Query Expansion for Web Information Retrieval

Armin Hust, Stefan Klink, Markus Junker, Andreas Dengel
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH)
Postfach 2080, D-67608 Kaiserslautern
{armin.hust, stefan.klink, markus.junker, andreas.dengel}@dfki.de

ABSTRACT
Information retrieval (IR) systems utilize user feedback for generating optimal queries with respect to a particular information need. However the methods that have been developed in IR for generating these queries do not memorize information gathered from previous search processes, and hence can not use such information in new search processes. Thus a new search process can not profit from the results of the previous processes. Web Information Retrieval (WIR) systems should be able to maintain results from previous search processes, thus learning from previous queries and improving overall retrieval quality. In our approach we are using the similarity of a new query to previously learned queries. We then expand the new query by extracting terms from documents which have been judged as relevant to these previously learned queries. Thus our method uses global feedback information for query expansion in contrast to local feedback information which has been widely used in previous work in query expansion methods.

KEY WORDS
Web Information Retrieval, Collaborative Information Retrieval, Query Expansion, Text Mining

1 Introduction

Gathering information for fulfilling the information need of a user is an expensive operation in terms of time required and resources used. Queries may have to be reformulated manually by the user or automatically by the IR system several times until the user is satisfied. The same expensive operation has to be carried out, if another user has the same information need and thus initiates the same or a similar search process.
How users can improve the original query formulation by means of (automatic) relevance feedback is an ongoing research activity in IR [MS99]. In our approach we are using global relevance feedback which has been learned from previous queries instead of local relevance feedback which is produced during execution of an individual query.
The motivation for our query expansion method is straightforward, especially in an environment where document collections are static:

- If documents are relevant to a query which has been issued previously by a user, then the same documents are relevant to the same query at a later time, when that

query is re-issued by the same or by a different user. This is the trivial case, where similarities between the two different queries is the highest.

- In the non-trivial case a new query is similar to a previously issued query only to a certain degree. Then our assumption is that documents which are relevant to the previously issued query will be relevant to the new query only to a certain degree.

In this work we do not consider learning methods for user relevance feedback, instead we expect that relevance judgements are available for use. A WIR system should be able to maintain information about previous search processes as well as information about relevance judgements (directly specified or derived from users actions). Then in processes called Collaborative Information Retrieval (CIR) the system may improve overall retrieval quality for all users, benefitting from previous search processes issued by different users.

2 Traditional Document Retrieval

The task of document retrieval is to retrieve documents relevant to a given query from a fixed set of documents. Documents as well as queries are represented in a common way using a set of index terms.
One of the simplest but most popular models used in IR is the vector space model (VSM) [MS99], [BYRN99]. Documents and queries are represented as M dimensional vectors, where different term weighting schemes may be used.
The result of the execution of a query is a list of documents ranked according to their similarity to the given query. The similarity $sim(d_j, q)$ between a document d_j and a query q is measured by the cosine of the angle between these two M dimensional vectors.
Several methods, called query expansion methods, have been proposed to cope with the problem that short queries rank only a limited number of documents according to their similarity [QF93]. These methods fall into three categories: usage of feedback information from the user, usage of information derived locally from the set of initially retrieved documents, and usage of information derived globally from the document collection.
The method called *pseudo relevance feedback* works in three stages: First documents are ranked according to their similarity to the original query. Then highly ranked documents are assumed to be relevant and their terms are used for expanding the original query. Then documents are ranked again according to the similarity to the expanded query. In this work we employ a simple variant of pseudo relevance feedback [KJDM01].

3 Query Similarity and Relevant Documents

In this paper we employ a query expansion method based on query similarities and relevant documents (QSD). Our method uses feedback information and information globally available from previous queries. Feedback information in our experimental environment is available in the ground truth data provided by the test document collections. The ground truth provides relevance information, i.e. for each query there exists a list of relevant documents.
Query expansion works as follows:

- compute the similarities between the new query and each of the existing old queries
- select the old queries having a similarity to the new query which is greater than or

equal to a given threshold

- from these selected old queries get the sets of relevant documents from the ground truth data
- from each set of relevant documents compute a new document vector
- use these document vectors and a weighting scheme to enrich the new query

The formal description is given here. The similarity $sim(q_k, q)$ between a query q_k and a new query q is measured by the cosine of the angle between these two M dimensional vectors: $sim(q_k, q) = q_k^T \cdot q$, where T indicates the transpose of the vector q_k. Let S be the set

$$S = \{q_k | sim(q_k, q) \geq \sigma, \ \ 1 \leq k \leq L\} \tag{1}$$

of existing old queries q_k having a similarity greater than or equal to a threshold σ to the new query q and let T_k be the sets of all documents d_j relevant to the queries q_k in S. Then the sums $r_k = \sum_{d_j \in T_k} d_j$ of the document vectors in each T_k are used as expansion terms for the original query. The expanded query vector $q^{'}$ is then obtained by

$$q^{'} = q + \sum_{k=1}^{L} \lambda_k \frac{r_k}{\parallel r_k \parallel}, \tag{2}$$

where the λ_k are parameter for weighting the expansion terms.
Notes:

- if σ in (1) is chosen to high the set S may be empty. Then the sets T_k will be empty and the document vectors r_k will be $(0, ..., 0)^T$. In this case the new query will not be expanded.
- even if a query q_k is in the set S, the corresponding set T_k may be empty (in case where no relevance judgements are contained in the ground truth data for query q_k). Then the corresponding document vector r_k will be $(0, ..., 0)^T$.
- parameters σ in (1) and λ_k in (2) are the tuning parameters for method QSD.

4 Experimental Design

We use standard document test collections and standard queries and questions provided by [Sma] and [Tre]. On the one hand by utilizing these collections we take advantage of the ground truth data for performance evaluation. On the other hand we do not expect to have queries having highly correlated similarities as we would expect in a real world application. So it is a challenging task to show performance improvements for our method. In our experiments we used the following eight collections:

- the CACM, CISI and CRAN collections available at [Sma].
- the CR collection available from the TREC test collections disk 4 [Tre] using queries of different length. The CR-titlecontains the title"queries, the CR-desccontains the "description"queries, the CR-narrcontains the narrative"queries.
- the FR collection available from the TREC test collections disk 2.
- the AP90 available from the TREC test collections disk 3 together with selected questions from the TREC-9 Question Answering track, where several questions are only a re-wording of some other questions, but specifying the same information need [VH01].

Tabelle 1: Average precision obtained in different methods

	CACM	CISI	CRAN	CR-desc	CR-narr	CR-title	FR	AP90
VSM	0.130	0.120	0.384	0.175	0.173	0.135	0.085	0.743
PRF	0.199	0.129	0.435	**0.204**	**0.192**	0.169	0.113	0.755
QSD	0.237	0.142	0.428	0.172	0.173	0.152	0.109	*0.811*
QSDPRF	**0.257**	*0.145*	*0.451*	0.195	*0.191*	*0.177*	**0.163**	*0.814*
PRFQSD	*0.255*	**0.151**	**0.463**	*0.196*	**0.192**	**0.180**	*0.139*	**0.814**

Terms used for document and query representation were obtained by stemming and eliminating stopwords. Statistics about these collections before stemming and stopword elimination can be found in [BYRN99] and [KJDM01]. In our experiments we employ the standard *tf-idf* scheme for weighting document and query terms.

5 Experimental Results

In this section the results of the experiments are presented. Results were evaluated using the average precision over all queries. Significance tests were applied to the results.
Methods VSM (vector space model), PRF (pseudo relevance feedback) and QSD (query similarity and relevant documents) were applied. Parameters for PRF and QSD are chosen such that average precision is highest. Experiments have shown that λ_k values in equation (2) have to be set to the similarity values $sim(q_k, q)$ for best average precision, i.e. QSD considers the query similarities for best performance.

In the next step we combined two methods of query expansion in this ways: First, after having expanded the new query using the QSD method, we applied the PRF method against the expanded query. This method is reported as the QSDPRF method. Second, after having expanded the new query using the PRF method, we applied the QSD method against the expanded query. This method is reported as the PRFQSD method. Best parameter value settings have again been obtained by experiment and are chosen such that average precision is highest. Also for PRFQSD the λ_k values have to be set to the similarity values $sim(q_k, q)$ for best average precision, i.e. PRFQSD considers the query similarities for best performance.
Table 1 shows the average precision obtained by using the best parameter values for different methods. For each collection the best value of average precision is indicated by bold font, the second best value is indicated by italic font. Using the 'paired t-test' from [Hul93] we compared the results of different methods in terms of average precision. Table 2 shows the results, where '++' ('+') indicates that method X is superior to method Y at the 0.01 (0.05) significance level, '—' ('–') indicates that method Y is superior to method X at the 0.01 (0.05) significance level, and 'o' indicates that there is no indication for method X or Y performing superior than the other.

6 Conclusions

We have experimentally compared a new query expansion method with two conventional information retrieval methods. From the results gathered from eight static test collections

Tabelle 2: Paired t-test results for significance levels $\alpha = 0.05$ and $\alpha = 0.01$

methods		CACM	CISI	CRAN	CR-desc	CR-narr	CR-title	FR	AP90
X	Y								
PRF	VSM	++	+	++	++	+	+	+	+
QSD	VSM	++	+	++	o	o	o	+	++
QSD	PRF	o	o	o	—	–	o	o	++
QSDPRF	VSM	++	+	++	+	o	+	++	++
QSDPRF	PRF	o	o	o	o	–	o	+	++
QSDPRF	QSD	o	+	++	+	o	+	++	o
PRFQSD	VSM	++	+	++	o	+	+	++	++
PRFQSD	PRF	++	+	++	o	o	o	+	++
PRFQSD	QSD	o	o	++	o	o	+	+	o

we have only one clear indication that the QSD method is superior to the conventional PRF method. But in contrast we also have only one clear indication that the conventional PRF method is superior to the QSD method.
From our results we think that we can combine this new method with the conventional PRF method. No performance degradation has been observed for this combination of the two methods. The results that have been obtained by combining the new QSD method with the conventional PRF method are promising.
Due to the construction method for the queries in the AP90 test collection (see section 4) where QSD significantly performs better than the other methods we think that we could utilize this new method in cases where old queries and their corresponding relevance information has been learned previously and where new queries have high similarities to old existing queries.

Literaturverzeichnis

[BYRN99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, 1999.

[Hul93] David Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of SIGIR-93*, pages 329–338, 1993.

[KJDM01] Koichi Kise, Markus Junker, Andreas Dengel, and Keinosuke Matsumoto. Experimental Evaluation of Passage-Based Document Retrieval. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, 2001.

[MS99] C.D. Manning and H. Schütze. *Foundations of Natural Language Processing*. MIT Press, 1999.

[QF93] Yonggang Qiu and Hans-Peter Frei. Concept-based query expansion. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, US, 1993.

[Sma] `ftp://ftp.cs.cornell.edu/pub/smart`.

[Tre] `http://trec.nist.gov`.

[VH01] Ellen M. Voorhees and Donna Harman. Overview of the Ninth Text Retrieval Conference (TREC-9). In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*, 2001.