A Mathematical Model for Improving Retrieval Performance in Collaborative Information Retrieval

Armin Hust (armin.hust@dfki.de),
Markus Junker (markus.junker@dfki.de) and
Andreas Dengel (andreas.dengel@dfki.de)
German Research Center for Artificial Intelligence
Erwin-Schrödinger-Straße 57
67663 Kaiserlautern
Germany

2000/04/29

Abstract. The accuracy of ad-hoc information retrieval (IR) systems has plateaued in the last few years. We are working on so-called collaborative information retrieval (CIR) systems which unobtrusively learn from their users' search processes. As a first step towards techniques, we focus on a restricted setting in CIR in which only old queries and correct answer documents to these queries are available for improving on a new query. For this restricted setting we propose two initial approaches, called DTW and QTW, for reweighting the document- resp. the query-terms. We then combine these approaches with the pseudo relevance feedback method. The approaches are evaluated experimentally on standard IR test collections. It turns out that in particular the hybrid approaches with pseudo relevance feedback give promising results. A bigger advantage of the proposed approaches is expected in real word test scenarios in which the overlap of user interests is larger than in our experimental set up.

Keywords: collaborative information retrieval, information retrieval, query expansion, text mining

Table of Contents

1. Introduction

Information Retrieval (IR) Systems have been studied in Computer Science for decades. The traditional ad-hoc task in Information Retrieval is to find all documents relevant for an ad-hoc given query. Much work has been done on improving this task, in particular in the Text Retrieval Conference series (TREC) (Trec, 2003). In 2000, it was decided at TREC-8 that this task should no longer be pursued within TREC, in particular because the accuracy has plateaued in the last few years (Voorhees and Harman, 1999). At German Research Center for Artificial Intelligence, we are working on approaches for Collaborative Information Retrieval (CIR) which learn to improve retrieval effectiveness from the interaction of different users with the retrieval engine. Such systems may have the potential to overcome the current plateau in ad-hoc retrieval.

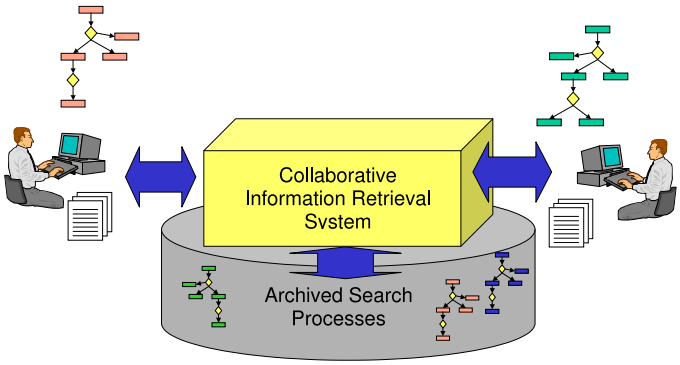

Figure 1. Scenario of Collaborative Information Retrieval.

Figure 1 illustrates the general scenario of CIR. An information retrieval system is typically used by many users. A typical search in a retrieval system consists of several query formulations. Often, the answer documents to the first query do not directly satisfy the user. Instead, the user has to reformulate his query taking the answer documents found into consideration. Such refinement may consist of specializations as well as generalizations of previous queries. In general, satisfying an information need requires to go through a search process with many decisions on query reformulations. The idea of CIR is to store these search processes as well as the ratings of documents returned by the system (if available) in an archive. Subsequent users with similar interests and queries should then benefit from knowledge automatically acquired by the CIR system based on the stored search processes. This should result in shorter search processes and better retrieval quality for subsequent users.

In this paper we focus on the investigation of a mathematical model for performance improvement in the CIR scenario. Given a number of old queries posed by different users and their corresponding answer documents, we try to improve on an arbitrary new query.

We call our approach, which learns to improve retrieval effectiveness from the interaction of different users with the retrieval engine, Collaborative Information Retrieval (CIR). CIR is the task where an IR system makes full use of all the information available in the system, especially

- the information from previous search processes, i.e. individual queries and complete search processes
- the relevance information which is gathered from previous search processes, independent from the method how this relevance information is obtained (explicitly by user relevance feedback or implicitly by unobtrusively detected relevance information)

The collaborative aspect here differs from other collaborative processes. We do not assume that different users from a working team or a specific community collaborate loosely or tightly through some information exchange or workflow processes. Instead we assume that users can benefit from search processes carried out at former times by other users (although those users may not know about the other users and their search processes), as long as the relevance information gathered from these previous users has some significant meaning.

This paper is organized as follows: Section 2 contains information about related work, in section 3 we introduce the vector space model and the pseudo relevance feedback model, since we are combining it with our new approaches. In section 4 we introduce, motivate and formalize the methods we are tackling. Section 5 contains an example for illustrating the methods. Sections 6 to 8 describe the experimental setup using standard IR test collections, present the results and compare them with the vector space model and the pseudo relevance feedback model. In section 9 we briefly review our new methods.

2. Related Work

Research in the field of query expansion (QE) procedures has been done for several years now. A lot of different procedures have been proposed for manual, automatic or interactive QE, leading to performance improvements. Some of the older procedures are available in (Harman, 1992b; Harman, 1992a). A comprehensive overview for newer procedures is available in (Efthimiadis, 1996). Another newer technique, called local context analysis, was introduced by (Xu and Croft,

1996; Xu and Croft, 2000), which analyzes the top ranked documents initially retrieved for a query.

Newest procedures in the field of query expansion are dealing with query bases, a set of persistent past optimal queries, for investigating similarity measures between queries. The query base can be used either to answer user queries or to formulate optimal queries (Raghavan and Sever, 1995). (Wen et al., 2001; Wen et al., 2002) are using query clustering techniques for discovering frequently asked questions or most popular topics on a search engine. This query clustering method makes use of user logs which allows to identify the documents the users have selected for a query. The similarity between two queries may be deduced from the common documents the users selected for them. (Cui et al., 2002) take into account the specific characteristics of web searching, where a large amount of user interaction information is recorded in the web query logs, which may be used for query expansion.

Gathering relevance feedback is another field of research in this area. Automatic acquisition of relevance information is necessary for improving IR performance, since users are not willing or do not intend to give feedback about the relevance of retrieved documents, although it has been shown by (Salton et al., 1985) that IR effectiveness does not improve any more after a few iterations of relevance feedback, and (Dominich, 2001) shows the mathematical structure of relevance effectiveness 'converging' to a stable limit. (White et al., 2002) compare two systems, where one is using explicit relevance feedback (explicitly marking relevant documents) and one is using implicit relevance feedback. They focus on the degree to which implicit evidence of document relevance can be substituted for explicit evidence. (Joachims, 2002) acquires relevance information by merely using the clickthrough data while the documents presented to the user have been ranked by two different IR systems.

Work in the field of term weighting procedures has been done ever since IR research. The dynamics of probabilistic term weights in different IR models have been discussed in (Crestani and van Rijsbergen, 1995a; Crestani and van Rijsbergen, 1995b; Crestani and van Rijsbergen, 1998), going back to the work of (van Rijsbergen, 1986; van Rijsbergen, 1989). The different models analyze the transfer of probabilities in the term space.

3. Basics and Terminology

In this section we briefly recall the vector space model for Information Retrieval on which all of our approaches rely (section 3.1) and describe

the pseudo relevance feedback model we are utilizing (section 3.2). We then formalize the CIR scenario we are focussing on in this paper (section 3.3).

3.1. The Vector Space Model

The basic retrieval model we rely on is the vector space model [VSM] (Baeza-Yates and Ribeiro-Neto, 1999), (Manning and Schütze, 1999). Documents as well as queries are represented by vectors. The set of N documents is denoted by $\mathbf{D} = \{d_j | 1 \leq j \leq N\}$, the set of L queries is denoted by $\mathbf{Q} = \{q_k | 1 \leq k \leq L\}$. Each individual document d_j is represented by its vector $d_j = (w_{1j}, w_{2j}, ..., w_{Mj})^T$, each individual query is represented by its vector $q_k = (w_{1k}, w_{2k}, ..., w_{Mk})^T$, where M is the number of terms in the collection and T denotes the transpose of the vector.

The similarity sim between a given query q_k and a document d_j is measured by the cosine of the angle between these two M dimensional vectors:

$$sim(d_j, q_k) = \frac{d_j^T \cdot q_k}{\parallel d_j \parallel \cdot \parallel q_k \parallel} \tag{1}$$

where $\parallel \cdot \parallel$ is the Euclidean norm of a vector. In the case that the vectors are already normalized (and hence have a unit length) the similarity is simply the dot product $sim(d_j, q_k) = d_j^T \cdot q_k = \sum_{i=1}^{M} w_{ij} \cdot w_{ik}$ between the two vectors d_j and q_k. For retrieving all documents to a given query, all documents of the underlying collection are ranked according to their similarity to the query and the top-ranked documents are given to the user.

3.2. Pseudo Relevance Feedback

The method called *pseudo relevance feedback* (also called *pseudo feedback*, [PRF]) is a method for overcoming the limitations in the number of documents ranked according to their similarity to the query when using short queries.

In this work we employ a simple variant of pseudo relevance feedback (Kise et al., 2001). Let E be the set of document vectors given by

$$E = \left\{ d_j \middle| \frac{sim(d_j, q_k)}{\max_{1 \leq i \leq N}\{sim(d_i, q_k)\}} \geq \theta \right\} \tag{2}$$

where q_k is the original query and θ is a threshold of the similarity. Then the sum D_k of the document vectors in E

$$D_k = \sum_{d_j \in E} d_j \tag{3}$$

is used as expansion terms for the original query. The expanded query vector q'_k is obtained by

$$q'_k = q_k + \alpha \frac{D_k}{\parallel D_k \parallel} \tag{4}$$

where α is a parameter for weighting the expansion terms. Then the documents are ranked again according to their similarity $sim(d_j, q'_k)$.

PRF is one of the models we are using in this work for comparison to our new query expansion method.

3.3. Restricted CIR Scenario

As we have stated in the introduction, in this paper we focus on a restricted CIR scenario. In this scenario we have a set of old (former) queries $\mathbf{Q} = \{q_1, q_2, \ldots, q_L\}$ available.

For each $q_k \in \mathbf{Q}$ the set of corresponding relevant documents is known (the ground truth data) and denoted by r_k, where each r_k is an N dimensional column vector

$$r_k = (r_{1k}, r_{2k}, ..., r_{Nk})^T, \quad 1 \leq k \leq L \tag{5}$$

where

$$r_{ik} = \begin{cases} 1 & \text{if document } i \text{ is relevant to query } k \\ 0 & \text{if document } i \text{ is not relevant to query } k \end{cases} \tag{6}$$

The goal now is to find all relevant documents for a new query $q^\star$ based on the old queries $\mathbf{Q}$ and their relevant documents r_k (in general $q^\star \notin \mathbf{Q}$). This is done by expanding the query $q^\star$ to a new query $q^\star_{exp}$ which is then used instead of $q^\star$. More formally, let $\mathbf{Q}$ be queries and $\mathbf{D}$ documents. We are searching for an expansion function

$$\begin{array}{rl} f_{exp} : \mathbf{Q} \times 2^{(\mathbf{Q} \times (2^{\mathbf{D}}))} & \rightarrow \mathbf{Q} \\ (q^\star, \{(q_1, r_1), (q_2, r_2), \ldots, (q_L, r_L)\}) & \mapsto q^\star_{exp} \end{array} \tag{7}$$

which maximizes the effectiveness of $q^\star_{exp}$.

4. Term Reweighting Methods

In this section we are motivating and defining two new global term-weighting functions, where one of them is based on reweighting document terms, the other is based on reweighting query terms. Both of these methods use the relevance information that is available from the ground truth data.

4.1. Term Reweighting Motivation

The basic idea for reweighting the terms is as follows:

- for reweighting document terms, we are giving those ambiguous terms of the documents more weight, that match the semantics of the same terms in the queries, i.e. a term like 'bank' having the meaning of 'financial institution' will be weighted higher than the term 'bank' having a meaning of 'dike' or 'wall', if the queries use the term 'bank' in conjunction with other terms related to financial topics
- for reweighting query terms, we are giving those ambiguous terms of the queries more weight, that match the semantics of the same terms in the documents, i.e. a term like 'bank' having the meaning of 'financial institution' will be weighted higher than the term 'bank' having a meaning of 'dike' or 'wall', if the documents use the term 'bank' in conjunction with other terms related to financial topics

Given the set **D** of N documents we write the document vectors as a matrix and denote it by $D = (d_j)_{1 \leq j \leq N}$, $D \in Mat(M \times N, \mathbb{R})$. For the set **Q** of L queries we write the query vectors as a matrix and denote it by $Q = (q_k)_{1 \leq j \leq L}$, $Q \in Mat(M \times L, \mathbb{R})$. We also have the relevance judgements matrix $R = (r_k)_{1 \leq j \leq L}$, $R \in Mat(N \times L, \{0, 1\})$.

Assuming that the document and query vectors are normalized, the similarity between all documents d_j and queries q_k is computed as

$$\begin{aligned} SIM &= sim(D, Q) \\ &= \begin{pmatrix} sim(d_1, q_1) & sim(d_1, q_2) & . & . & sim(d_1, q_L) \\ sim(d_2, q_1) & sim(d_2, q_2) & . & . & sim(d_2, q_L) \\ . & . & . & . & . \\ . & . & . & . & . \\ sim(d_N, q_1) & sim(d_N, q_2) & . & . & sim(d_N, q_L) \end{pmatrix} \end{aligned} \tag{8}$$

or in matrix notation

$$SIM = sim(D, Q) = D^T \cdot Q \tag{9}$$

where $SIM \in Mat(N \times \text{Ł}, \mathbb{R})$ is a matrix of N rows and L columns containing the similarity values.

Sorting each of the column vectors in SIM descending on the similarity values gives the ranking of the documents according to their similarity to the given queries. Then the column vectors in SIM and the relevance information r_k are used for calculating the recall and precision values during evaluation of retrieval performance.

The SIM matrix and the matrix R containing relevance judgements are of same size. Both matrices have N rows and L columns. In the best case the SIM matrix computed according to equation (9) would be identical to the given R matrix. In this case the precision for

each query would be 100% at every recall level, because every relevant document has a similarity of 1 to the query and every non-relevant document has a similarity of 0 to the query.

These considerations lead to the term reweighting methods described below.

4.2. Term Reweighting for Document Terms

Since the similarity matrix SIM computed in (9) will almost never be identical to the relevance judgements matrix R we try to find a transformation matrix to achieve this goal. Assuming that a transformation matrix $W_D \in Mat(N \times N, \mathbb{R})$ exists that satisfies

$$W_D \cdot SIM = R \tag{10}$$

we can rewrite equation (10) as follows

$$\begin{aligned} R &= W_D \cdot SIM \\ &= W_D \cdot (D^T \cdot Q) \\ &= (W_D \cdot D^T) \cdot Q \\ &= D_W^T \cdot Q \end{aligned} \tag{11}$$

where the transformation matrix W_D is used for reweighting the document terms, giving a new document term matrix D_W, which is independent from new queries.

Equation (10) specifies systems of several linear equations, where the unknowns are the N column vectors of matrix W_D. Rewriting of equation (10)

$$\begin{aligned} (W_D \cdot SIM)^T &= \\ SIM^T \cdot W_D^T &= \\ SIM^T \cdot (w_{D1}, w_{D2}, \ldots, w_{DN})^T &= R^T \end{aligned} \tag{12}$$

gives the standard form of N linear equation systems, which have to be solved for each column vector of the transformation matrix W_D^T.

In our case SIM is normally singular ($N \neq L$), depending on the number of documents and queries contained in the collection. In this situation we try to find column vectors $(\hat{w}_{D1}, \hat{w}_{D2}, ...\hat{w}_{DN})$ such that they provide a closest fit (in some sense) to the equation for the under- or overdetermined system.

Our approach is to minimize the Euclidean norm of all of the column vectors

$$SIM^T \cdot (w_{D1}, w_{D2}, \ldots, w_{DN})^T - R^T \tag{13}$$

i.e. we solve

$$(\hat{w}_{D1}, \hat{w}_{D2}, \ldots, \hat{w}_{DN})^T = argmin_{(w_{D1}, w_{D2}, \ldots, w_{DN})^T} \| SIM^T \cdot W_D^T - R^T \| \tag{14}$$

where $\hat{W}_D^T = (\hat{w}_{D1}, \hat{w}_{D2}, \ldots, \hat{w}_{DN})^T$ is called the least squares solution for the system $SIM^T \cdot W_D^T = R^T$.

After computation of the transformation matrix $\hat{W}_D$ we can now reweight the document terms according to equation (11), i.e. we now compute

$$\hat{D}_W^T = \hat{W}_D \cdot D^T \tag{15}$$

or equivalently

$$\hat{D}_W = D \cdot \hat{W}_D^T \tag{16}$$

and have a reweighted document term matrix $\hat{D}_W$.

4.3. Term Reweighting for Query Terms

Since the similarity matrix SIM computed in (9) will almost never be identical to the relevance judgements matrix R we try to find a transformation matrix to achieve this goal. Assuming that a transformation matrix $W_Q \in Mat(L \times L, \mathbb{R})$ exists that satisfies

$$SIM \cdot W_Q = R \tag{17}$$

we can rewrite equation (17) as follows

$$\begin{aligned} R &= SIM \cdot W_Q \\ &= (D^T \cdot Q) \cdot W_Q \\ &= D^T \cdot (Q \cdot W_Q) \\ &= D^T \cdot Q_W \end{aligned} \tag{18}$$

where the transformation matrix W_Q is used for reweighting the query terms, giving a new query term matrix Q_W, which is independent from new queries to be issued in the system.

Equation (17) specifies systems of several linear equations, where the unknowns are the L column vectors of matrix W_Q. Equation (17) is in the standard form of L linear equation systems, which have to be solved for each column vector of the transformation matrix W_Q.

In our case SIM is normally singular ($N \neq L$), depending on the number of documents and queries contained in the collection. We try to find column vectors $(\hat{w}_{Q1}, \hat{w}_{Q2}, \ldots \hat{w}_{QL})$ such that they provide a closest fit (in some sense) to the equation for the under- or overdetermined system.

Our approach is to minimize the Euclidean norm of all of the column vectors

$$SIM \cdot (w_{Q1}, w_{Q2}, \ldots, w_{QL}) - R \tag{19}$$

i.e. we solve

$$(\hat{w}_{Q1}, \hat{w}_{Q2}, \ldots, \hat{w}_{QL}) = argmin_{(w_{Q1}, w_{Q2}, \ldots, w_{QL})} \|SIM \cdot W_Q - R\| \quad (20)$$

where $\hat{W}_Q = (\hat{w}_{Q1}, \hat{w}_{Q2}, \ldots, \hat{w}_{QL})$ is called the least squares solution for the system $SIM \cdot W_Q = R$.

After computation of the transformation matrix $\hat{W}_Q$ we can now reweight the query terms according to equation (18), i.e. we now compute

$$\hat{Q}_W = Q \cdot \hat{W}_Q \quad (21)$$

and have a reweighted query term matrix $\hat{Q}_W$.

4.4. Transformation matrices

In equations (10) and (17)

$$W_D \cdot SIM = R$$

resp.

$$SIM \cdot W_Q = R$$

we assume that transformation matrices W_D respectively W_Q exist that satisfy the equations.

The existence of these matrices follows from the following definition, theorem and lemma. Refer to (Ben-Israel and Greville, 1974).

DEFINITION 1. *The matrix $A^\dagger$ is called a pseudo inverse (or Moore-Penrose inverse) matrix of A if*

1. $A = A \cdot A^\dagger \cdot A$
2. $A^\dagger = A^\dagger \cdot A \cdot A^\dagger$
3. $A^\dagger \cdot A = (A^\dagger \cdot A)^T$
4. $A \cdot A^\dagger = (A \cdot A^\dagger)^T$

THEOREM 2. *For each matrix there exists exactly one pseudo inverse matrix.* ♦

Using this definition and theorem we can show that the matrices W_D and W_Q exist and are uniquely defined.

THEOREM 3. *In equations (10) and (17)*

$$W_D \cdot SIM = R$$

resp.

$$SIM \cdot W_Q = R$$

the matrices W_D and W_Q exist and are defined by:

1. $W_D = R \cdot SIM^{\dagger}$
2. $W_Q = SIM^{\dagger} \cdot R$

Proof. From theorem (2) follows that the pseudo inverse $SIM^{\dagger}$ *of matrix* SIM *exists and is uniquely defined. Then*

$$\begin{aligned}
& W_D \cdot SIM && = R \\
\Leftrightarrow\ & W_D \cdot SIM \cdot (SIM^{\dagger} \cdot SIM) && = R \cdot (SIM^{\dagger} \cdot SIM) \\
\Leftrightarrow\ & W_D \cdot (SIM \cdot SIM^{\dagger} \cdot SIM) && = R \cdot SIM^{\dagger} \cdot (SIM \cdot SIM^{\dagger} \cdot SIM) \\
\Leftrightarrow\ & W_D && = R \cdot SIM^{\dagger}
\end{aligned}$$

and

$$\begin{aligned}
& SIM \cdot W_Q && = R \\
\Leftrightarrow\ & (SIM \cdot SIM^{\dagger}) \cdot SIM \cdot W_Q && = (SIM \cdot SIM^{\dagger}) \cdot R \\
\Leftrightarrow\ & (SIM \cdot SIM^{\dagger} \cdot SIM) \cdot W_Q && = (SIM \cdot SIM^{\dagger} \cdot SIM) \cdot SIM^{\dagger} \cdot R \\
\Leftrightarrow\ & W_Q && = SIM^{\dagger} \cdot R
\end{aligned}$$

⧫

5. Illustrating the Term Reweighting Methods

In this section we show an example which makes intuitively clear, how the reweighting methods work on a given set of documents and queries.

Let us start the example showing the terms used in the documents and queries. The terms are numbered from t_1 to t_{17}, which are contained in the term-vector denoted by T:

$$\begin{aligned}
T \ = \ & (\prime annuity\prime, \prime bank\prime, \prime blood\prime, \prime bogus\prime, \prime bottle\prime, \prime capital\prime, \prime cash\prime, \prime credit\prime, \\
& \prime debt\prime, \prime deposit\prime, \prime earth\prime, \prime food\prime, \prime interest\prime, \prime loan\prime, \prime note\prime, \prime sand\prime, \prime stock\prime)^T
\end{aligned}$$

In this example the terms t_1 and t_3 to t_{17} have been selected such that they have a relation and meaning together with the term $t_2 =\prime bank\prime$.

We have created six documents using these terms as follows, where d_1 and d_2 have been created such that they use the term $t_2 =\prime bank\prime$ with a meaning of a financial institution, which is indicated by the co-occurrence of the terms related to financial affairs in these documents:

$$\begin{aligned}
d_1 \ &= \ (\prime bank\prime, \prime credit\prime, \prime debt\prime, \prime interest\prime, \prime loan\prime, \prime note\prime)^T \\
d_2 \ &= \ (\prime annuity\prime, \prime bank\prime, \prime capital\prime, \prime cash\prime, \prime deposit\prime, \prime stock\prime)^T \\
d_3 \ &= \ (\prime bank\prime, \prime blood\prime, \prime bogus\prime, \prime earth\prime)^T \\
d_4 \ &= \ (\prime bank\prime, \prime bottle\prime, \prime food\prime, \prime sand\prime)^T \\
d_5 \ &= \ (\prime blood\prime, \prime bogus\prime, \prime earth\prime)^T \\
d_6 \ &= \ (\prime bottle\prime, \prime food\prime, \prime sand\prime)^T
\end{aligned}$$

We have created five queries using these terms as follows, where all queries have been created such that they use the term $t_2 =' bank'$ with a meaning of a financial institution, which is indicated here by the additional term in each query referring to financial affairs:

$$\begin{aligned} q_1 &= ('bank','interest')^T \\ q_2 &= ('bank','credit')^T \\ q_3 &= ('bank','note')^T \\ q_4 &= ('bank','deposit')^T \\ q_5 &= ('bank','capital')^T \end{aligned}$$

Now we can write the documents and the queries in the matrices D_{raw} and Q_{raw} as follows, where the entries in the matrices just show the raw term frequencies:

$$D_{raw} = \begin{pmatrix} 0&1&0&0&0&0 \\ 1&1&1&1&0&0 \\ 0&0&1&0&1&0 \\ 0&0&1&0&1&0 \\ 0&0&0&1&0&1 \\ 0&1&0&0&0&0 \\ 0&1&0&0&0&0 \\ 1&0&0&0&0&0 \\ 1&0&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&0&1&0&1&0 \\ 0&0&0&1&0&1 \\ 1&0&0&0&0&0 \\ 1&0&0&0&0&0 \\ 1&0&0&0&0&0 \\ 0&0&0&1&0&1 \\ 0&1&0&0&0&0 \end{pmatrix}, \quad Q_{raw} = \begin{pmatrix} 0&0&0&0&0 \\ 1&1&1&1&1 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&1 \\ 0&0&0&0&0 \\ 0&1&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 1&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{pmatrix}$$

In addition we have created our relevance information such that documents d_1 and d_2 are relevant to each of the five queries, and documents d_3 to d_6 are not relevant to any of the five queries. This gives our relevance matrix R as follows:

$$R = \begin{pmatrix} 1&1&1&1&1 \\ 1&1&1&1&1 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \end{pmatrix}$$

Weighting and normalization of the document and query vectors according to our $tfidf$ scheme (see section 6.2 for detailed description of the weighting scheme) and computing the similarity matrix SIM in the standard vector space model according to equation (9) gives the following matrix:

$$SIM = \begin{pmatrix} 0.3858 & 0.3858 & 0.3858 & 0.0712 & 0.0712 \\ 0.0712 & 0.0712 & 0.0712 & 0.3858 & 0.3858 \\ 0.1474 & 0.1474 & 0.1474 & 0.1474 & 0.1474 \\ 0.1474 & 0.1474 & 0.1474 & 0.1474 & 0.1474 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

This similarity matrix leads to an interpolated average precision of 0.75 and gives the following recall/precision graph:

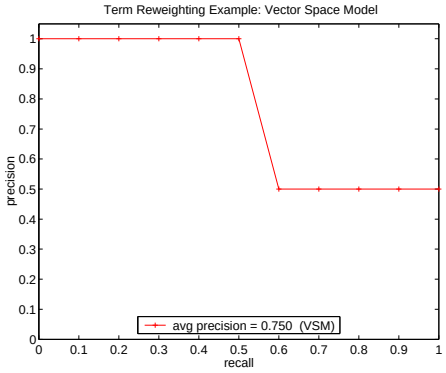

Figure 2. Term reweighting methods: Vector Space Model

For method DTW we compute the document term reweighting matrix $\hat{W}_D$ according to equation (14), and according to equation (16) the reweighted document term matrix $\hat{D}_W$ which leads to similarity matrix SIM according to equation (9)

$$SIM = \begin{pmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

giving an interpolated average precision of 1.0.

For method QTW we compute the query term reweighting matrix $\hat{W}_Q$ according to equation (20), and according to equation (21) the reweighted query term matrix $\hat{Q}_W$ which leads to similarity matrix SIM according to equation (9)

$$SIM = \begin{pmatrix} 2.8251 & 2.8251 & 2.8251 & 2.8251 & 2.8251 \\ 2.8251 & 2.8251 & 2.8251 & 2.8251 & 2.8251 \\ 1.8219 & 1.8219 & 1.8219 & 1.8219 & 1.8219 \\ 1.8219 & 1.8219 & 1.8219 & 1.8219 & 1.8219 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

giving an interpolated average precision of 1.0.

6. Experimental Design

6.1. The Test Collections

In this section we describe the test collections used in the experimental comparison. We use standard document test collections and standard queries and questions provided by (Smart, 1988) and (Trec, 2003). 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 nine collections:

- the CACM, CISI and CRAN collections are available at (Smart, 1988).
- the CR (congressional record) collection. We created three test cases for the CR collection, using the TREC queries of different length in order to investigate the influence of query length. The "CR-title" contains the "title" queries (the shortest query representation), the "CR-desc" contains the "description" queries (the medium length query representation), the "CR-narr" contains the "narrative" queries (the longest query representation).
- the FR (federal register) collection.
- the QA (question answering) collection. From the TREC-9 Question Answering track (QA) we selected the question set 201-893. Because of the construction method for this set (Voorhees and Harman, 2000) we expected to get higher similarities between different questions. We selected only those documents being relevant to at least one of the questions. Thus we reduced our test data to 6025 documents and 693 questions.
- the QA-AP90 collection. From the QA collection described above we selected only those questions having a relevant answer document in the AP90 document collection, and only those documents being relevant to at least one of the questions. Thus we reduced our test data to 723 documents and 353 questions.

6.2. Preparation of the Test Collections

Terms used for document and query representation were obtained by stemming and eliminating stopwords. Table I lists statistics about the collections after stemming and stopword elimination has been carried out, statistics about these collections before stemming and stopword elimination can be found in (Baeza-Yates and Ribeiro-Neto, 1999) and (Kise et al., 2001).

The best known term weighting schemes use weights according to the so-called *tfidf* schemes. In our experiments we employ a standard scheme as follows: For document vectors the weights w_{ij} are calculated as

$$w_{ij} = tf_{ij} \cdot idf_i \tag{22}$$

where tf_{ij} is a weight computed from the raw frequency f_{ij} of a term t_i (the number of occurrences of term t_i in document d_j)

$$tf_{ij} = \sqrt{f_{ij}} \tag{23}$$

and idf_i is the inverse document frequency of term t_i given by

$$idf_i = \log \frac{N}{n_i} \tag{24}$$

where n_i is the number of documents containing term t_i. For query vectors the weights w_{ik} are calculated as

$$w_{ik} = \sqrt{f_{ik}} \tag{25}$$

where f_{ik} is the raw frequency of a term t_i in a query q_k (the number of occurrences of term t_i in query q_k).

Table I. Statistics about test collections

	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN
size(MB)	1.2	1.4	93	93	93	1.4
number of documents	3204	1460	27922	27922	27922	1400
number of terms	3029	5755	45717	45717	45717	2882
mean document length	18.4 (short)	38.2 (med)	188.2 (long)	188.2 (long)	188.2 (long)	49.8 (med)
number of queries	52	112	34	34	34	225
mean query length	9.3 (med)	23.3 (long)	7.2 (med)	22.8 (long)	2.9 (short)	8.5 (med)
mean relevant documents per query	15.3 (med)	27.8 (high)	24.8 (high)	24.8 (high)	24.8 (high)	8.2 (med)

Table I.

	FR	QA	QA-AP90
size(MB)	69	28.2	3.7
number of documents	19860	6025	723
number of terms	50866	48381	17502
mean document length	189.7 (long)	230.7 (long)	201.8 (long)
number of queries	693	353	112
mean query length	3.1 (short)	3.2 (short)	9.2 (med)
mean relevant documents per query	16.4 (med)	2.8 (low)	8.4 (med)

7. Document Term Reweighting Experiments

In this section we describe a query expansion method based on query similarity and document term reweighting and relevant documents. The method is denoted as DTW from now on.

We will first give a high level description of the method, then the detailed mathematical description is given.

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 and non-relevant documents from the ground truth data
- from this set of relevant documents compute a term reweighting matrix such that relevant documents are ranked higher than non-relevant documents
- use this term reweighting matrix to compute the ranking of the documents for the new query

The formal description is given here. Let S be the set

$$S = \{q_k | sim(q_k, q_l) \geq \sigma, \quad 1 \leq k \leq L\} \tag{26}$$

of existing old queries q_k having a similarity greater than or equal to a threshold σ to the new query q_l and let R_k be the set

$$R_k = \{r_k | q_k \in S, \quad 1 \leq k \leq L\} \tag{27}$$

of all relevance judgements for the queries q_k in S. Write R_k as a matrix of N rows and $|S|$ columns

$$R = (r_k)_{1 \leq k \leq |S|} \tag{28}$$

and compute the document term reweighting matrix $\hat{W}_D$ according to equations (11) to (14) such that

$$\hat{W}_D \cdot SIM = R \tag{29}$$

Then do a document term reweighting

$$\hat{D}_W = D \cdot \hat{W}_D^T \tag{30}$$

according to equations (15) and (16) and compute the new document ranking for query q_l according to equation (8)

$$sim_l = sim(\hat{D}_W, q_l) = \hat{D}_W^T \cdot q_l \tag{31}$$

Notes:

- if σ in (26) is chosen to high the set S may be empty. Then the set R_k will be empty and we will not compute any document term reweighting matrix $\hat{W}_D$. In this case we will have our original similarity vector sim_l computed from $sim_l = sim(D, q_l) = D^T \cdot q_l$.
- parameter σ in (26) is a tuning parameter for method DTW.

7.1. Experimental Results

In this section the results of the experiments are presented. Results were evaluated using the average precision over all queries. Recall/precision graphs were generated according to (Baeza-Yates and Ribeiro-Neto, 1999). Then significance tests were applied to the results.

7.1.1. *Results*

The methods VSM (vector space model), PRF (pseudo relevance feedback) and DTW (document term reweighting) were applied. Best parameter value settings for parameters α and θ for method PRF had been obtained previously by experiment and those which give the highest average precision were selected and used (see table II and (Kise et al., 2001)).

Table II. Best parameter values for method PRF

	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
α	1.7	0.7	0.5	0.4	0.6	1.3	0.6	0.3	0.2
θ	0.35	0.7	0.85	0.95	0.75	0.9	0.55	0.6	0.75

Method DTW has been evaluated using different settings for parameters σ. From the set of queries contained in each collection we selected each query one after the other and treated it as a new query $q_l, 1 \leq l \leq L$. Then for each fixed query q_l we computed the similarity $\sigma_k := sim(q_k, q_l)$ for all queries $q_k, 1 \leq k \leq L, k \neq l$ according to equation (1). Then σ in equation (26) has been varied from 0.0 up to 1.0 in steps of 0.01. Finally we got our document term reweighting matrix according to equation (29) and issued the query. Best parameter values for σ are reported in table III.

In the next step we combined two methods of query expansion in this way: After having expanded the new query using the PRF method, we applied the DTW method against the expanded query. This method is reported as the PRFDTW method. Best parameter values settings have again been obtained by experiment and are chosen

Table III. Best parameter values for method DTW

	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
σ	0.26	0.12	0.41	0.34	0.44	0.83	0.36	0.75	0.82

Table IV. Best parameter values for method PRFDTW

		CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
PRFDTW	α, θ	same as in table II								
	σ	0.72	0.12	0.42	0.35	0.48	0.95	0.40	0.81	0.84

such that average precision is highest. These parameter value settings are reported in table IV.

Table V 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.

Table V. Average precision obtained in different methods

	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
VSM	0.130	0.120	0.175	0.173	0.135	0.384	0.085	0.645	0.745
PRF	*0.199*	*0.129*	**0.204**	**0.192**	**0.169**	**0.435**	*0.113*	0.685	0.757
DTW	0.142	0.122	0.150	0.173	0.132	0.386	0.098	*0.727*	*0.785*
PRFDTW	**0.208**	**0.133**	*0.200*	*0.191*	*0.154*	*0.431*	**0.123**	**0.752**	**0.791**

Table VI compares the average precision obtained by the different methods. The ratio of difference is calculated as follows: let X be the average precision obtained by one of the methods and let Y be the average precision obtained by another method. Then the ratio is calculated by $(X-Y)/Y$. A positive value for the ratio indicates an improvement, a negative value indicates a degradation in average precision. Note, that average precision values in table V are rounded to three digits, while average precision values are not rounded before computation of ratios of differences.

Figures 3 and 4 show some of the recall/precision graphs for the test collections. Each figure contains the graphs for methods VSM, PRF, DTW and PRFDTW.

Table VI. Average precision comparison

methods X	methods Y	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR
PRF	VSM	+52.7%	+7.3%	+16.6%	+10.9%	+25.5%	+13.4%	+33.3%
DTW	VSM	+9.2%	+1.7%	-14.5%	-0.1%	-2.5%	+0.5%	+15.2%
DTW	PRF	-28.5%	-5.2%	-26.7%	-9.9%	-22.3%	-11.4%	-13.6%
PRFDTW	VSM	+59.2%	+10.3%	+14.2%	+10.8%	+14.1%	+12.2%	+45.2%
PRFDTW	PRF	+4.2%	+2.8%	-2.1%	-0.1%	-9.1%	-1.1%	+8.9%
PRFDTW	DTW	+45.8%	+8.5%	+33.5%	+10.9%	+17.0%	+11.6%	+26.0%

Table VI.

methods X	methods Y	QA	QA-AP90
PRF	VSM	+6.2%	+1.7%
DTW	VSM	+12.7%	+5.4%
DTW	PRF	+6.1%	+3.7%
PRFDTW	VSM	+16.6%	+6.2%
PRFDTW	PRF	+9.8%	+4.5%
PRFDTW	DTW	+3.5%	+0.7%

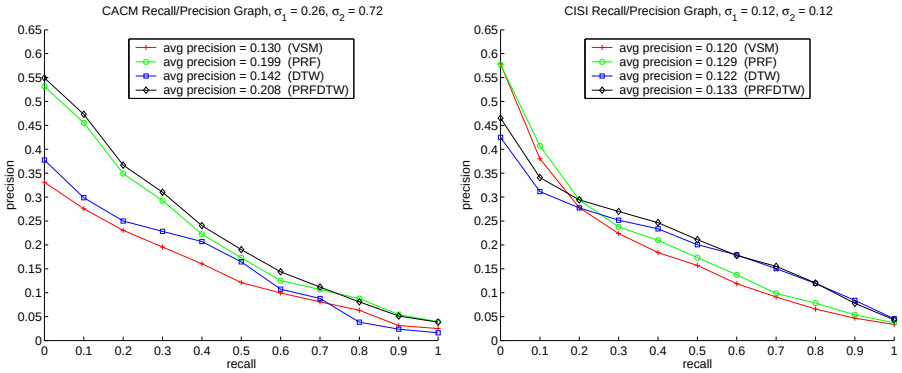

Figure 3. CACM and CISI: recall/precision graph for DTW methods

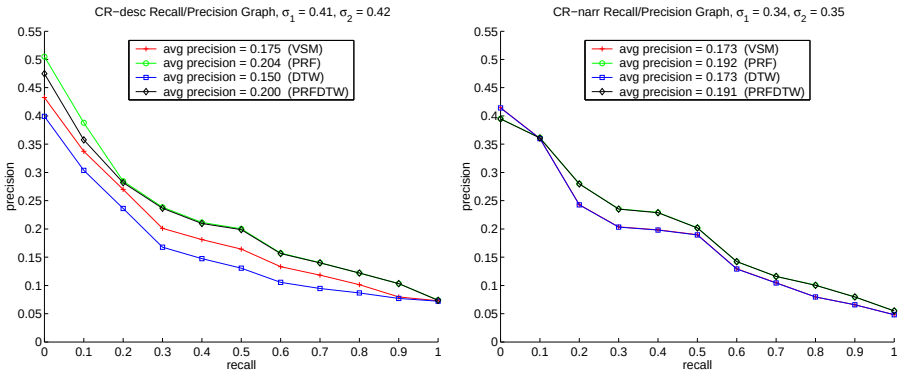

Figure 4. CR-desc and CR-narr: recall/precision graph for DTW methods

7.1.2. *Significance Testing*

The next step for the evaluation is the comparison of the values of the average precision obtained by different methods. Statistical tests provide information about wether observed differences in different methods are really significant or just by chance. Several statistical tests have been used in IR (Hull, 1993), (Yang and Liu, 1999). We employ the "paired t-test" described in (Hull, 1993).

The results are shown in table VII. Each row contains the results of two tests, i.e. test method X against method Y and vice versa. We tested each method X against each other method Y and vice versa. We used significance levels $\alpha = 0.05$ and $\alpha = 0.01$.

- An entry of ++ in a table cell indicates that the null hypothesis is rejected for testing X against Y at significance level $\alpha = 0.01$. This means that method X is almost guaranteed to perform better than method Y.
- An entry of + in a table cell indicates that the null hypothesis is rejected for testing X against Y at significance level $\alpha = 0.05$, but can not be rejected at significance level $\alpha = 0.01$. This means that method X is likely to perform better than method Y.
- An entry of o in a table cell indicates that the null hypothesis can not be rejected in both test. This means that there is low probability that one of the methods is performing better than the other method.
- An entry of − in a table cell indicates that the null hypothesis is rejected for testing Y against X at significance level $\alpha = 0.05$, but can not be rejected at significance level $\alpha = 0.01$. This means that method Y is likely to perform better than method X.
- An entry of −− in a table cell indicates that the null hypothesis is rejected for testing Y against X at significance level $\alpha = 0.01$. This means that method Y is almost guaranteed to perform better than method X.

Table VII. Paired t-test results for significance levels $\alpha = 0.05$ and $\alpha = 0.01$

methods X	methods Y	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
PRF	VSM	++	++	++	+	+	++	+	++	+
DTW	VSM	o	o	o	o	o	o	o	++	++
DTW	PRF	−	o	−	−	—	—	o	++	++
PRFDTW	VSM	++	o	+	o	o	++	++	++	++
PRFDTW	PRF	o	o	o	o	o	o	o	++	++
PRFDTW	DTW	++	o	+	+	o	++	+	++	o

8. Query Term Reweighting Experiments

In this section we describe a query expansion method based on query similarity and query term reweighting and relevant documents. The method is denoted as QTW from now on.

We will first give a high level description of the method, then the detailed mathematical description is given.

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 and non-relevant documents from the ground truth data
- from this set of relevant documents compute a term reweighting matrix such that relevant documents are ranked higher than non-relevant documents
- use this term reweighting matrix to compute the ranking of the documents for the new query

The formal description is given here. Let S be the set

$$S = \{q_k | sim(q_k, q_l) \geq \sigma, \quad 1 \leq k \leq L\} \tag{32}$$

of existing old queries q_k having a similarity greater than or equal to a threshold σ to the new query q_l and let R_k be the set

$$R_k = \{r_k | q_k \in S, \quad 1 \leq k \leq L\} \tag{33}$$

of all relevance judgements for the queries q_k in S. Write R_k as a matrix of N rows and $|S|$ columns

$$R = (r_k)_{1 \leq k \leq |S|} \tag{34}$$

and compute the query term reweighting matrix $\hat{W}_Q$ according to equations (18) to (20) such that

$$SIM \cdot \hat{W}_Q = R \tag{35}$$

Then do a query term reweighting

$$\hat{Q}_W = Q \cdot \hat{W}_Q \tag{36}$$

according to equation (21) and compute the new document ranking for query q_l according to equation (8)

$$sim_l = sim(D, \hat{Q}_W) \tag{37}$$

Notes:

- if σ in (32) is chosen to high the set S may be empty. Then the set R_k will be empty and we will not compute any query term reweighting matrix $\hat{W}_Q$. In this case we will have our original similarity vector sim_l computed from $sim_l = sim(D, q_l) = D^T \cdot q_l$.
- parameter σ in (32) is a tuning parameter for method QTW.

8.1. Experimental Results

In this section the results of the experiments are presented. Results were evaluated using the average precision over all queries. Recall/precision graphs were generated according to (Baeza-Yates and Ribeiro-Neto, 1999). Then significance tests were applied to the results.

8.1.1. *Results*

Methods VSM, PRF and QTW (query term reweighting and relevant documents) were applied. Best parameter value settings for parameters α and θ for method PRF had been obtained previously by experiment and those which give the highest average precision were selected and used (see table II and (Kise et al., 2001)).

Method QTW has been evaluated using different settings for parameters σ. From the set of queries contained in each collection we selected each query one after the other and treated it as a new query $q_l, 1 \leq l \leq L$. Then for each fixed query q_l we computed the similarity $\sigma_k := sim(q_k, q_l)$ for all queries $q_k, 1 \leq k \leq L, k \neq l$ according to equation (1). Then σ in equation (32) has been varied from 0.0 up to 1.0 in steps of 0.01. Finally we got our query term reweighting matrix according to equation (35) and issued the query. Best parameter values for σ are reported in table VIII.

Table VIII. Best parameter values for method QTW

	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
σ	0.25	0.41	0.41	0.34	0.41	0.49	0.36	0.78	0.68

In the next step we combined two methods of query expansion in this way: First, after having expanded the new query using the QTW method, we applied the PRF method against the expanded query. This method is reported as the QTWPRF method. Second, after having expanded the new query using the PRF method, we applied the QTW method against the expanded query. This method is reported as the PRFQTW method. Best parameter values settings have again been

obtained by experiment and are chosen such that average precision is highest. These parameter value settings are reported in table IX.

Table IX. Best parameter values for methods PRFQTW and QTWPRF

		CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
QTWPRF	σ	same as in table VIII								
	α	1.70	0.50	0.50	0.40	0.20	0.90	0.40	0.30	0.30
	θ	0.45	0.80	0.85	0.95	0.80	0.85	0.00	0.65	0.80
PRFQTW	α, θ	same as in table II								
	σ	0.65	0.40	0.36	0.35	0.46	0.65	0.46	0.80	0.69

Table X 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.

Table X. Average precision obtained in different methods

	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
VSM	0.130	0.120	0.175	0.173	0.135	0.384	0.085	0.645	0.745
PRF	0.199	0.129	*0.204*	**0.192**	*0.169*	0.435	0.113	0.685	0.757
QTW	0.155	0.133	0.150	0.173	0.133	0.420	0.106	0.716	0.808
QTWPRF	*0.212*	**0.138**	0.179	*0.192*	0.163	*0.452*	**0.154**	**0.740**	**0.815**
PRFQTW	**0.231**	*0.136*	**0.221**	0.191	**0.180**	**0.454**	*0.127*	*0.739*	*0.810*

Table XI compares the ratios of average precision improvements or degradations obtained by the different methods. Refer to subsection 7.1.1 on page 18 for a description of the table contents.

Table XI. Average precision comparison

methods		CACM	CISI	CR-desc	CR-narr	CR-title	CRAN
X	Y						
PRF	VSM	+52.7%	+7.3%	+16.6%	+10.9%	+25.5%	+13.4%
QTW	VSM	+18.5%	+10.7%	-14.5%	-0.1%	-1.6%	+9.5%
QTW	PRF	-22.4%	+3.2%	-26.7%	-9.9%	-21.6%	-3.5%
QTWPRF	VSM	+62.3%	+15.0%	+2.1%	+10.8%	+21.1%	+17.8%
QTWPRF	PRF	+6.3%	+7.2%	-12.5%	-0.1%	-3.5%	+3.9%
QTWPRF	QTW	+37.0%	+3.9%	+19.4%	+11.0%	+23.1%	+7.6%
PRFQTW	VSM	+77.2%	+13.3%	+26.0%	+10.8%	+33.6%	+18.2%
PRFQTW	PRF	+16.0%	+5.6%	+8.0%	-0.1%	+6.4%	+4.2%
PRFQTW	QTW	+49.6%	+2.4%	+47.4%	+10.9%	+35.8%	+7.9%
PRFQTW	QTWPRF	+9.2%	-1.5%	+23.4%	-0.0%	+10.3%	+0.3%

Table XI.

methods X	methods Y	FR	QA	QA-AP90
PRF	VSM	+33.3%	+6.2%	+1.7%
QTW	VSM	+24.7%	+10.9%	+8.5%
QTW	PRF	-6.5%	+4.5%	+6.7%
QTWPRF	VSM	+81.6%	+14.6%	+9.4%
QTWPRF	PRF	+36.2%	+8.0%	+7.6%
QTWPRF	QTW	+45.6%	+3.3%	+0.9%
PRFQTW	VSM	+49.3%	+14.6%	+8.7%
PRFQTW	PRF	+12.0%	+7.9%	+6.9%
PRFQTW	QTW	+19.7%	+3.3%	+0.2%
PRFQTW	QTWPRF	-17.8%	-0.1%	-0.7%

Figures 5 and 6 show some of the recall/precision graphs for the test collections. Each figure contains the graphs for methods VSM, PRF, QTW, PRFQTW and QTWPRF.

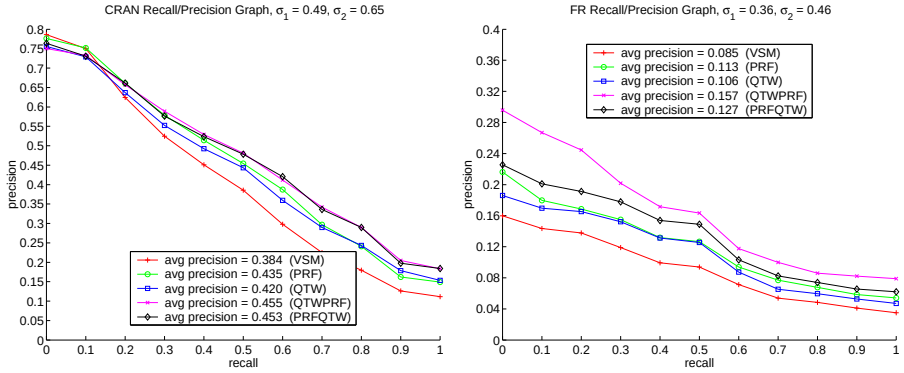

Figure 5. CRAN and FR: recall/precision graph for QTW methods

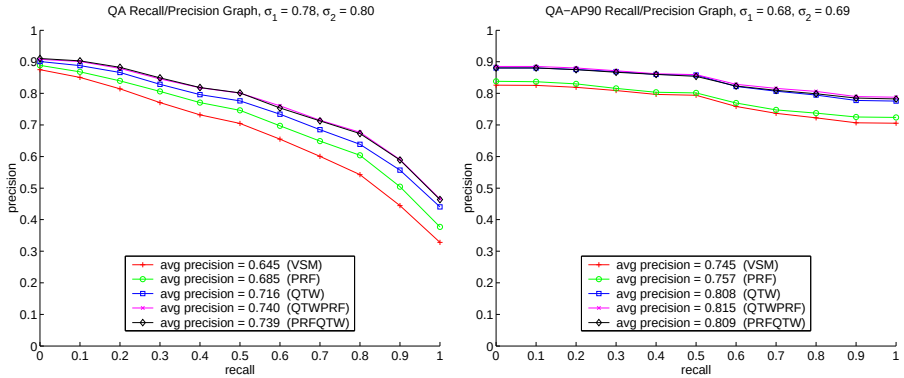

Figure 6. QA and QA-AP90: recall/precision graph for QTW methods

8.1.2. *Significance Testing*
The results are shown in table XII. Each row contains the results of two tests, i.e. test method X against method Y and vice versa. We tested each method X against each other method Y and vice versa. We used significance levels $\alpha = 0.05$ and $\alpha = 0.01$.

Table XII. Paired t-test results for significance levels $\alpha = 0.05$ and $\alpha = 0.01$

methods X	methods Y	CACM	CISI	CR-desc	CR-narr	CR-title	CRAN	FR	QA	QA-AP90
PRF	VSM	++	++	++	+	+	++	+	++	+
QTW	VSM	o	o	o	o	o	++	+	++	++
QTW	PRF	–	o	–	–	—	o	o	++	++
QTWPRF	VSM	++	+	o	+	+	++	++	++	++
QTWPRF	PRF	o	o	o	o	o	o	+	++	++
QTWPRF	QTW	++	+	++	+	+	++	+	++	o
PRFQTW	VSM	++	o	++	o	+	++	++	++	++
PRFQTW	PRF	+	o	o	o	o	+	o	++	++
PRFQTW	QTW	++	o	+	+	++	++	o	++	o
PRFQTW	QTWPRF	o	o	o	o	+	o	o	o	–

9. Summary

Our findings are as follows:

- The results show that with the optimal parameter setting PRF always performs better than the pure VSM model in our test sets.
- The DTW method outperforms the VSM and the PRF method in two cases, but is also significantly worse than the PRF method in 5 cases.
- The QTW method outperforms the VSM method in four cases and the PRF method in two cases, but is also significantly worse than the PRF method in 4 cases.
- The combined method PRFDTW outperforms the PRF method in two cases, and in no case it is significantly worse than the PRF method.
- The combined methods PRFQTW and QTWPRF outperform the PRF method in 4 and 3 cases, and in no case they are significantly worse than the PRF method.
- QTW slightly outperforms DTW in most cases. This is also true for the combined version PRFQTW which mostly outperforms PRFDTW.
- Generally, the best results are obtained with the approaches combining QTW with PRF.

The significant improvements of the new methods are mostly in the QA and QA-AP90 collections. We assume that the special construction of the queries in these collections (see section 6.1) is the main factor for these improvements, since our tuning parameter σ (see equations 26 and 32) has a much higher value than in the other collections (see tables III and VIII).

We did some analysis in order to explain the different performances of our approaches in the collections taking the properties of the test set into account. So far, we have not been able to find a clear correlation between measurable properties of the test sets and the results. We expect the following factors to be crucial for good results:

- the query lengths
- the overlap of "query content" in a test collection (number of queries and number of pairs of queries with high similarities)
- the number of relevant documents for queries with some "overlap in content" (some documents being relevant to different queries)

We expect that a real world information retrieval system has a relatively large overlap in user interests and queries. In Collaborative Information Retrieval (CIR) we want to benefit from this overlap by exploiting users' search processes for subsequent searches. As an initial step towards techniques we focused on a restricted CIR scenario in which only user queries and the relevant answer documents to these queries are known. The two approaches DTW and QTW that we developed for this scenario were tested on queries of standard Information Retrieval test collections. Although in these collections we do not have the query and interest distribution that we assume to have in real world systems, the approaches show relatively good results, in particular if they are combined with pseudo relevance feedback. It turns out that the QTW performs slightly better.

As one of our next steps for the DTW/QTW approaches we want to learn the similarity measure between queries based on training examples. Two more topics that we will work on in the future are the removal of the explicit parameters in the DTW and QTW approach as well as shifting towards real world retrieval systems. For the latter we are cooperating with a search engine provider.

10. Acknowledgements

This work was supported by the German Ministry for Education and Research, bmb+f (Grants 01 IN 902 B8 and 01 IW C01).

References

Baeza-Yates, R. and B. Ribeiro-Neto: 1999, *Modern Information Retrieval*. Addison-Wesley Publishing Company.

Ben-Israel, A. and T. N. E. Greville: 1974, *Generalized inverses: theory and applications*. New York: Wiley-Interscience [John Wiley & Sons]. (reprinted by Robert E. Krieger Publishing Co. Inc., Huntington, NY, 1980.).

Crestani, F. and C. J. van Rijsbergen: 1995a, 'Information Retrieval by Imaging'. *Journal of Documentation* **51**(1), 1–15.

Crestani, F. and C. J. van Rijsbergen: 1995b, 'Probability Kinematics in Information Retrieval: A Case Study'. In: E. A. Fox, P. Ingwersen, and R. Fidel (eds.): *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, Washington, USA, pp. 291–299, ACM Press, New York, NY, USA.

Crestani, F. and C. J. van Rijsbergen: 1998, 'A study of probability kinematics in information retrieval'. *ACM Transactions on Information Systems (TOIS)* **16**(3), 225–255.

Cui, H., J.-R. Wen, J.-Y. Nieand, and W.-Y. Ma: 2002, 'Probabilistic Query Expansion Using Query Logs'. In: *Eleventh International World Wide Web Conference*. Honolulu, Hawaii, USA.

Dominich, S.: 2001, *Relevance Effectiveness in Information Retrieval*, chapter 5, pp. 215–232, Mathematical Foundations of Information Retrieval. Dordrecht, Boston, London: Kluwer Academic Publishers.

Efthimiadis, E. N.: 1996, 'Query expansion'. *Annual Review of Information Science and Technology* **31**, 121–187.

Harman, D.: 1992a, 'Relevance Feedback and Other Query Modification Techniques'. In: W. B. Frakes and R. Baeza-Yates (eds.): *Information Retrieval - Data Structures & Algorithms*. New Jersey, pp. 241–263, Prentice Hall.

Harman, D.: 1992b, 'Relevance feedback revisited'. In: N. Belkin, P. Ingwersen, and A. M. Pejtersen (eds.): *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Copenhagen, Denmark, pp. 1–10, ACM Press, New York, NY, USA.

Hull, D.: 1993, 'Using Statistical Testing in the Evaluation of Retrieval Experiments'. In: R. Korfhage, E. Rasmussen, and P. Willett (eds.): *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburgh, Pennsylvania, USA, pp. 329–338, ACM Press, New York, NY, USA.

Joachims, T.: 2002, 'Unbiased Evaluation of Retrieval Quality using Clickthrough Data'. Technical report, Cornell University, Department of Computer Science.

Kise, K., M. Junker, A. Dengel, and K. Matsumoto: 2001, 'Experimental Evaluation of Passage-Based Document Retrieval'. In: *Proceedings of the Sixth International Conference on Document Analysis and Recognition ICDAR-01*. Seattle, WA, pp. 592–596.

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

Raghavan, V. V. and H. Sever: 1995, 'On the Reuse of Past Optimal Queries'. In: E. A. Fox, P. Ingwersen, and R. Fidel (eds.): *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, Washington, USA, pp. 344–350, ACM Press, New York, NY, USA.

Salton, G., E. A. Fox, and E. M. Voorhees: 1985, 'Advanced feedback methods in Information Retrieval'. *Journal of the American Society for Information Science* **36**(3), 200–210.

Smart: 1968–1988, 'FTP Directory at Cornell University'. Homepage. `ftp://ftp.cs.cornell.edu/pub/smart`.

Trec: 1992–2003, 'Text REtrieval Conference (TREC)'. Homepage. `http://trec.nist.gov`.

van Rijsbergen, C. J.: 1986, 'A non classical logic for Information Retrieval'. *The Computer Journal* **29**(6), 481–485.

van Rijsbergen, C. J.: 1989, 'Towards an information logic'. In: N. J. Belkin and C. J. van Rijsbergen (eds.): *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Cambridge, Massachusetts, USA, pp. 77–86, ACM Press, New York, NY, USA.

Voorhees, E. M. and D. K. Harman: 1999, 'Overview of the Eighth Text Retrieval Conference (TREC-8)'. In: E. M. Voorhees and D. K. Harman (eds.): *NIST Special Publication 500-246: Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. Gaithersburg, MD, pp. 1–23, National Institute of Standards and Technology.

Voorhees, E. M. and D. K. Harman: 2000, 'Overview of the Ninth Text Retrieval Conference (TREC-9)'. In: E. M. Voorhees and D. K. Harman (eds.): *NIST Special Publication 500-249: Proceedings of the Ninth Text Retrieval Conference (TREC-9)*. Gaithersburg, MD, pp. 1–13, National Institute of Standards and Technology.

Wen, J.-R., J.-Y. Nie, and H.-J. Zhang: 2001, 'Clustering user queries of a search engine'. In: *Proceedings of the 10th International World Wide Web Conference*. Hong Kong, pp. 162–168.

Wen, J.-R., J.-Y. Nie, and H.-J. Zhang: 2002, 'Query clustering using user logs'. *ACM Transactions on Information Systems* **20**(1), 59–81.

White, R. W., I. Ruthven, and J. M. Jose: March 2002, 'The Use of Implicit Evidence for Relevance Feedback in Web Retrieval'. In: F. Crestani, M. Girolami, and C. J. van Rijsbergen (eds.): *Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research, ECIR 2002, Proceedings*, Vol. 2291 of *Lecture Notes in Computer Science*. Glasgow, UK, pp. 93–109, Springer.

Xu, J. and W. B. Croft: 1996, 'Query Expansion Using Local and Global Document Analysis'. In: H.-P. Frei, D. Harman, P. Schabie, and R. Wilkinson (eds.): *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Zurich, Switzerland, pp. 4–11, ACM Press, New York, NY, USA.

Xu, J. and W. B. Croft: 2000, 'Improving the effectiveness of information retrieval with local context analysis'. *ACM Transactions on Information Systems* **18**(1), 79–112.

Yang, Y. and X. Liu: 1999, 'A re-examination of text categorization methods'. In: F. Gey, M. Hearst, and R. Tong (eds.): *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley, California, USA, pp. 42–49, ACM Press, New York, NY, USA.

Address for Offprints:
KLUWER ACADEMIC PUBLISHERS PrePress Department,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands
e-mail: TEXHELP@WKAP.NL
Fax: +31 78 6392500