Introducing Query Expansion Methods for Collaborative Information Retrieval

Armin Hust

German Research Center for Artificial Intelligence (DFKI GmbH)
P.O. Box 2080, 67608 Kaiserslautern, Germany
armin.hust@dfki.de

Abstract. The accuracy of ad-hoc document retrieval systems has plateaued in the last few years. At DFKI, we are working on so-called collaborative information retrieval (CIR) systems which unobtrusively learn from their users' search processes. We focus on a restricted setting in CIR in which only old queries and correct answer documents to these queries are available for improving a new query. For this restricted setting we propose new approaches for query expansion procedures.
This paper describes query expansion methods to be used in collaborative information retrieval. We define collaborative information retrieval as a task, where an information retrieval system uses information gathered from previous search processes from one or several users to improve retrieval performance for the current user searching for information. We show how collaboration of individual users can improve overall information retrieval performance. Performance in this case is expressed in terms of quality and utility of the retrieved information regardless of specific user groups.

1 Introduction

In this section we introduce the research area of Collaborative Information Retrieval (CIR). We motivate and characterize the primary goals of this work, query expansion procedures for CIR and outline the structure and contents of this work.

1.1 Information Retrieval

Although Information Retrieval has now been studied for decades there is no clear and comprehensive definition for Information Retrieval.

One of the older definitions, referenced by Cornelius J. van Rijsbergen [50], refers to the book of F.W. Lancaster [31], where it is stated that "Information retrieval is the term conventionally, though somewhat inaccurately, applied to the type of activity discussed in this volume. An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request."

A newer definition, according to the IR-group of the German Informatics Society [14], states that "IR considers information systems according to their role in the knowledge transfer process from a human knowledge producer to an information seeker. The problems arising from vague queries and uncertain knowledge are the main focus of the IR-group. Vague Queries are characterized by the fact that answers to these queries are a priori not uniquely defined (...). The uncertainty and/or the incompleteness of the knowledge often results from a restricted representation of its semantics, since the representation of the knowledge is not limited to some special forms (e.g. text documents, multimedia documents, facts, rules, semantic nets). Additionally IR considers applications where the stored knowledge itself may be uncertain or incomplete (e.g. technical or scientific data sets)" and states that "From these problems the necessity for an evaluation of the quality of the answers of an information system arises, where the utility of the system according to the support for the users with respect to solving their problems has to be considered."

This definition is very general. It stresses the vagueness and uncertainty of stored knowledge and queries. It also stresses the utility of the retrieved information for the users, helping them to solve their problems.

Utility is an idea introduced by the von Neumann-Morgenstern utility theory [53] and is closely connected with their idea of preference relations, both of which come from the field of economics. A preference relation and a utility function can be seen, from a global point of view, as equivalent for an informal and a formal description of the same concept. A preference relation is a partial order on a set of elements with a binary relation between each two elements. A preference relation can describe such statements like "situation B is better than situation A". Whereas a preference relation is only a qualitative measure, the idea of the utility function introduces the quantitative measure. The utility function assigns a number to each element of the set and allows us to compare the utility of these elements, i.e. the utility function adds a cardinality aspect to the preference relation aspect, such that one can say how useful the choice is. If we can formalize the utility function an a set of elements, then it naturally induces a preference relation on that set. In this sense the quality of the answers of an IR system can be measured.

Let us state an example to show the different preference relations users may have. A physician, a chemist and a lawyer may query an IR system for information about the medicament "Lipobay" or its American name "Baycol". While the physician may be interested in medication, indication and contra-indication, the chemist may by interested in chemical structure and undergoing reactions of the active ingredient; the lawyer may be interested in legal cases, lawsuits, court decisions and compensations. It is clear that each of these users has his or her own personal preferences as to which documents an IR system presents in response to the query. These preferences may also be influenced by the context the user is working in and it is clearly possible that they may change somehow

over time.

Another research area overlapping with the IR area (see figure 1) is the usage of context knowledge for a more detailed specification of the information need. Because queries can be vague, it might be possible to use knowledge about the context the user is working in to influence the query processing and achieve better retrieval results. Some of the aspects of the user's context (according to [18]) are: which tasks the user is busy with at the time of the query, which documents have been viewed within the last few minutes, which document is currently being processed by the user. Research in this area integrates modelling and representation of the context information, and integrates this information into the IR processes.

Much work has been done on improving IR systems, in particular in the Text Retrieval Conference series (TREC) [49]. 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 [55]. We are working on new approaches 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.

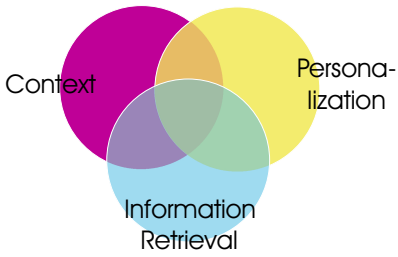

Fig. 1. Overlap of Research Areas

1.2 Collaborative Information Retrieval

The ultimate goal in IR is finding the documents that are useful to the information need expressed as a query. There is a natural preference relation, namely "document A is more useful than document B" or "documents A and B are equally useful" for the information need stated by the user. An approach for formalizing this preference relation is the concept of "relevance" introduced for quality measures of IR systems.

Here relevance has two meanings: On the one hand it means the judgements made by assessors to determine the documents which are satisfying the information need. As we will see later (in section 4.1) relevance judgements are established after experiments are carried out. Assessors review the documents that are retrieved by an experiment and judge the documents as either relevant or non-relevant to the information need expressed as a query. In doing so they neglect personal interests or context related information, thus ignoring personal preference relations a user may have (as stated in the example above) [55]. On the other hand a relevance value is computed as a numerical value and assigned to each of the documents presented by an IR system in response to a user's query. Then documents are ranked according to their relevance in the same way as currently done by web search engines. This so computed relevance value should reflect the user's preference relation.

We call our approach Collaborative Information Retrieval (CIR), learning to improve retrieval effectiveness from the interaction of different users with the retrieval engine. CIR on top of an IR system uses all the methodologies that have been developed in this research field. Moreover, CIR is a methodology where an IR system makes full use of all the additional information available in the system, especially

- the information from previous search processes, i.e. individual queries and complete search processes
- the relevance information gathered during previous search processes, independent of the method used to obtain this relevance information i.e. explicitly by user relevance feedback or implicitly by unobtrusively detected relevance information.

The collaborative aspect here differs from other collaborative processes. We don't 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.

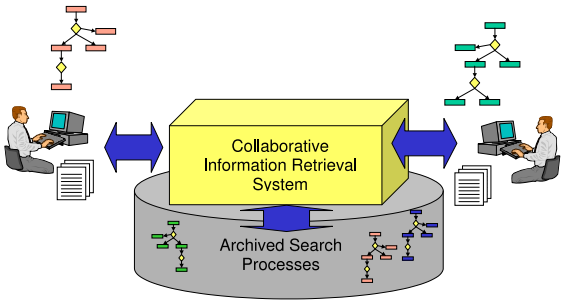

Fig. 2. Scenario of CIR

Figure 2 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 so that he has to reformulate his query taking into consideration the answer documents found. Such refinement may consist of specializations as well as generalizations of previous queries. In general, satisfying an information need means going through a search process with many decisions on query reformulations. Hence gathering information for fulfilling the information need of a user is an expensive operation in terms of time required and resources used. 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.

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 if the following basic assumptions can be fulfilled by an CIR system:

- relevance judgements for retrieved documents can be derived from users' actions

– previous queries by some users will be useful to improve new queries for other users

Subject to these assumptions we expect that collaborative searches will improve overall retrieval quality for all users.

Thus we can see a CIR system as a trusted and experienced advisor where we request help for fulfilling our information need. We explain the new task to the system and hope that the advisor has gathered previous experiences with similar information needs.

Our query expansion methods for CIR realize the advisor: the users describe their information need as a query. We then find previous queries similar to the new query and the documents that have been judged as relevant to the previous queries. Instead of running the original query entered by the users we expand the query by terms gathered from relevant documents of previous queries.

1.3 Delimitation of Collaborative Information Filtering

The description here follows the papers of Alsaffar et al. [2], Olsson [34] and Tian et al. [48]. A comprehensive overview of research papers in the field of Information Filtering is available from Thornton [47]. The objective of information filtering (IF) is to classify/categorize documents as they arrive in the system. IF is the process that monitors documents as they enter the system and selects only those that match the user query (also known as a user profile). Thus, IF makes decisions about relevance or non-relevance rather than providing a ranked output list. In IF the document collection can be seen as a stream of documents trying to reach the user, and unwanted documents are removed from the stream. The collaborative approach, called Collaborative Information Filtering (CIF) takes into account user preferences of other "like-minded" users.

While in CIR as described above the user query is the central focus point, in CIF the documents are central. CIF can be described as a "push" technology, where documents are pushed against the user query (or user profile), while CIR is a "pull" technology, drawing the relevant documents from the collection.

1.4 Outline of this work

In this work we limit ourselves to the text retrieval field, sometimes also called text mining, which is only a part of the information retrieval research area. As a first approach to CIR we also limit ourselves to developing, analyzing and evaluating algorithms which can be used for IR effectiveness improvements, based on individual queries which may be stated by different users. In this work we don't consider complete search processes of users, we especially ignore such vague queries, which can only be answered in dialogue by iterative reformulations of the queries (depending on the previous answers of the system).

This paper is organized as follows:

– section 2 describes related work in the field of query expansion.

– section 3 introduces the vector space model and query expansion procedures that have been developed for use in the vector space model.
– section 4 describes the document collections we use for evaluating our new algorithms and includes the evaluation of some basic IR procedures.
– section 5 introduces the environment of Collaborative Information Retrieval and describes the methodology used in the experiments and in the evaluation.
– section 6 shortly describes the algorithms that have been developed to be used in CIR.
– section 7 summarizes the improvements that we have achieved by our different algorithms.
– section 8 summarizes this paper, draws some conclusions, and shows the essential factors for improving retrieval performance in CIR.

2 Related Work

Research in the field of query expansion (QE) procedures has been done for several years now. Usage of short queries in IR produces a shortcoming in the number of documents ranked according to their similarity to the query. Users issuing short queries retrieve only a few relevant documents, since the number of ranked documents is related to the number of appropriate query terms. The more query terms, the more documents are retrieved and ranked according to their similarity to the query [36]. In cases where a high recall is critical, users seldom have many ways to restate their query to retrieve more relevant documents.

Thus IR systems try to reformulate the queries in a semi-automatic or automatic way. Several methods, called query expansion methods, have been proposed to cope with this problem [3], [32]. 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 goal of all query expansion methods is to finally find the optimal query which selects all the relevant documents.

Research in the field of query expansion procedures has been done for several years now. Query expansion is the process of supplementing the original query with additional terms and should lead to an improvement in the performance of IR systems. This process also includes the reweighting of terms in a query after it has been enriched by additional terms. A lot of different procedures have been proposed for manual, automatic or interactive query expansion. Some of the first publications describing query expansion procedures are Sparck-Jones [46], Minker et al. [33] and Rijsbergen [50]. Some of the older procedures are described by Donna Harman in [17] and [16], experiments in the SMART systems have been described by Salton [42] and Buckley [5]. A comprehensive overview of newer procedures is available from Efthimiadis in [12]. Another newer technique, called local context analysis (LCA), was introduced by Xu and Croft in [60] and [61]. While pseudo relevance feedback assumes that all of the highly ranked documents are relevant, LCA assumes that only some of the top ranked documents initially retrieved for a query are relevant and analyzes these documents for term

co-occurrences.

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 (refer to Raghavan, Sever and Alsaffar et al. in [37], [44] [2]). Wen et al. ([57] and [58]) 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. [10] 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. Agichtein et al. [1] are learning search engine specific query transformations for question answering in the web.

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. [59] compare two systems, where one is using explicit relevance feedback (where searchers explicitly have to mark documents relevant) 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. [24] 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 term weights in different IR models have been discussed in [7], [8] and [6], going back to the work of [51] and [52]. The different models analyze the transfer of probabilities in the term space, mainly for, but not limited to, the probabilistic IR models.

3 Basics and Terminology

In this section we introduce the vector space model (VSM) which is employed in our work. We motivate the different techniques that have been applied to the VSM for performance improvements, since this is also the basic model for the development of our CIR methods. For further reading we recommend the books of Cornelius J. van Rijsbergen [50], Ricardo Baeza-Yates and Berthier Ribeiro-Neto [3] and Christopher D. Manning and Hinrich Schütze [32] and the new book from Reginald Ferber [13].

3.1 Vector Space Model

The vector space model, introduced by Salton [40], assigns weights to index terms in queries and in documents. These term weights are ultimately used to compute the degree of similarity between each document stored in the system

and the user query. By sorting the retrieved documents in decreasing order of this degree of similarity, the vector space model takes into consideration documents which match the query terms only partially.

Definition 1 (Vector Space Model). *Documents as well as queries are represented by vectors in a vector space. The set of N documents is denoted by*

$$D = \{d_j | 1 \leq j \leq N\}, \tag{1}$$

the set of L queries is denoted by

$$Q = \{q_k | 1 \leq k \leq L\}. \tag{2}$$

Each individual document d_j is represented by its vector

$$d_j = (d_{1j}, d_{2j}, ..., d_{Mj})^T, \tag{3}$$

each individual query q_k is represented by its vector

$$q_k = (q_{1k}, q_{2k}, ..., q_{Mk})^T, \tag{4}$$

where M is the number of terms in the collection and T denotes the transpose of the vector.
Each position i in the vectors corresponds to a specific term t_i in the collection. The values d_{ij} or q_{ik} respectively indicate the weighted presence or absence of the respective term in the document d_j or query q_k. The weights d_{ij} and q_{ik} are all greater than or equal to 0.
Term weights d_{ij} and q_{ik} in equations (3) and (4) can be computed in many different ways. Different weighting schemes, so called tf-idf weighting schemes, have been developed by Salton and Buckley [41], the older work by Salton and McGill [43] reviews various term-weighting techniques. A newer work by Kolda [29], [30] evaluates different weighting schemes and compares the results achieved by each of the weighting methods. The main idea behind the most effective term weighting schemes is related to the basic principles of clustering techniques [43]. Moreover it allows the usage of different weighting schemes for the document representation and the query representation.

Despite its simplicity, the vector space model is a resilient ranking strategy with general collections. It yields ranked answer sets which are difficult to improve upon without query expansion or relevance feedback within the framework of the vector space model. A large variety of alternative ranking methods have been compared to the vector space model but the consensus seems to be that, in general, the vector space model is either superior or almost as good as the known alternatives. Furthermore, it is simple and fast. For these reasons, the vector space model is a popular retrieval model nowadays.

Definition 2 (Cosine Similarity). *The ranking function normally used in the vector space model is the so called cosine-similarity. The vector space model proposes to evaluate the degree of similarity of the document $d_j = (d_{1j}, d_{2j}, ..., d_{Mj})^T$*

with regard to the query $q_k = (q_{1k}, q_{2k}, ..., q_{Mk})^T$ *as the correlation between the vectors* d_j *and* q_k*. This correlation can be quantified, for instance, by the cosine of the angle between these two vectors. That is, the similarity* sim *between a document* d_j *and a given query* q_k *is measured by the cosine of the angle between these two* M *dimensional vectors:*

$$\begin{aligned} \text{sim} : \mathbb{R}^M \times \mathbb{R}^M &\rightarrow \mathbb{R}^+ \\ (d_j, q_k) &\mapsto \text{sim}(d_j, q_k) \end{aligned} \tag{5}$$

with

$$\text{sim}(d_j, q_k) = \frac{d_j^T \cdot q_k}{\|d_j\| \cdot \|q_k\|} = \frac{\sum_{i=1}^{M} d_{ij} \cdot q_{ik}}{\|d_j\| \cdot \|q_k\|} \tag{6}$$

where $\|\cdot\|$ *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 scalar product between the two vectors* d_j *and* q_k

$$\text{sim}(d_j, q_k) = d_j^T \cdot q_k = \sum_{i=1}^{M} d_{ij} \cdot q_{ik} \tag{7}$$

Figure 3 illustrates (in the 2-dimensional space) the document and query vectors of unit length lying on the surface of the unit-hypersphere. The cosine of angle α measures the similarity between a document and a query.

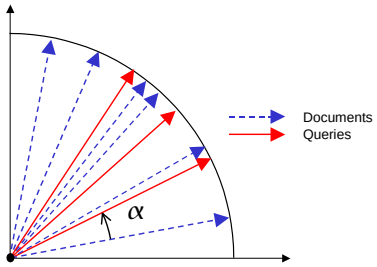

Fig. 3. Similarity and Distance of Documents and Queries

For our purposes we also need to measure the similarity between documents (called inter-document similarity) and between queries (called inter-query similarity). The definitions are comparable to definition 2.

Definition 3 (Inter-Document and Inter-Query Similarity). *The similarity* sim *between two documents* d_j *and* d_k *and between two queries* q_k *and* q_l *is measured by the cosine of the angle between these two* M *dimensional vectors*

$$\begin{aligned} \text{sim} : \mathbb{R}^M \times \mathbb{R}^M &\rightarrow \mathbb{R}^+ \\ (d_j, d_k) &\mapsto \text{sim}(d_j, d_k) \end{aligned} \tag{8}$$

$$\begin{aligned} \text{sim} : \mathbb{R}^M \times \mathbb{R}^M &\rightarrow \mathbb{R}^+ \\ (q_k, q_l) &\mapsto \text{sim}(q_k, q_l) \end{aligned} \tag{9}$$

according to equations 5, 6 and 7.

3.2 Pseudo Relevance Feedback

Pseudo relevance feedback (PRF) avoids the interaction of the IR system with the user after the list of the retrieved documents is created in the first stage. PRF works in three stages: First documents are ranked according to their similarity to the original query. Then highly ranked documents are all assumed to be relevant (refer to [61]) and their terms (all of them or some highly weighted terms) are used for expanding the original query. Then documents are ranked again according to their similarity to the expanded query.

In this work we employ a variant of pseudo relevance feedback described by Kise et al. [25], [26]. In our comparisons with the newly developed methods, we will use the PRF method.

Let E be the set of document vectors given by

$$E = \left\{ d_j \middle| \frac{sim(d_j, q_k)}{\max_{1 \leq i \leq N}\{sim(d_i, q_k)\}} \geq \theta \right\} \tag{10}$$

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

$$D_k = \sum_{d_j \in E} d_j \tag{11}$$

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

$$q'_k = q_k + \alpha \frac{D_k}{\parallel D_k \parallel} \tag{12}$$

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

Parameters θ in equation (10) and α in equation (12) are tuning parameters. During evaluation best parameter value settings have been obtained by experiment and those which give the highest average precision were selected for comparison against other methods.

4 The Text Collections

In this section we describe the contents of the text collections used in the evaluation. We then show some properties of the collections which are limiting factors for the retrieval performance of an IR system.

4.1 Contents of the Text Collections

We use standard document collections and standard queries and questions provided by the SMART project [45] and the TREC (Text REtrieval Conferences) conferences series [49]. In addition we use some special collections that we have generated from the TREC collections to show special effects of our algorithms.

Additionally we use two real world collections that have been gathered especially for these experiments by a company providing a web search engine [35].

In our experiments we used the following 16 collections:

- The SMART collections ADI (articles about information sciences), CACM (articles from 'Communications of the ACM' journal), CISI (articles about information sciences), CRAN (abstracts from aeronautics articles), MED (medical articles) and NPL (articles about electrical engineering).
- The TREC collections CR with 34 queries out of topics 251 - 300 using the "title", "description" and "narrative" topics to investigate the influence of query length, FR with 112 queries out of topics 51 - 300.
- the TREC QA (question answering) collection prepared for the Question Answering track held at the TREC-9 conference [56], the QA-AP90 collection containing only those questions having a relevant answer document in the AP90 (Associated Press articles) document collection, the QA-AP90S collection (extracted from the QA-AP90 collection) having questions with similarity of 0.65 or above to any other question, and the QA-2001 collection prepared for the Question Answering track held at the TREC-10 conference [54].
- The PHIBOT collections PHYSICS (articles about physics) and SCIENCE (articles about sciences except physics) are real world collections gathered by a web search engine [35]. Ground truth data has been gathered from documents the user has clicked on from the list which is presented to the user after the query has been executed.

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 (see section 4.3 for a description of some properties of the collections). So it is a challenging task to show performance improvements for our method.

4.2 Preparation of the Text Collections

Terms used for document and query representation were obtained by stemming and eliminating stopwords. Then document and query vectors were created according to the so called tf-idf weighting scheme (see section 3.1), where the document weights d_{ij} are computed as

$$d_{ij} = \frac{1}{n_j} \cdot tf_{ij} \cdot idf_i \tag{13}$$

where n_j is the normalization factor $n_j = \sqrt{\sum_{i=1}^{M} (tf_{ij} \cdot idf_i)^2}$ and 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)

$$tf_{ij} = \sqrt{f_{ij}} \tag{14}$$

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

$$idf_i = \log \frac{N}{n_i} \tag{15}$$

where n_i is the number of documents containing term t_i and N is the number of documents in the collection and the query weights q_{ik} are computed as

$$q_{ik} = \frac{1}{n_k} \cdot \sqrt{f_{ik}} \tag{16}$$

where n_k is the normalization factor $n_k = \sqrt{\sum_{i=1}^{M} f_{ik}}$ and 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).

4.3 Properties of the Text Collections

Table 1 lists statistics about the collections after stemming and stopword elimination has been carried out, statistics about some of these collections before stemming and stopword elimination can be found in Baeza-Yates [3] and Kise et al. [25], [26].

	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
size(MB)	0.1	1.2	1.4	1.4	1.1	3.8	4.9	20.6
number of documents	82	3204	1460	1400	1033	11429	375	2175
number of terms	340	3029	5755	2882	4315	4415	35312	104891
mean number of terms per document	17.9 (short)	18.4 (short)	38.2 (med)	49.8 (med)	46.6 (med)	17.9 (short)	308.2 (long)	322.1 (long)
number of queries	35	52	112	225	30	93	230	1108
mean number of terms per query	5.7 (med)	9.3 (med)	23.3 (long)	8.5 (med)	9.5 (med)	6.5 (med)	1.9 (short)	2.0 (short)
mean number of relev. documents per query	4.9 (low)	15.3 (med)	27.8 (high)	8.2 (med)	23.2 (high)	22.4 (high)	1.7 (low)	2.0 (low)

	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90	QA-AP90S	QA-2001
size(MB)	93	93	93	69	28.2	3.7	3.7	20.1
number of documents	27922	27922	27922	19860	6025	723	723	4274
number of terms	45717	45717	45717	50866	48381	17502	17502	40626
mean number of terms per document	188.2 (long)	188.2 (long)	188.2 (long)	189.7 (long)	230.7 (long)	201.8 (long)	201.8 (long)	220.5 (long)
number of queries	34	34	34	112	693	353	161	500
mean number of terms per query	7.2 (med)	22.8 (long)	2.9 (short)	9.2 (med)	3.1 (short)	3.2 (short)	3.5 (short)	2.7 (short)
mean number of relev. documents per query	24.8 (high)	24.8 (high)	24.8 (high)	8.4 (med)	16.4 (med)	2.8 (low)	3.2 (low)	8.9 (med)

Table 1. Statistics about the test collections

Evaluation of the Basic Models We used the vector space model (VSM) and the pseudo relevance feedback (PRF) model in our evaluation. Additionally we used the Okapi-BM25 model (OKAPI), and from the dimensionality reduction models we used the latent semantic indexing (LSI) model (refer to [15], [11], [4] and [9]) in the following evaluation (no description of these models is given here because of space shortage).

The OKAPI model is evaluated using the BM25 weighting scheme [39] and the Roberston-Sparck Jones term weights as described in [38]. For the evaluation of the LSI model we used the dimensionality $k = 300$.

First we show the average precision obtained by each of the methods in table 2. For each collection the best value of average precision is indicated by bold font, the second best value is indicated by italic font. Then a recall/precision graph is presented in figure 4.

	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
VSM	0.375	0.130	0.120	0.384	0.525	0.185	*0.616*	*0.569*
PRF	*0.390*	*0.199*	*0.129*	**0.435**	**0.639**	**0.224**	**0.638**	**0.587**
OKAPI	**0.421**	**0.290**	0.128	0.339	0.480	*0.200*	0.535	0.489
LSI	0.376	0.122	**0.132**	*0.424*	*0.597*	0.163	0.615	0.495

	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90	QA-AP90S	QA-2001
VSM	*0.175*	*0.173*	0.135	0.085	*0.645*	0.745	0.643	*0.603*
PRF	**0.204**	**0.192**	**0.169**	*0.113*	**0.685**	**0.757**	*0.661*	**0.614**
OKAPI	0.078	0.055	*0.136*	**0.236**	0.633	*0.751*	**0.666**	0.536
LSI	0.106	0.106	0.096	0.051	0.508	0.709	0.601	0.482

Table 2. Average precision obtained by basic methods

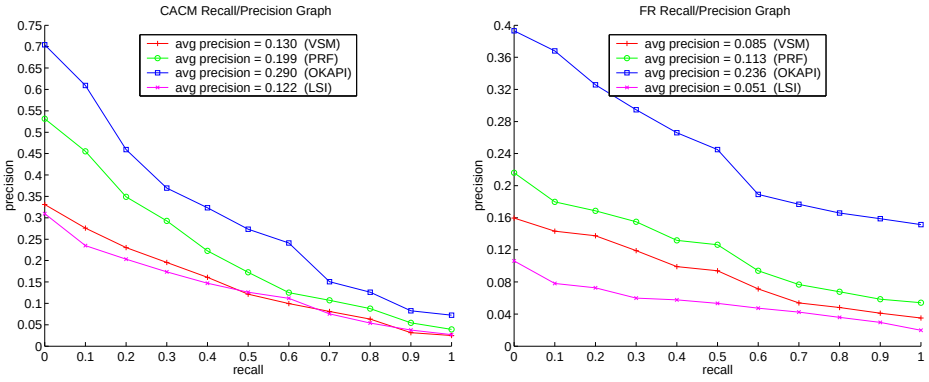

Fig. 4. CACM and FR: recall/precision graphs of basic models

Statistical tests provide information about whether observed differences in different methods are really significant or just by chance. Several statistical tests have been proposed [19], [62]. We employ the "paired t-test" described in [19]. In table 3 we show the significance indicators from statistical testing of the experimental results.

- An entry of $++$ $(--)$ in a table cell indicates that the null hypothesis is rejected for testing X against Y (Y against X) at significance level $\alpha = 0.01$. This means that method X (Y) is almost guaranteed to perform better than method Y (X).
- An entry of $+$ $(-)$ in a table cell indicates that the null hypothesis is rejected for testing X against Y (Y against X) at significance level $\alpha = 0.05$, but can not be rejected at significance level $\alpha = 0.01$. This means that method X (Y) is likely to perform better than method Y (X).

– An entry of o in a table cell indicates that the null hypothesis can not be rejected in both tests. This means that there is low probability that one of the methods is performing better than the other method.

methods X	methods Y	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
PRF	VSM	+	++	++	++	++	++	+	++
OKAPI	VSM	o	++	o	−−	−−	o	−−	−−
OKAPI	PRF	o	++	o	−−	−−	o	−−	−−
LSI	VSM	++	o	++	++	++	−	o	−−
LSI	PRF	o	−−	o	o	−−	−−	−	−−
LSI	OKAPI	o	−−	o	++	++	−	++	o

methods X	methods Y	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90	QA-AP90S	QA-2001
PRF	VSM	++	+	+	+	++	+	o	++
OKAPI	VSM	−−	−−	o	++	o	o	o	−−
OKAPI	PRF	−−	−−	o	++	−−	o	o	−−
LSI	VSM	−	−	o	−	−−	−−	−−	−−
LSI	PRF	−−	−−	−	−−	−−	−−	−−	−−
LSI	OKAPI	o	o	o	−−	−−	−−	−−	−−

Table 3. Paired t-test results for basic methods for significance levels $\alpha = 0.05$ and $\alpha = 0.01$

Analysis of the Results Results achieved by the basic models are non-uniform. From paired t-test results we can see that in most cases PRF performs significantly better than the VSM model. The OKAPI model performs better than VSM and PRF in only two cases, but performs significantly worse in 7 (8) cases than VSM (PRF). LSI performs better than VSM in only 4 cases and worse in 9 cases, and in no case does it perform better than PRF but performs worse than PRF in 13 cases. LSI performs significantly better than OKAPI in only 3 cases, but performs significantly worse in 7 cases.

From average precision analysis we can see that pseudo relevance feedback seems to be the top performer in this evaluation of the basic models. In 11 cases it has the best average precision, and in the other 5 cases it has the second best average precision.

Similarities of Queries to Documents Some of the current limitations of IR can easily be shown. The following graph 5 shows for each query the similarity between the query and the documents. The graph on the left side shows the similarity of each query to its relevant documents. The graph on the right side shows the similarity of each query to its non-relevant documents. The dots show the similarity of an individual document to a query. The thin connecting line shows the average similarity of all relevant (or non-relevant) documents for each query. The thick line averages these similarities over all queries.

From the graph we can see that average similarity of a query to its relevant documents is higher than average similarity of a query to its non-relevant documents. But very often it occurs that there are non-relevant documents having a higher similarity to a query than relevant documents. From this observation it follows that retrieval precision is decreasing if similarity between a query and non-relevant documents is high.

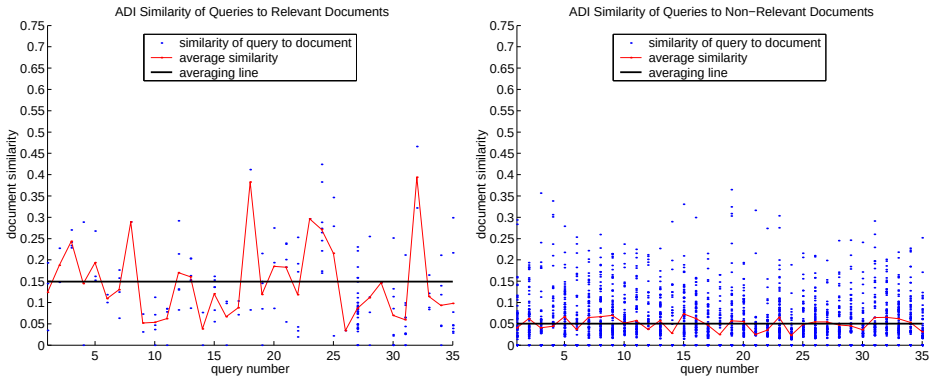

Fig. 5. ADI: similarities of relevant and non-relevant documents

5 Collaborative Information Retrieval

In this section we first explain the motivation for our new approaches to Collaborative Information Retrieval. Then we show the general methodology we are using in our algorithms and delimit the new algorithms from existing procedures. Then we show the general principle of evaluation of our new algorithms.

5.1 Motivation for Collaborative Information Retrieval

In our approach we use 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, and personal preferences and context knowledge are ignored:

- 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.

It does not necessarily follow that if a new query is dissimilar to a previously issued query, the documents which are relevant to the previously issued query are not relevant to the new query.

We will illustrate this fact in a short example with queries taken from the TREC QA text collection. Some of these queries are shown below:

1. What was the name of the first Russian astronaut to do a spacewalk?
2. How many astronauts have been on the moon?

3. What is the name of the second space shuttle?
4. Who was the first woman in space?
5. Name the first Russian astronaut to do a spacewalk.
6. Who was the first Russian astronaut to walk in space?
7. Who was the first Russian to do a spacewalk?

From these queries it is very clear that

- documents being relevant to query 1 are necessarily relevant to queries $5-7$ and vice versa
- documents being relevant to queries $2-4$ are not necessarily relevant to queries 1 and $5-7$

In the following table 4 we show the similarities between the queries (also called the inter-query similarity) with the corresponding similarities between the relevant documents (also called the inter-document similarity) in parenthesis in the second line of each row.

query (document)	1	2	3	4	5	6	7
1	1.0 (1.0)	0.408 (0.138)	0.0 (0.0)	0.0 (0.259)	1.0 (1.0)	0.577 (1.0)	0.816 (1.0)
2	0.408 (0.138)	1.0 (1.0)	0.0 (0.0)	0.0 (0.192)	0.408 (0.183)	0.353 (0.183)	0.0 (0.183)
3	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.5 (0.0)	0.0 (0.0)	0.353 (0.0)	0.0 (0.0)
4	0.0 (0.259)	0.0 (0.192)	0.5 (0.0)	1.0 (1.0)	0.0 (0.259)	0.353 (0.259)	0.0 (0.259)
5	1.0 (1.0)	0.408 (0.183)	0.0 (0.0)	0.0 (0.259)	1.0 (1.0)	0.577 (1.0)	0.816 (1.0)
6	0.577 (1.0)	0.353 (0.183)	0.353 (0.0)	0.353 (0.259)	0.577 (1.0)	1.0 (1.0)	0.353 (1.0)
7	0.816 (1.0)	0.0 (0.183)	0.0 (0.0)	0.0 (0.259)	0.816 (1.0)	0.353 (1.0)	1.0 (1.0)

Table 4. Similarities of sample queries and their relevant documents

Note that for query 3 there are no documents marked as relevant in the text collection and thus the inter-document similarity of the documents relevant to query 3 is defined to be 0.

Our approach to CIR is to find the exact degree of similarity between queries (which of course includes finding the exact degree of dissimilarity) that maximizes the improvements in retrieval performance. We do this by expanding the newly issued query to include terms from previous issued queries and/or documents known as being relevant to the previously issued queries.

5.2 Inter-Query Similarities

In our preliminary considerations for usage of similarities between different queries for retrieval performance improvements we decided to analyze the inter-query similarities. As we already stated at the end of the text collections description (refer to section 4.1) we did not expect to have queries having highly correlated similarities as we would expect in real world applications. In the event, however, the results were even worse than expected.

Indeed, the following histogram in figure 6 shows very low inter-query similarity (as for most of the text collections). The graph on the left side shows the distributions of the query-to-query similarity, including those similarities which are 0. Since this is the dominating factor in each of the graphs, we also produced the same histogram leaving out query-to-query similarities of 0. This is the graph on the right side in each row. For each distribution we also computed the mean and the median value as well as the variance and the standard deviation. These values are shown in the header line of each graph. The vertical lines in the graphs are: the mean similarity (solid line), and the values of the mean similarity ± the standard deviation (dotted lines).

The SMART collections do not incorporate queries having a high similarity to any other query. From the TREC collections only those especially prepared have some high inter-query similarities (refer to the description of the QA, QA-AP90 and QA-AP90S collections in section 4.1). In our real world collections obtained from PHIBOT (the PHYSICS and SCIENCE collection) we also have some high inter-query similarities.

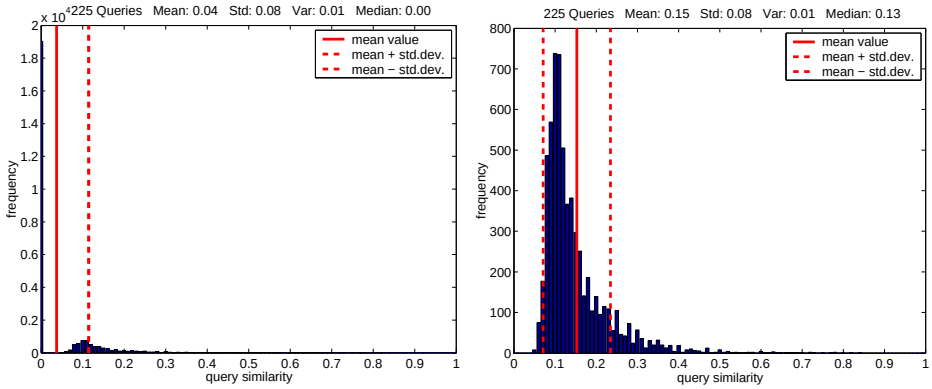

Fig. 6. CRAN: distribution of query similarities

5.3 Correlation between Query Similarities and Document Similarities

In our considerations about similarities between different queries we considered to analyze to inter-query similarities as opposed to the inter-document similarity of the relevant documents. If there were a direct correlation between inter-query similarities to inter-document similarity of the relevant documents, it would be easy to derive the relevant documents for a given new query from the documents being relevant to the existing old queries. This would directly match the expectations stated in the motivation for this chapter (see 5.1).

From these considerations we derived the creation of the following graphs (see figure 7): each graph shows the inter-query similarities for each two pairwise different queries on the x-axis and the inter-document similarity of the relevant documents on the y-axis as a dot. A dot at coordinates $(0.5, 0.9)$ shows that there

are two queries having an inter-query similarity of 0.5 and their relevant documents have an inter-document similarity of (0.9). Another example is a dot at coordinates $(0.4, 0.0)$, which means that there are two queries having an inter-query similarity of 0.4 and their relevant documents have an inter-document similarity of (0.0). The line in each graph is the least-squares estimator for the polynomial of degree 1 fitting best to the clouds of dots.

Here we see that there is no simple correlation between inter-query similarity and inter-document similarity. There are low inter-query similarity and their relevant documents have a high inter-document similarity and vice versa. From the TREC collections only those collections especially prepared have a high correlation between inter-query and inter-document similarity. Also for the PHIBOT collections there seems to be a correlation between a few inter-query similarities to the inter-document similarity, and the least-squares estimator has a low slope.

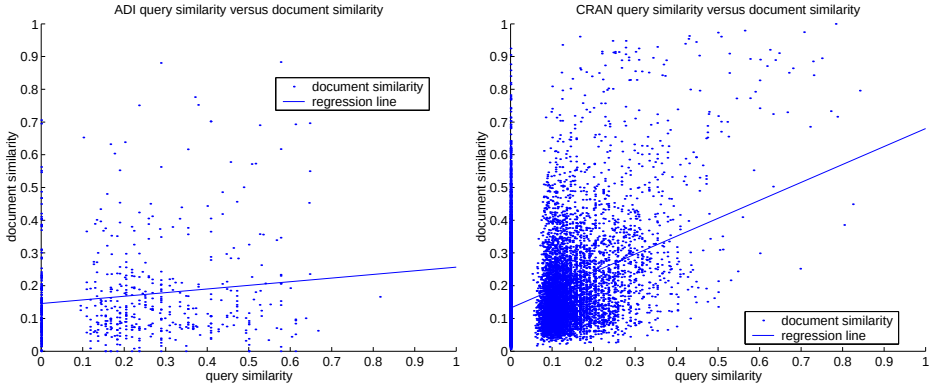

Fig. 7. ADI and CRAN: query similarity vs. document similarity

5.4 Overlap of Relevant Documents

It is essential for achieving retrieval performance improvements to have some "overlapping" relevant documents for pairs of queries. Thus we define the overlap of relevant documents as follows:

Definition 4 (Overlap of Relevant Documents). *Let $q_k, q_l \in Q,\ k \neq l$ be two different queries. Let RD_k, RD_l be the sets of documents being relevant to the queries q_k and q_l respectively. Then the overlap of relevant documents for these two queries is the number of documents in the set $O_{kl} = RD_k \cap RD_l = \{d_j |\ d_j \in RD_k \wedge d_j \in RD_l\}$.*

For all our new query expansion procedures we expect retrieval performance improvements if the overlap of relevant documents is high. The following table 5 gives some statistics about the overlap of relevant documents. Graph 8 shows the individual overlap for each pair of queries. The x-axis and y-axis show the query number and the z-axis shows the overlap for each pair of queries. Since

the overlap between each pair of queries is symmetric ($|O_{kl}| = |O_{lk}|$), we left out the symmetric part for clarity.

	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
pairs of queries	595	1326	6216	25200	435	4278	26335	613278
max overlap	7	17	70	18	0	36	1	4
query pairs with overlap	90	134	1154	686	0	181	25	75
percentage of query pairs with overlap	15.1%	10.1%	18.6%	2.7%	0.0%	4.2%	0.1%	0.01%

	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90S	QA-AP90	QA-2001
pairs of queries	561	561	561	6216	239778	12880	62128	124750
max overlap	27	27	27	10	140	16	16	11
query pairs with overlap	35	35	35	385	760	195	237	259
percentage of query pairs with overlap	6.2%	6.2%	6.2%	6.2%	0.3%	1.5%	0.4%	0.2%

Table 5. Statistics about overlap of relevant documents

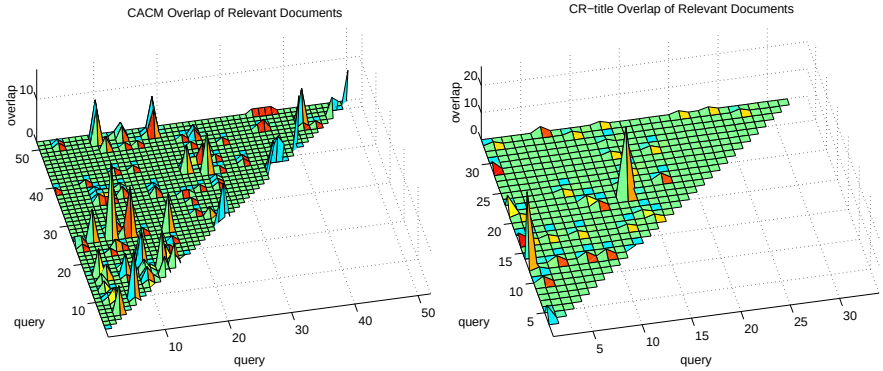

Fig. 8. CACM and CR-title: overlap of relevant documents

5.5 Methodology of Collaborative Information Retrieval Methods

As stated in the motivation for CIR we use global relevance feedback which has been learned from previous queries. Thus we here first describe the query expansion procedures based on query similarities and their relevant documents from a high level approach and will then give a more algorithmic description followed by the formal description.

All our new query expansion procedures work as follows:

- for each new query to be issued 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 this set of relevant documents compute some terms for expansion of the new query
- use this terms to expand the new query and issue the new expanded query

The algorithmic description is given here:

```
for each new query q do
    compute the set S = {q_k | sim(q_k, q) ≥ σ, 1 ≤ k ≤ L}
    compute the sets RD_k = {d_j | q_k ∈ S ∧ d_j is relevant to q_k}
    compute the expanded query q' = cirf(q, S, RD_k)
        by some function cirf
end
```

where S is the set of existing old queries q_k with a similarity of σ or higher to the new query q, RD_k are the sets of the documents being relevant to the queries q_k and f is a function for query expansion.

Our methods in Collaborative Information Retrieval are characterized by a generalized function

$$\begin{aligned} cirf : Q \times 2^{Q \times 2^D} &\rightarrow Q \\ (q, ((q_1, RD_1), (q_2, RD_2), \cdots, (q_L, RD_L))) &\mapsto q' \end{aligned} \tag{17}$$

where the sets RD_i are the sets of documents which are relevant to query q_i.

The goal now is to find suitable functions $cirf$ which can be efficiently computed and which maximize the effectiveness of the new query q' in terms of recall and precision. As is shown in figure 9 our approach is searching for neighbors of the new query. If suitable neighbors of a query q within a given distance are found, we try to derive information about the documents which are relevant to q from its nearest neighbors.

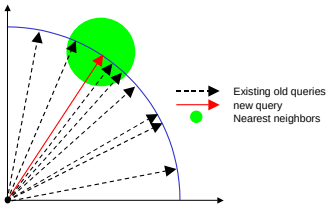

Fig. 9. Motivation for CIR methods: usage of the nearest neighbors

These functions introduce a new level of quality in the IR research area: while the term weighting functions such as tf-idf only work on documents and document collections, and relevance feedback works on a single query and uses information from their assumed relevant and non-relevant documents only, CIR now works on a single query, and uses the information of all other queries and their known relevant documents.

5.6 Methodology of Evaluation Method

The evaluation follows the "leave one out" technique used in several areas such as document classification, machine learning etc.

From the set of L queries contained in each text collection we select each query one after the other and treat it as a new query $q_l, 1 \leq l \leq L$. Then for each fixed query q_l we use the algorithm as described in section 5.5. Of course the now fixed query q_l itself does not take part in the computation of the query expansion. We vary parameters of the algorithms according to suitable values, and select those parameters where highest performance improvements (in terms of average precision over all queries) has been achieved.

6 Query Expansion Methods for CIR

In this section we shortly describe the query expansion methods which we have developed for CIR. For detailed information refer to [21], [22], [23] and [20].

6.1 Methods Description

Query Similarity and Relevant Documents Method QSD (refer to [21] and [23]) uses the relevant documents of the most similar queries for query expansion of a new query. The new query is rewritten as a sum of selected relevant documents of existing old queries, which have a high similarity to the new query.

Query Linear Combination and Relevant Documents Method QLD (refer to [22] and [23]) uses the relevant documents of the most similar queries, which are used in re-writing the new query as a linear combination of the most similar queries. This query expansion method reconstructs the new query as a linear combination of existing old queries. Then the terms of the relevant documents of these existing old queries are used for query expansion.

Document Term Reweighting Method DTW [20] uses the relevant documents of the most similar queries for giving more weight to those ambiguous terms in the documents, that match the semantics of the same terms in the queries. If, for example, the queries use the term 'bank' in conjunction with other terms related to financial topics, then the term 'bank' meaning 'financial institution' will be weighted higher than the term 'bank' meaning 'dike' or 'wall'.

Query Term Reweighting Method QTW [20] uses the relevant documents of the most similar queries for giving more weight to those ambiguous terms in the queries, that match the semantics of the same terms in the documents.

6.2 Other Collaborative Methods

Other methods, denoted as Term Concept Learning (TCL), have been developed in the field of CIR. These methods are used to learn the concept of a term in a new query from the usage of the term in documents which are relevant to existing old queries. These methods are not evaluated herein, refer to [27], [28] and [23] for further information.

6.3 Experiments Description

Methods VSM (vector space model), PRF (pseudo relevance feedback) and the newly developed methods QSD, QLD, DTW and QTW were applied. Best parameter value settings for method PRF had been obtained previously by experiment and those which give the highest average precision were selected and used.

The newly developed methods were evaluated using different settings for parameters σ (see section 5.5) and following our evaluation methodology (see section 5.6). During evaluation best parameter value settings have been obtained by

experiment and those which give the highest average precision were selected for comparison against other methods.

In the next steps we combined two methods of query expansion in this ways: First, after having expanded the new query using the PRF method, we applied one of the methods QSD, QLD, DTW and QTW against the expanded query. These methods are reported as the PRFxxx methods. Second, after having expanded the new query using the QSD, QLD and QTW methods, we applied the PRF method against the expanded query. These methods are reported as the xxxPRF methods.

A recall/precision graph showing results from methods QSD and QLD is shown in figure 10.

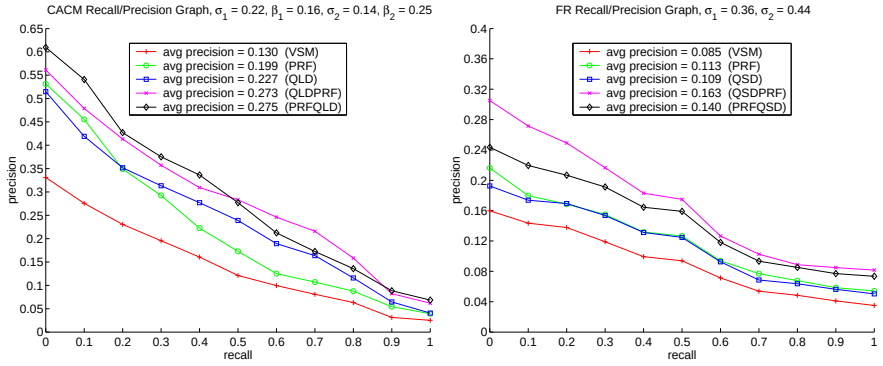

Fig. 10. CACM and FR: recall/precision graphs for QLD and QSD method

7 Improvements

In this section we summarize the results of the query expansion methods developed in this work and analyze the results.

Tables 6 and 7 summarize the results of the newly developed basic methods QSD, QLD, DTW and QTW and show the results from significance testing. For each collection the best value of average precision is indicated by bold font, the second best value is indicated by italic font.

From average precision analysis we can see that methods QSD and QLD perform better than PRF in 7 out of 16 cases, and PRF performs better than any of the newly developed methods in 9 out of 16 cases. Methods DTW and QTW never perform better than QSD or QLD, but in 4 (resp. 5) cases perform better than PRF.

From significance testing we can see that QLD outperforms QSD in 3 cases and in no case performs significantly worse than QSD. QLD performs significantly better than DTW and QTW in 6 cases and performs in no case worse than DTW or QTW.

	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
VSM	*0.375*	0.130	0.120	0.384	*0.525*	0.185	*0.616*	0.645
PRF	**0.390**	0.199	0.129	*0.435*	**0.639**	**0.224**	**0.638**	0.685
QSD	0.374	**0.237**	*0.142*	0.428	0.503	0.184	0.612	*0.727*
QLD	0.369	*0.227*	**0.171**	**0.436**	0.507	*0.185*	0.614	**0.734**
DTW	0.356	0.142	0.122	0.386	0.494	0.182	0.599	0.727
QTW	0.364	0.154	0.131	0.420	0.500	0.183	0.611	0.716

	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90	QA-AP90S	QA-2001
VSM	*0.175*	0.173	0.135	0.085	0.645	0.745	0.643	0.603
PRF	**0.204**	**0.192**	**0.169**	**0.113**	0.685	0.757	0.661	**0.614**
QSD	0.172	0.173	0.152	*0.109*	*0.727*	*0.810*	*0.786*	0.603
QLD	0.175	*0.175*	*0.164*	0.108	**0.734**	**0.812**	**0.789**	*0.603*
DTW	0.150	0.173	0.132	0.098	0.727	0.785	0.732	0.601
QTW	0.150	0.173	0.144	0.106	0.716	0.808	0.762	0.601

Table 6. Average precision obtained in different methods

methods X	methods Y	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
QLD	QSD	o	o	++	o	o	o	o	++
QLD	DTW	o	++	++	++	o	o	++	o
QLD	QTW	o	++	++	+	o	o	o	++
QSD	DTW	o	++	o	++	o	o	++	o
QSD	QTW	o	++	++	+	o	o	o	+
QTW	DTW	o	o	o	++	o	o	++	−−

methods X	methods Y	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90	QA-AP90S	QA-2001
QLD	QSD	o	o	o	o	++	o	o	o
QLD	DTW	o	o	o	o	o	++	++	o
QLD	QTW	o	o	o	o	++	o	++	o
QSD	DTW	o	o	o	o	o	++	++	o
QSD	QTW	o	o	o	o	+	o	++	o
QTW	DTW	o	o	+	o	−−	++	+	o

Table 7. Paired t-test results for significance levels $\alpha = 0.05$ and $\alpha = 0.01$ in different methods

Tables 8 and 9 summarize the results of the combined methods PRFQSD, PRFQLD, PRFDTW, PRFQTW, QSDPRF, QLDPRF and QTWPRF, and show the results from significance testing. In significance testing we only tested each PRFxxx method against each other PRFxxx method as well as we tested each xxxPRF method against each other xxxPRF method. We did no significance testing for testing methods PRFxxx against xxxPRF methods.

From average precision analysis we can see that QLDPRF performs best in 3 out of 16 cases, and PRFQLD performs best in 7 out of 16 cases. For the other 6 cases, PRF performs best in 3 cases, QSDPRF performs best in 2 cases, and in 1 case PRFQTW performs best. Methods PRFQTW and PRFDTW never perform best or second best, except for the 1 case mentioned above. QTWPRF performs second best in 2 cases. In those 9 cases where PRFQLD is not performing best, it is the second best method in 6 cases.

From significance testing we can see that PRFQLD performs significantly better than PRFQSD in 1 case, and in all cases where PRFQSD outperforms the

PRFDTW method, PRFQLD also outperforms this method. In all but for 2 cases where PRFQSD outperforms the PRFQTW method, PRFQLD also outperforms PRFQTW.

	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
VSM	0.375	0.130	0.120	0.384	0.525	0.185	0.616	0.569
PRF	0.390	0.199	0.129	0.435	**0.639**	0.224	**0.638**	**0.587**
PRFQSD	*0.391*	0.256	0.151	*0.463*	0.611	0.223	0.634	0.584
PRFQLD	**0.394**	**0.275**	*0.169*	**0.470**	*0.631*	*0.225*	*0.638*	*0.587*
PRFDTW	0.372	0.208	0.133	0.431	0.602	0.221	0.627	0.575
PRFQTW	0.388	0.231	0.133	0.453	0.606	0.222	0.635	0.583
QSDPRF	0.388	0.257	0.145	0.451	0.609	**0.225**	0.636	0.582
QLDPRF	0.385	*0.273*	**0.173**	0.453	0.613	0.207	0.611	*0.587*
QTWPRF	0.380	0.206	0.137	0.455	0.609	0.224	0.635	0.582

	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90	QA-AP90S	QA-2001
VSM	0.175	0.173	0.135	0.085	0.645	0.745	0.643	0.603
PRF	0.204	0.192	0.169	0.113	0.685	0.757	0.661	*0.614*
PRFQSD	0.196	0.192	0.180	0.140	*0.754*	0.813	0.781	0.613
PRFQLD	*0.208*	**0.193**	**0.190**	0.144	**0.757**	0.814	0.782	**0.615**
PRFDTW	0.200	0.191	0.154	0.123	0.752	0.791	0.733	0.611
PRFQTW	**0.221**	0.191	0.180	0.127	0.739	0.809	0.755	0.612
QSDPRF	0.195	0.191	0.177	**0.163**	0.739	0.813	*0.786*	0.614
QLDPRF	0.204	*0.192*	0.184	*0.161*	0.747	**0.815**	**0.789**	0.613
QTWPRF	0.179	0.192	*0.189*	0.157	0.740	*0.815*	0.764	0.613

Table 8. Average precision obtained in different methods

methods X	methods Y	ADI	CACM	CISI	CRAN	MED	NPL	PHY-SICS	SCI-ENCE
PRFQLD	PRFQSD	o	o	+	o	o	o	o	o
PRFQLD	PRFDTW	+	++	++	++	o	+	+	++
PRFQLD	PRFQTW	o	o	++	+	o	o	o	+
PRFQSD	PRFDTW	o	++	o	++	o	o	+	++
PRFQSD	PRFQTW	o	+	++	++	o	o	o	o
PRFQTW	PRFDTW	+	++	o	++	o	o	+	++
QLDPRF	QSDPRF	o	o	++	o	o	−	−−	o
QLDPRF	QTWPRF	o	++	++	o	o	o	−	++
QSDPRF	QTWPRF	o	++	++	o	o	o	o	o

methods X	methods Y	CR-desc	CR-narr	CR-title	FR	QA	QA-AP90	QA-AP90S	QA-2001
PRFQLD	PRFQSD	o	o	o	o	o	o	o	o
PRFQLD	PRFDTW	o	o	+	+	o	++	++	+
PRFQLD	PRFQTW	o	o	+	+	++	+	++	o
PRFQSD	PRFDTW	o	o	o	+	o	++	++	o
PRFQSD	PRFQTW	o	o	o	+	++	o	++	o
PRFQTW	PRFDTW	o	o	+	o	−−	++	+	o
QLDPRF	QSDPRF	o	o	o	o	++	o	o	o
QLDPRF	QTWPRF	o	o	o	o	o	o	++	o
QSDPRF	QTWPRF	o	o	o	o	o	o	++	o

Table 9. Paired t-test results for significance levels $\alpha = 0.05$ and $\alpha = 0.01$ in different methods

Methods QLDPRF and QSDPRF seem to perform similar. In 12 cases there is no significant improvement or degradation, and in each 2 cases one of these methods is outperforming the other. The QLDPRF method outperforms QTW-PRF in only 4 cases, and performs significantly worse then QTWPRF in 1 case. QSDPRF outperforms QTWPRF in 3 cases, and is in no case significantly worse than QTWPRF.

For a quick overview figure 11 shows the average precision achieved by each method in bar graphs.

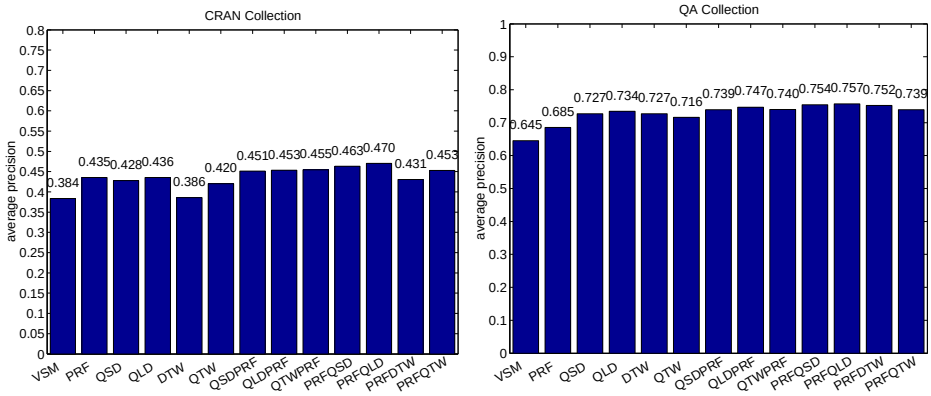

Fig. 11. CRAN and QA: average precision comparison

8 Conclusions

We have studied methods for improving retrieval performance in a restricted Collaborative Information Retrieval environment where information about relevant documents from previous search processes carried out by several users is available for the current query.

Specifically, we developed, evaluated and analyzed new algorithms for query expansion, since query expansion methods are known to be successful in improving retrieval performance.

Results of the newly developed methods are encouraging. Retrieval performance improvements were achieved in most cases. From the basic methods QSD, QLD, DTW and QTW best results were achieved in the combination with the Pseudo Relevance Feedback (PRF) method.

For some text collections no significant retrieval performance improvements could be achieved, neither in the basic methods nor in applying the methods after learning similarity functions.

In the analysis of the results we identified three essential factors for retrieval performance improvements:

1. similarity between queries, also called inter-query similarity (refer to section 5.2)
2. similarity of queries to their relevant documents and similarity of queries to their non-relevant documents (refer to section 4.3)

3. the overlap of relevant documents for pairs of queries (refer to section 5.4)

We think that the first two factors are more important for achieving improvements than the last factor. Best performance improvements have been achieved in text collections, were the inter-query similarity is high, although the overlap in relevant documents is not high.

Low or no retrieval performance improvements were achieved in those cases were the inter-query similarity in the average is low. Also for text collections were similarity of queries to their non-relevant documents is high in the average, we could not achieve high performance improvements.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research, bmb+f (Grant: 01 IN 902 B8).

References

1. Eugene Agichtein, Steve Lawrence, and Luis Gravano. Learning search engine specific query transformations for question answering. In *Proceedings of the 10th International World Wide Web Conference*, pages 169–178, Hong Kong, 2001.
2. Ali H. Alsaffar, Jitender S. Deogun, and Hayri Sever. Optimal queries in information filtering. In *Foundations of Intelligent Systems, 12th International Symposium, ISMIS 2000, Charlotte, NC, USA, October 11-14, 2000, Proceedings*, volume 1932 of *Lecture Notes in Computer Science*, pages 435–443. Springer, 2000.
3. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, 1999.
4. Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup. Matrices, vector spaces, and information retrieval. *Society for Industrial and Applied Mathematics Review*, 41(2):335–362, 1999.
5. Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using smart. In Donna Harman, editor, *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 69–80, Gaithersburg, MD, 1995.
6. Fabio Crestani and Cornelis J. van Rijsbergen. A study of probability kinematics in information retrieval. *ACM Transactions on Information Systems (TOIS)*, 16(3):225–255, 1998.
7. Fabio Crestani and Cornelius J. van Rijsbergen. Information retrieval by imaging. *Journal of Documentation*, 51(1):1–15, 1995.
8. Fabio Crestani and Cornelius J. van Rijsbergen. Probability kinematics in information retrieval: A case study. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–299, Seattle, Washington, USA, July 1995. ACM Press, New York, NY, USA.
9. N. Cristianini, H. Lodhi, and J. Shawe-Taylor. Latent semantic kernels for feature selection, 2000.
10. Hang Cui, Ji-Rong Wen, Jian-Yun Nieand, and Wei-Ying Ma. Probabilistic query expansion using query logs. In *Eleventh International World Wide Web Conference*, Honolulu, Hawaii, USA, May 2002.

11. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science and Technology*, 41(6):391–407, 1990.
12. Efthimis N. Efthimiadis. Query expansion. *Annual Review of Information Science and Technology*, 31:121–187, 1996.
13. Reginald Ferber. *Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. dpunkt.verlag, Heidelberg, 2003.
14. Norbert Fuhr. Goals and tasks of the IR-group. Homepage of the IR-group of the German Informatics Society, 1996. `http://ls6-www.cs.uni-dortmund.de/ir/fgir/mitgliedschaft/brochure2.html`.
15. G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In Yves Chiaramella, editor, *Proceedings of the 11th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 465–480, Grenoble, France, May 1988. ACM Press, New York, NY, USA.
16. Donna Harman. Relevance feedback and other query modification techniques. In William B. Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval - Data Structures & Algorithms*, pages 241–263, New Jersey, 1992. Prentice Hall.
17. Donna Harman. Relevance feedback revisited. In Nicholas Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, Copenhagen, Denmark, June 1992. ACM Press, New York, NY, USA.
18. Andreas Henrich. IR research at university of bayreuth. Homepage of the IR-research group, 2002. `http://ai1.inf.uni-bayreuth.de/forschung/forschungsgebiete/ir_mmdb`.
19. David Hull. Using statistical testing in the evaluation of retrieval experiments. In Robert Korfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, Pittsburgh, Pennsylvania, USA, June 1993. ACM Press, New York, NY, USA.
20. Armin Hust, Markus Junker, and Andreas Dengel. A mathematical model for improving retrieval performance in collaborative information retrieval. 2003. to appear.
21. Armin Hust, Stefan Klink, Markus Junker, and Andreas Dengel. Query expansion for web information retrieval. In Sigrid Schubert, Bernd Reusch, and Norbert Jesse, editors, *Proceedings of Web Information Retrieval Workshop, 32nd Annual Conference of the German Informatics Society*, volume P-19 of *Lecture Notes in Informatics*, pages 176–180, Dortmund, Germany, October 2002. German Informatics Society.
22. Armin Hust, Stefan Klink, Markus Junker, and Andreas Dengel. Query reformulation in collaborative information retrieval. In Marc Boumedine, editor, *Proceedings of the International Conference on Information and Knowledge Sharing, IKS 2002*, pages 95–100, St. Thomas, U.S. Virgin Islands, November 2002. ACTA Press.
23. Armin Hust, Stefan Klink, Markus Junker, and Andreas Dengel. Towards collaborative information retrieval: Three approaches. In Ingrid Renz Jürgen Franke, Gholamreza Nakhaeizadeh, editor, *Text Mining - Theoretical Aspects and Applications*. Physica-Verlag, 2003.
24. Thorsten Joachims. Unbiased evaluation of retrieval quality using clickthrough data. Technical report, Cornell University, Department of Computer Science, 2002.

25. 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*, pages 592–596, Seattle, WA, September 2001.
26. Koichi Kise, Markus Junker, Andreas Dengel, and Keinosuke Matsumoto. Passage-based document retrieval as a tool for text mining with user's information needs. In Klaus P. Jantke and Ayumi Shinohara, editors, *Proceedings of Discovery Science, 4th International Conference DS-2001*, volume 2226 of *Lecture Notes in Computer Science*, pages 155–169, Washington, DC, USA, November 2001. Springer.
27. Stefan Klink, Armin Hust, Markus Junker, and Andreas Dengel. Collaborative learning of term-based concepts for automatic query expansion. In *Proceedings of ECML 2002, 13th European Conference on Machine Learning*, volume 2430 of *Lecture Notes in Artificial Intelligence*, pages 195–206, Helsinki, Finland, August 2002. Springer.
28. Stefan Klink, Armin Hust, Markus Junker, and Andreas Dengel. Improving document retrieval by automatic query expansion using collaborative learning of term-based concepts. In *Proceedings of DAS 2002, 5th International Workshop on Document Analysis Systems*, volume 2423 of *Lecture Notes in Computer Science*, pages 376–387, Princeton, NJ, USA, August 2002. Springer.
29. Tamara G. Kolda. Limited-memory matrix methods with applications. Technical Report CS-TR-3806, University of Maryland, 1997.
30. Tamara G. Kolda and Dianne P. O'Leary. A semidiscrete matrix decomposition for latent semantic indexing information retrieval. *ACM Transactions on Information Systems*, 16(4):322–346, 1998.
31. F. W. Lancaster. *Information Retrieval Systems: Characteristics, Testing and Evaluation*. Wiley, New York, 1968.
32. Christopher D. Manning and Hinrich Schütze. *Foundations of Natural Language Processing*. MIT Press, 1999.
33. Jack Minker, Gerald Wilson, and Barbara Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8:329–348, 1972.
34. Tomas Olsson. Information filtering with collaborative agents. Master's thesis, Department of Computer and Systems Sciences, Royal Institute of Technology, Sweden, 1998.
35. Phibot search engine. Homepage, 2002. `http://phibot.org`.
36. Yonggang Qiu and Hans-Peter Frei. Concept-based query expansion. In Robert Korfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, Pennsylvania, USA, June 1993. ACM Press, New York, NY, USA.
37. Vijay V. Raghavan and Hayri Sever. On the reuse of past optimal queries. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 344–350, Seattle, Washington, USA, July 1995. ACM Press, New York, NY, USA.
38. Stephen E. Robertson and Karen Sparck-Jones. Relevance weighting of search terms. In *Journal of the American Society for Information Science*, volume 27, pages 129–146, 1976.
39. Stephen E. Robertson, Stephen Walker, and Micheline Hancock-Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, 1998.

40. Gerard Salton. *The SMART retrieval system - experiments in automatic document processing*. Prentice Hall, Englewood Cliffs, New Jersey, 1971.
41. Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
42. Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science and Technology*, 41(4):288–297, 1990.
43. Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, 1983.
44. Hayri Sever. *Knowledge Structuring for Database Mining and Text Retrieval Using Past Optimal Queries*. PhD thesis, University of Louisiana, Lafayette, LA, May 1995.
45. Ftp directory at cornell university. Homepage, 1968–1988. `ftp://ftp.cs.cornell.edu/pub/smart`.
46. Karen Sparck-Jones and Roger M. Needham. Automatic term classification and retrieval. *Information Storage and Retrieval*, 4:91–100, 1968.
47. James Thornton. Collaborative Filtering Research Papers. Homepage of James Thornton, 2003. `http://jamesthornton.com/cf/`.
48. Lily F. Tian and Kwok-Wai Cheung. Learning user similarity and rating style for collaborative recommendation. In Fabrizio Sebastiani, editor, *Advances in Information Retrieval, 25th European Conference on IR Research, ECIR 2003*, volume 2633 of *Lecture Notes in Computer Science*, pages 135–145, Pisa, Italy, April 2003. Springer.
49. Text REtrieval Conference (TREC). Homepage, 1992–2003. `http://trec.nist.gov`.
50. Cornelius J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
51. Cornelius J. van Rijsbergen. A non classical logic for information retrieval. *The Computer Journal*, 29(6):481–485, 1986.
52. Cornelius J. van Rijsbergen. Towards an information logic. In N. J. Belkin and C. J. van Rijsbergen, editors, *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 77–86, Cambridge, Massachusetts, USA, June 1989. ACM Press, New York, NY, USA.
53. John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, New Jersey, 1944.
54. Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*, 2002.
55. Ellen M. Voorhees and Donna Harman. Overview of the eighth text retrieval conference (TREC-8). In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, 2000.
56. 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.
57. Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Clustering user queries of a search engine. In *Proceedings of the 10th International World Wide Web Conference*, pages 162–168, Hong Kong, May 2001.
58. Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1):59–81, January 2002.
59. Ryen W. White, Ian Ruthven, and Joemon M. Jose. The use of implicit evidence for relevance feedback in web retrieval. In Fabio Crestani, M. Girolami, and Cornelis J.

van Rijsbergen, editors, *Advances in Information Retrieval, 24th BCS-IRSG European Colloquium on IR Research, ECIR 2002, Proceedings*, volume 2291 of *Lecture Notes in Computer Science*, pages 93–109, Glasgow, UK, March 2002. Springer.

60. Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In Hans-Peter Frei, Donna Harman, Peter Schabie, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland, August 1996. ACM Press, New York, NY, USA.
61. Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.
62. Yiming Yang and Xin Liu. A re-examination of text categorization methods. In Fredric Gey, Marti Hearst, and Richard Tong, editors, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, Berkeley, California, USA, August 1999. ACM Press, New York, NY, USA.