected significant characteristics orthogonal to each other within a document, because orthogonal features are able to represent the main themes of a document separately. In this paper, we extend the framework to deal with structured documents.

Nowadays, with the increased number of documents formatted in the eXtensible Markup Language (XML), it makes sense to investigate whether the structure, as captured by XML, can also be used to generate useful query-biased concepts. We suggest using the structural relations between paragraphs for partitioning. The

paragraphs belonging to the same section or subsection can be partitioned. Through this restriction, we can reduce the non-desirable unification caused by common terms and/or specific terms like title terms or query terms within one document. Depending on the level of partitioning, the number of features in a document can be increased or decreased. The results with different levels of partitioning are presented in section 3.

After a local analysis, it is necessary to integrate these features across all the documents to build the concepts. We adopt a single pass method based in early work on clustering analysis [2]. The main purpose of this step, the global analysis, is to prevent the duplication of similar features. The clustering makes it possible to generate the primitive concepts that are approximately orthogonal. Analyzing a set of documents locally and globally has been used in previous studies [1, 4]. The final stage of constructing the query-biased concepts is to combine the generated concepts with the initial query. We compute the similarity of between the initial query and concepts. The concept that has the maximum similarity with the initial query is selected as the best query-biased concepts ( $QBC_{best}$ ). For a new query, the original query terms are expanded with those associated terms in  $QBC_{best}$ . Finally, the new query is resubmitted to the retrieval system.

## 2 Experimental set up

We use the test collection developed at INEX 2005 [3], which consists of a set of XML documents, topics and relevance assessments. The document collection is made up of the full-texts, marked up in XML, of 16,819 articles of the IEEE Computer Society's publications. Generally, one article consists of a front matter (<fm>), body (<body>) and back matter (<bm>). The opening and closing tags enclose the main content, which is structured into sections (<sec>), subsections (<ss1>) and sub-subsections (<ss2>, <ss3>). Each of these logical units starts with a title followed by a number of paragraphs (<p>). We use the 23 content-only topics provided by INEX, as we are focusing on document retrieval. The <title> part of the topic is used as an initial query. Although the relevance assessment in INEX is done at element level, we can derive the assessment at document level. Any document that has any relevant content for a query is set as relevant to that query.

All experiments are performed using the HySpirit [7] retrieval system. We carry out a number of query reformulation experiments in the context of a relevance feedback scenario. First, the documents are retrieved using the initial query. To examine the impact of sampling a subset of the top ranked documents, we restrict the set of returned documents to the top 5, 10, 15, and 20 documents, respectively (as this reflects more realistic scenarios). Then we assess the relevance of retrieved documents. Practically, the user's relevance judgment is the most accurate but we use the relevance assessments provided by INEX 2005 to simulate the feedback process. With the selected documents (the retrieved and relevant documents), we construct the query-biased concepts with various approaches. The  $QBC_{CO}$  approach constructs the query-biased concepts with content information only. Neither the font information nor the structural information is used. The  $QBC_{CS}$  approach applies all the

techniques with structural information to construct the concepts. Finally, the query-biased concepts from both approaches are combined with the initial query.

We examine four different types of results: one baseline, one pseudo-relevance feedback (PRF), one classis relevance feedback (RF), and various QBC approach (QBC\_CO and QBC\_CS). For the baseline, we use a traditional  $tf*idf$  ranking. For PRF and RF, we use Rocchio's formula [6], where we use the top 5, 10, 15, and 20 retrieved documents of the baseline. For PRF, we use all such documents, whereas for RF, we use those that are relevant. We only use a positive feedback strategy (we only consider relevance), and choose the top 20 terms to expand the query. QBC\_CO and QBC\_CS are also performed with the top 5, 10, 15, and 20 retrieved documents of the baseline. For QBC\_CS, we considered a hierarchical relation between paragraphs such as sections and subsections in the partitioning step. We partition the paragraphs belonging to the same section (QBC\_CS\_SEC), subsection (QBC\_CS\_SS1), or sub-subsection (QBC\_CS\_SS2). We also choose the top 20 terms to form the new query. Finally, we evaluate the results with the full freezing method [8]. There, the rank positions of the top  $n$  documents ( $n = 5, 10, 15, \text{ and } 20$ ), the ones used to modify the query, are frozen. The remaining documents are re-ranked.

### 3 Results and analysis

For space reason, we only compare the results using mean average precision (MAP) over the whole ranking, which we calculate using `trec_eval`. Table 1 shows all the results. The results of PRF are inferior to the baseline. Since PRF assumes that all top  $n$ -ranked documents ( $n = 5, 10, 15, 20$ ) are relevant, this indicates that we need to use relevance information to find appropriate terms for expanding the query. Although the results of RF are higher than the PRF, they are still lower than the baseline. This indicates that we need better techniques to extract appropriate terms. We can see that QBC\_CO and QBC\_CS outperform the baseline, PRF, and RF. In the QBC\_CO approaches, using both local and global analysis such as summarization, partitioning, and clustering shows the best result. In the QBC\_CS approaches, the cases without global analysis (i.e., clustering) show the best result. Here, we only report the best results of QBC\_CS where we partition the paragraphs belonging to the same section and do not apply any clustering. We discuss the other results in Table 3.

**Table 1.** Mean average precision (MAP) of baseline, PRF, RF, and QBC runs. TOP5, TOP10, TOP15, and TOP20 represent the number of retrieved documents.

| Baseline |       | PRF    | RF     | QBC_CO | QBC_CS |
|----------|-------|--------|--------|--------|--------|
| 0.2073   | TOP5  | 0.1793 | 0.2056 | 0.2369 | 0.2325 |
|          | TOP10 | 0.1793 | 0.2049 | 0.2354 | 0.2211 |
|          | TOP15 | 0.1815 | 0.2057 | 0.2380 | 0.2190 |
|          | TOP20 | 0.1717 | 0.2045 | 0.2376 | 0.2193 |

It is known that the success of a query reformulation process depends on how the initial query performs. We thus classify the 23 topics into two groups: poor (P), and good (G) performing queries. We investigate whether the QBC approaches are

particularly effective in the case of the poorly performing queries. The good/poor decision is based on the MAP achieved by our baseline. If the MAP of the query is above 0.2073, we consider the query to be good. 14 queries with the MAP under 0.2073 are identified to be poor. For simplicity, we only compare the results of RF and QBC approaches with respect to the baseline in Table 2. Due to space limitation, we choose the same results of QBC\_CO and QBC\_CS approaches with TOP5 in Table 1 again to compare the retrieval performance for the two types of queries.

**Table 2.** Retrieval Performance of QBC\_CO and QBC\_CS runs in poor(P) and good(G) performing queries. There are 14 poorly performing queries and 9 good performing queries.

|             | Baseline |        | RF     |        | QBC_CO |        | QBC_CS |        |
|-------------|----------|--------|--------|--------|--------|--------|--------|--------|
|             | P(14)    | G(9)   | P(14)  | G(9)   | P(14)  | G(9)   | P(14)  | G(9)   |
| MAP         | 0.1061   | 0.3647 | 0.1132 | 0.3493 | 0.1599 | 0.3566 | 0.1399 | 0.3767 |
| %chg        |          |        | +6.2   | -4.4   | +33.6  | -2.3   | +24.2  | +3.2   |
| R-precision | 0.1581   | 0.3934 | 0.1903 | 0.3981 | 0.1985 | 0.3828 | 0.1694 | 0.4125 |
| %chg        |          |        | +16.9  | +1.2   | +20.4  | -2.8   | +6.7   | +4.6   |

The MAP of QBC\_CO is improved by 33.6% and that of QBC\_CS\_SEC is also improved by 24.2% in poorly performing queries over the baseline. This indicates that expanding terms by query-biased concepts has a positive effect in the case of poorly performing queries.

Then, we investigate the effect of structural information by comparing the results of various QBC\_CS approaches in Table 3. QBC\_CS\_SEC, QBC\_CS\_SS1, and QBC\_CS\_SS2 represent different kinds of partitioning. The cases of QBC\_CS\_SEC with non-clustering (NONCL) show the best result. Generally, the results of SS1 and SS2 are lower than those of SEC. In QBC\_CS approaches, a global analysis does not affect the improved performance for generating query-biased concepts.

**Table 3.** MAP of QBC\_CS runs in different levels of partitioning. CL denotes a clustering and NONCL denotes a non-clustering.

|       | QBC_CS_SEC |        | QBC_CS_SS1 |        | QBC_CS_SS2 |        |
|-------|------------|--------|------------|--------|------------|--------|
|       | CL         | NONCL  | CL         | NONCL  | CL         | NONCL  |
| TOP5  | 0.2035     | 0.2325 | 0.2028     | 0.2290 | 0.2304     | 0.2290 |
| TOP10 | 0.2043     | 0.2211 | 0.2036     | 0.1900 | 0.1930     | 0.1900 |
| TOP15 | 0.2079     | 0.2190 | 0.2072     | 0.2037 | 0.2090     | 0.2037 |
| TOP20 | 0.2079     | 0.2193 | 0.2072     | 0.2041 | 0.2090     | 0.2041 |

Finally, we examine the effect of the number of retrieved documents used for generating concepts. In Table 1, using 15 documents with QBC\_CO leads to the best result. In Table 3, using 5 documents with QBC\_CS\_SEC (NONCL) leads to the best result. This indicates that the number of retrieved documents does not directly affect the performance. This is because we did not use all the retrieved documents but only used those retrieved documents that were relevant. As long as some relevant documents are highly ranked, we are able to generate appropriate concepts for query expansion.

## 4 Conclusions

In this paper, we proposed an approach for constructing query-biased concepts from retrieved and relevant documents. The generated query-biased concepts were used to expand the queries in a relevance feedback (RF) process. The experimental results showed the improvement of retrieval performance with our various approaches. Particularly, we found an increase of performance when QBC was applied to the poorly performing queries. We also investigated the effect of structural information to construct the query-biased concepts and the number of retrieved documents used for generating the concepts. The use of structural information in a local analysis was effective to select the significant features of documents. But the use of clustering for a global analysis was not beneficial for query reformulation. In QBC\_CS approaches, those which generated the query-biased concepts with the content and structural information (without a global analysis) led to the best performance. The retrieval performance of QBC approaches does not seem to rely on the number of retrieved documents. This is because we only used the relevance information of retrieved documents. It is not necessary to have a large number of relevant documents for generating appropriate query-biased concepts. However, it is essential to have them highly ranked for generating appropriate query-biased concepts.

**Acknowledgments.** This work was financially supported by IT Scholarship Program supervised by IITA (Institute for Information Technology Advancement) & MKE (Ministry of Knowledge Economy), Republic of Korea.

## References

1. Chang, Y., Kim, M., Raghavan, V.V.: Construction of query concepts based on feature clustering of documents. *Information Retrieval*. 9(3), 231 -- 248 (2006)
2. Frakes, W. B., Baeza-Yates, R.: *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs, NJ (1992)
3. Malik, S., Lalmas, M., Fuhr, N: Overview of INEX 2005. In: *Advances in XML Information Retrieval and Evaluation: 4th Workshop of the Initiative for the Evaluation of XML Retrieval*, pp. 1--15. Springer-Verlag (2005)
4. Nakata, K., Voss, A., Juhnke, M., Kreifelts, T.: Collaborative concept extraction from documents. In: *2nd International Conference on Practical Aspects of Knowledge management*, pp. 29--30. Basel (1998)
5. Qiu, Y., Frei, H.P.: Concept based query expansion. In: *16th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 160--170. ACM press, Pittsburgh (1993)
6. Rocchio J.J.: Relevance Feedback in Information retrieval. *The SMART retrieval system – experiments in automatic document processing*, (G.Salton ed.) pp.313—323. (1971)
7. Rölleke, T., Lübeck, R., Kazai, G.: *The HySpirit Retrieval Platform*, In: *ACM SIGIR Demonstration*, New Orleans (2001)
8. Ruthven, I., Lalmas, M.: A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*. 18 (1). 95--145 (2003)