1 Towards Collaborative Information Retrieval: Three Approaches

Armin Hust, Stefan Klink, Markus Junker, and Andreas Dengel

German Research Center for Artificial Intelligence (DFKI GmbH),
P.O. Box 2080, 67608 Kaiserslautern, Germany
{armin.hust, stefan.klink, markus.junker, andreas.dengel}@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 unintrusively 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 three initial approaches, called QSD, QLD, and TCL as well as combinations of these approaches with pseudo relevance feedback. The approaches are evaluated experimentally on standard Information Retrieval 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.

1.1 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 Evaluation Conference series (TREC) [10]. In 1998, 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. At DFKI, 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 1.1 illustrates the general scenario of CIR. A document 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, he 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.

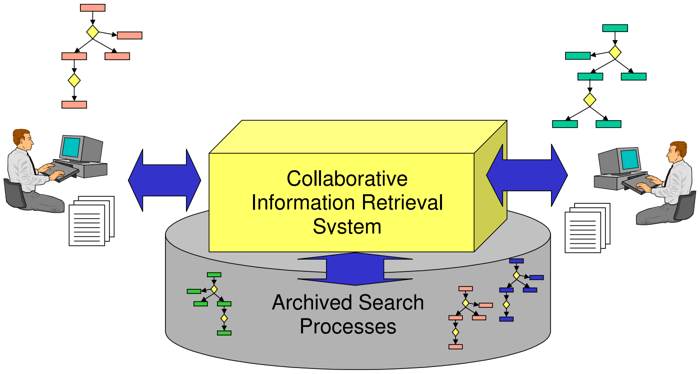

Fig. 1.1. Scenario of Collaborative Information Retrieval.

In this paper we focus on the investigation of a simple 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.

In section 1.2 we first introduce the well-known vector space model and the formal description of the task we are tackling. The first two approaches described in section 1.3 both rely on measuring the similarity of a new query to former ones for which relevant documents are known. In section 1.4 we propose an approach which extends queries by extending the query terms individually. In section 1.5 we introduce pseudo relevance feedback and its combination with our three approaches. Section 1.6 describes the experimental setup using standard IR test collections and section 1.7 presents the results and compares them with the vector space model and the pseudo relevance feedback model. In section 1.8 we briefly review our new methods.

1.2 Basics and Terminology

In this section we briefly recall the vector space model for Information Retrieval on which all of our approaches rely (section 1.2.1). We then formalize the CIR scenario we are focussing on in this paper (section 1.2.2).

1.2.1 The Vector Space Model

The basic retrieval model we rely on is the vector space model (VSM)[1], [6]. In this model documents as well as queries are represented by vectors. The relevance of a document with respect to a query is mapped to a similarity function between the query vector and all document vectors. More formally, documents as well as queries are represented by vectors $(w_1, w_2, ..., w_n)$. Each position i in the vectors corresponds to a specific term w_i in the collection. The value w_i indicates the weighted presence or absence of the respective term in the document or query. For weighting we rely on a variant of the standard tf-idf weighting schema [1]. In this weighting schema a word is weighted higher in a document/query if it occurs more often in the document/query. It is also weighted higher if the word is rare in the document collection. The similarity sim between a given query q and a document d is computed by

$$sim(d, q) = \frac{d \cdot q}{\parallel d \parallel \cdot \parallel q \parallel} \tag{1.1}$$

where $\parallel \cdot \parallel$ is the Euclidean norm of a vector. For retrieving all documents to a given query, all documents D of the underlying collection are ranked according to their similarity to the query and the top-ranked documents are given to the user.

1.2.2 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 $Q = \{q_1, \ldots, q_m\}$ available. For each $q \in Q$ the set of corresponding relevant documents is known and denoted by R_q. The goal now is to find all relevant documents for a new query $q^\star$ based on the old queries Q and their relevant documents $R_q, q \in Q$ (in general $q^\star \notin Q$). This is done by expanding the query $q^\star$ to a new query $q^\star_{\text{exp}}$ 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

$$\begin{aligned} f_{\text{exp}} : \mathbf{Q} \times 2^{(\mathbf{Q} \times (2^{\mathbf{D}}))} &\rightarrow \mathbf{Q} \\ (q^\star, \{(q_1, R_{q_1}), (q_2, R_{q_2}), \ldots, (q_m, R_{q_m})\}) &\mapsto q^\star_{\text{exp}} \end{aligned} \tag{1.2}$$

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

1.3 Query similarity-based approaches

The first two approaches presented in this section both rely on measuring the similarity between a new query and the old queries for which the answer documents are known.

1.3.1 QSD

The simple QSD approach (abbreviating "Query Similarity and relevant Documents") selects those former queries which are most similar to the new one. The relevant documents to each of these queries are represented by averaged document vectors. The expanded query is obtained by adding the weighted averaged document vectors to the original query.

The formal description is given here. The similarity $sim(q^\star, q)$ between the new query $q^\star$ and a query q is measured by the cosine of the angle between these two M dimensional vectors (cmp. equation 1.1):

$$sim(q^\star, q) = \frac{q^\star \cdot q}{\|q^\star\| \cdot \|q\|} \tag{1.3}$$

The set

$$Q_{q^\star} = \{q \in Q \mid sim(q, q^\star) \geq \vartheta\} \tag{1.4}$$

denotes all known queries which have a similarity to $q^\star$ greater than or equal to ϑ. The document vector r_q averages all documents which are relevant for query q:

$$r_q = \frac{\sum_{d \in R_q} d}{\| \sum_{d \in R_q} d\|} \tag{1.5}$$

Using $Q_{q^\star}$ and r_q, the expanded query vector $q^\star_{\mathrm{exp}}$ is computed by:

$$q^\star_{\mathrm{exp}} = q^\star + \sum_{q \in Q_{q^\star}} sim(q^\star, q) \cdot r_q \tag{1.6}$$

Thus, $q^\star_{\mathrm{exp}}$ is the expansion by representatives of the answer documents to the queries most similar to $q^\star$. Note that the QSD method has the parameter ϑ.

1.3.2 QLD

In general, a new query to be expanded by QSD consists of multiple words. The expansion depends on those former queries which have a high similarity to the new query. It can be that the high similarity of these most similar queries completely relies on just a subset of the terms in the original query. Thus, other terms of the new query will not be taken into account for the expansion. The QLD (Query Linear combination and relevant Documents) approach does not suffer from this potential drawback. It approximates the new query as a linear combination of similar former queries. The new query is then expanded by the answer documents to those queries which constitute the linear combination.

More formally, in the QLD approach a new query $q^\star$ is to be represented as a linear combination of the most similar former queries $Q_{q^\star} = \{q_1, q_2, \ldots, q_{m^\star}\}$:

$$q^\star = \sum_{i=1}^{m^\star} \lambda_i q_i \tag{1.7}$$

with λ_i being linear coefficients. In most cases we cannot represent the new query $q^\star$ exactly as a linear combination of the old queries $Q_{q^\star}$, i.e. equation 1.7 will have no solution. In order to solve this problem we write 1.7 as

$$q^\star = (q_1, q_2, \ldots, q_{m^\star})\lambda \tag{1.8}$$

where $(q_1, q_2, \ldots, q_{m^\star})$ is a matrix with n rows and $m^\star$ columns and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{m^\star})$ is a column vector of dimension $m^\star$. In this situation we have to find a vector $\hat{\lambda}$ which provides the closest fit to the equation in some sense. Our approach is to minimize the Euclidean norm of the vector $(q_1, q_2, \ldots, q_{m^\star})\lambda - q^\star$, i.e we solve

$$\hat{\lambda} = \text{argmin}_\lambda \, \|(q_1, q_2, \ldots, q_{m^\star})\lambda - q\| \tag{1.9}$$

with $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{m^\star})$ being the least squares solution for the equation system.

For the expansion of the new query $q^\star$ only those queries $q_i \in Q_{q^\star}$ are taken into account which significantly contribute to the representation of $q^\star$, i.e., the corresponding $|\hat{\lambda}_i|$ must exceed some threshold $\vartheta_{\hat{\lambda}}$:

$$\tilde{\lambda}_i = \begin{cases} \hat{\lambda}_i \text{ if } |\hat{\lambda}_i| \geq \vartheta_{\hat{\lambda}} \\ 0 \text{ if } |\hat{\lambda}_i| < \vartheta_{\hat{\lambda}} \end{cases} \tag{1.10}$$

The expanded query $q^\star_{\text{exp}}$ is obtained by adding the weighted representatives of the answer documents of those queries whose linear combination approximates the original query $q^\star$ best:

$$q^\star_{\text{exp}} = q^\star + \sum_{q \in Q_{q^\star}} \tilde{\lambda}_q r_q \tag{1.11}$$

In addition to the parameter ϑ as in the QSD approach, the QLD approach has the parameter $\vartheta_{\hat{\lambda}}$.

1.4 Query Term-based Approach

The two approaches QSD and QLD rely on measuring the similarity of a new query and old queries with known answer documents. In this section we present an approach which expands a query by expanding each query term individually. The TCL (Term Concept Learning) approach learns concept vectors for all terms that occur in former queries. These concepts are represented by concept vectors. They are computed by the documents relevant to all former queries containing the respective term. For the expansion of a new query, for each term in the query the corresponding concept vector is added. More formally, the TCL approach uses the sets $Q_i \subset Q$ of all former queries which contain the term with the index i:

$$Q_i = \{(w_1, ..., w_i, ..., w_n) \in Q \mid w_i \neq 0\} \tag{1.12}$$

The document set C_i denotes all documents relevant for at least one query in Q_i, i.e., all documents which are relevant to at least one query containing the term with index i:

$$C_i = \bigcup_{q \in Q_i} R_q \tag{1.13}$$

A representative concept vector for all documents contained in C_i is built by

$$c_i = \sum_{d \in C_i} d \tag{1.14}$$

The expansion $q^\star_{\text{exp}}$ of the new query $q^\star = (w^\star_1, \ldots, w^\star_m)$ is now given by

$$q^\star_{\text{exp}} = q^\star + \sum_{w^\star_i \neq 0} c_i \tag{1.15}$$

i.e., to all terms occurring in $q^\star$ the corresponding concept vector c_i is added. In contrast to QSD and QLD, the TCL approach has no parameters.

1.5 Pseudo Relevance Feedback

Pseudo Relevance Feedback (PRF) is a well-known technique in the VSM. It can improve the effectiveness of the original VSM approach as described in section 1.2. We use it in two different ways. First, it is used as a stand-alone retrieval technique (as described in section 1.5.1) for experimental comparisons. Secondly, we combine it with our query expansion approaches QSD, QLD, and TCL to new hybrid approaches. The combination and its motivation is described in section 1.5.2.

1.5.1 Stand-alone technique

Pseudo relevance feedback (PRF) is a well-known query expansion technique [9,2]. In contrast to our approaches, it relies purely on the document collection and does not use former queries for expanding a new query. PRF enriches a new query $q^\star$ by the terms of the top-ranked documents with respect to $q^\star$.

We are using a variation of PRF as described in [5]. Let $D_{q^\star}$ be the set of document vectors given by

$$D_{q^\star} = \left\{ d \in D \middle| \frac{\mathrm{sim}(d, q^\star)}{\max_{d' \in D}\{sim(d', q^\star)\}} \geq \theta \right\} \tag{1.16}$$

where $q^\star$ is the original query and θ is a similarity threshold. The expanded query vector $q^\star_{\mathrm{exp}}$ is obtained by

$$q^\star_{\mathrm{exp}} = q^\star + \alpha \frac{p}{\|p\|}, \text{ with } p = \sum_{d \in D_{q^\star}} d \tag{1.17}$$

Note that the PRF approach has two parameters, θ and α.

1.5.2 Combination with new Approaches

We expect our approaches to be particulary useful if there is some overlap in the user queries. In contrast, PRF cannot exploit previous queries but just uses the document collection to be retrieved for query expansion. It is obvious that a combination of both approaches might be fruitful.

Our combination of PRF with QSD, QLD, and TCL is straight-forward. First, QSD/QLD/TCL is used to expand the query $q^\star$. The expanded query $q^\star_{exp}$ is then expanded again by PRF as described above. We denote the combined approaches as PRF(QSD), PRF(QLD), and PRF(TCL).

1.6 Experimental Setup

We use standard Information Retrieval test collections for our experiments as provided by [8] and [10]. These collections were originally made for evaluating the effectiveness of different IR systems in a wide range of (artificial) queries. The collections offer the advantage that for all queries the correct answer documents are known. On the other hand the queries are very likely not typical for real world document retrieval systems. In particular, as opposed to real world systems, we expect little overlap in query topics. This may harm the performance of our approaches.

In the experiments we used three collections of the SMART system: "CACM" (titles and abstracts from the journal 'Communications of the "ACM"'), "CISI" (texts from the Institute of Scientific Information) and CRAN ("Cranfield" collection, abstracts of Aerodynamics). The CR collection (congressional reports) of the TREC conference provides three different lengths for the queries from which we generated three test sets: "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). Also from TREC we used the collection "FR" (federal register entries) and a modified version of the "AP90" collection. The questions in

AP90 have one particularity: they were built by first defining 500 individual queries and then adding 193 reformulations to some of the original queries [11]. The modified version, which we call ”AP90*” was generated by taking only the questions 201-893 which had at least one answer document in AP90 and only those documents of ”AP90” which were relevant to at least one question. We thus resulted in 353 queries and 723 documents.

Table 1.1 lists some statistics about the collections we used after stemming and stop word elimination has been carried out.

Table 1.1. Statistics about test collections

	CACM	CISI	CRAN	CR-title	CR-desc	CR-narr	FR	AP90*
size(MB)	1.2	1.4	1.4	93	93	93	69	3.7
number of docs	3204	1460	1400	27922	27922	27922	19860	723
number of terms	3029	5755	2882	45717	45717	45717	50866	17502
mean doc length	18.4 (short)	38.2 (med)	49.8 (med)	188.2 (long)	188.2 (long)	188.2 (long)	189.7 (long)	201.8 (long)
number of queries	52	112	225	34	34	34	112	353
mean query length	9.3 (med)	23.3 (long)	8.5 (med)	2.9 (short)	7.2 (med)	22.8 (long)	9.2 (med)	3.2 (short)
mean num. of rel. docs per query	15.3 (med)	27.8 (high)	8.2 (med)	24.8 (high)	24.8 (high)	24.8 (high)	8.4 (med)	2.8 (low)

The evaluation of the new learning methods (QSD, QLD, TCL, PRF(QSD), PRF(QLD), and PRF(TCL)) was done using a leave-one-out technique: From the set of queries $Q = \{q_1, \dots q_n\}$ in a collection, we selected each $q_k \in Q$ and expanded it based on the queries $Q \setminus q_k$. Effectiveness was measured using macro-averaged recall/precision and averaged precision. Details on these measures can be found in [1]. In order to identify significant differences among methods we use the macro-averaged t-test (see, e.g., [12]) on the averaged precision.

1.7 Results

Some of the methods we compared contain parameters. For all methods with parameters we first searched for the parameter setting maximizing the average precision for a collection. The following parameters were tested:

- for PRF: $\alpha \in \{0, 0.1, \dots, 2.0\}$ and $\theta \in \{0, 0.05, \dots, 1.0\}$
- for QSD: $\vartheta \in \{0, 0.01, \dots, 1.00\}$
- for QLD: $\vartheta \in \{0, 0.01, \dots, 1.00\}$ and $\vartheta_{\hat\lambda} \in \{0, 0.01, \dots, max\{\hat{\lambda_i}\}\}$

Table 1.2 shows the respective parameter values. Please note that finding the best parameter settings for QSD and QLD has to be taken with some

Table 1.2. Optimal Parameter Settings

		CACM	CISI	CRAN	CR-title	CR-desc	CR-narr	FR	AP90*
PRF	α	1.7	0.7	1.3	0.6	0.5	0.4	0.6	0.2
	θ	0.35	0.7	0.9	0.75	0.85	0.95	0.55	0.75
QSD	ϑ	0.24	0.41	0.49	0.42	0.44	0.33	0.36	0.68
QLD	ϑ	0.22	0.25	0.37	0.41	0.36	0.17	0.37	0.68
	$\vartheta_{\hat{\lambda}}$	0.16	0.23	0.41	0.43	0.37	0.20	0.13	0.48
TCL	-	-	-	-	-	-	-	-	-
PRF(QSD)	ϑ	same as for QSD							
	α	0.6	0.3	0.7	0.6	0.5	0.4	0.4	0.1
	θ	0.65	0.7	0.95	0.75	0.85	0.95	0.0	0.9
PRF(QLD)	$\vartheta, \vartheta_{\hat{\lambda}}$	same as for QLD							
	α	0.8	0.2	0.6	0.4	0.5	0.4	0.4	0.1
	θ	0.7	0.85	0.95	0.75	0.85	0.95	0.0	0.9
PRF(TCL)	α	1.9	0.2	0.4	0.6	0.3	0.0	0.3	0.1
	θ	0.70	0.85	0.90	0.85	0.95	0.00	0.45	0.95

caution. With this optimization we can only show that these methods have the potential to obtain a specific performance.

Table 1.3 shows the average precision obtained by using the best parameter values for different methods. The best value of the average precision is indicated in bold face. Figures 1.2 to 1.4 show the corresponding recall/precision graphs. The results of the significance tests are shown in table 1.4. The entries have the following meanings with respect to the two methods X and Y:

- $\gg$ ($>$) indicates that X performs better than Y at the significance level of $\alpha = 0.01$ ($\alpha = 0.05$)
- $\ll$ ($<$) indicates that X performs worse than Y at the significance level of $\alpha = 0.01$ ($\alpha = 0.05$).
- $\circ$ indicates that no statement on differences can be made on the significance levels.

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 QSD and QLD approaches generally also outperform the VSM with just one exception being not significant (QSD on "CR-desc"). The comparison between TCL and VSM gives no clear picture.
- The combined methods generally outperform VSM and PRF as expected. Two exceptions are PRF(QSD) as compared to PRF on "CR-desc" and PRF(TCL) as compared to PRF on "CR-narr". In these two settings the combined methods are significantly worse than PRF.

Table 1.3. Average precision obtained in different methods

	CACM	CISI	CRAN	CR-title	CR-desc	CR-narr	FR	AP90*
VSM	13.0	12.0	38.4	13.5	17.5	17.3	8.5	74.3
PRF	19.9	12.9	43.5	16.9	20.4	19.2	11.3	75.5
QSD	23.7	14.2	42.8	15.2	17.2	17.3	10.9	81.1
QLD	22.7	17.1	43.6	16.4	17.5	17.5	10.8	81.1
TCL	28.2	10.0	34.2	16.0	16.7	8.2	14.0	62.9
PRF(QSD)	25.7	14.5	45.1	19.5	19.1	17.7	16.3	**81.4**
PRF(QLD)	27.3	**17.3**	**45.3**	**20.4**	19.2	**18.4**	16.1	**81.4**
PRF(TCL)	**30.4**	12.7	42.6	18.0	**20.6**	17.3	**19.9**	77.3

Table 1.4. Comparison of approaches using significance test

X	Y	CACM	CISI	CRAN	CR-title	CR-desc	CR-narr	FR	AP90*
PRF	VSM	$\gg$	$\gg$	$\gg$	$>$	$\gg$	$>$	$>$	$>$
QSD	VSM	$\gg$	$>$	$\gg$	$\circ$	$\circ$	$\circ$	$>$	$\gg$
QLD	VSM	$\gg$	$\gg$	$\gg$	$\circ$	$\circ$	$\circ$	$>$	$\gg$
TCL	VSM	$\gg$	$\ll$	$\ll$	$\gg$	$\circ$	$\ll$	$\gg$	$\ll$
QSD	PRF	$\circ$	$\circ$	$\circ$	$\circ$	$\ll$	$<$	$\circ$	$\gg$
QLD	PRF	$\circ$	$\gg$	$\circ$	$\circ$	$\ll$	$\circ$	$\circ$	$\gg$
TCL	PRF	$\gg$	$\ll$	$\ll$	$\circ$	$\circ$	$\ll$	$\circ$	$\ll$
PRF(QSD)	PRF	$\circ$	$\circ$	$\circ$	$\circ$	$<$	$\circ$	$>$	$\gg$
PRF(QLD)	PRF	$>$	$\gg$	$\circ$	$\circ$	$\circ$	$\circ$	$>$	$\gg$
PRF(TCL)	PRF	$\gg$	$\circ$	$\circ$	$\circ$	$\circ$	$<$	$\gg$	$>$

- As expected QLD slightly outperforms QSD in most cases. This is also true for the combined version PRF(QLD) which mostly outperforms PRF(QSD).
- Generally, the best results are obtained with the approaches combining QLD, QSD, and TCL with PRF.

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 extent of overlap)
- the number of relevant documents for queries with some ”overlap in content”

1.8 Summary

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 three approaches QSD, QLD, and TCL 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 QLD, an extension of the QSD approach, performs slightly better. The differences between QLD and TCL give no clear picture. We are still lacking on an explanation when and why TCL or QLD performs better or worse.

As one of our next steps for the QSD/QLD approaches we want to learn the similarity measure between queries based on training examples. In future, for the TCL approach, we do not always want to extend query terms by learned concepts but only if the learned concepts are reliable in some way. Two more topics that we will work on in the future are the removal of the explicit parameters in the QSD and QLD approach as well as shifting towards real world retrieval systems. For the latter we are cooperating with a search engine provider.

Acknowledgements

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

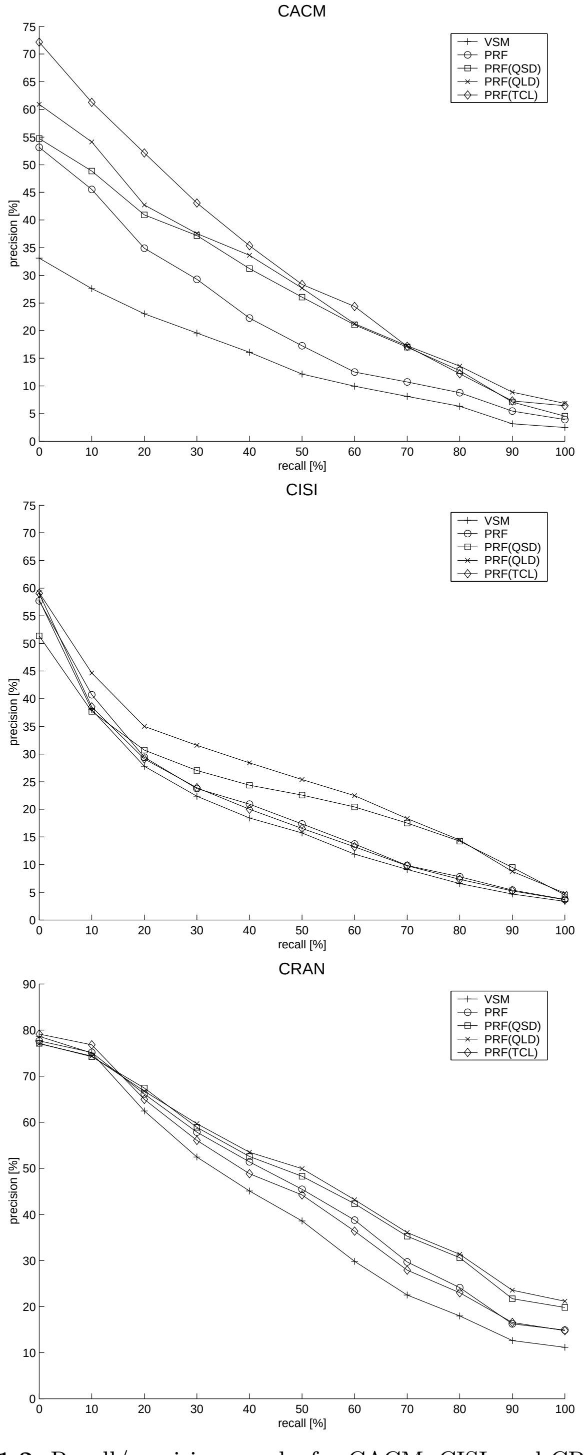

Fig. 1.2. Recall/precision graphs for CACM, CISI, and CRAN.

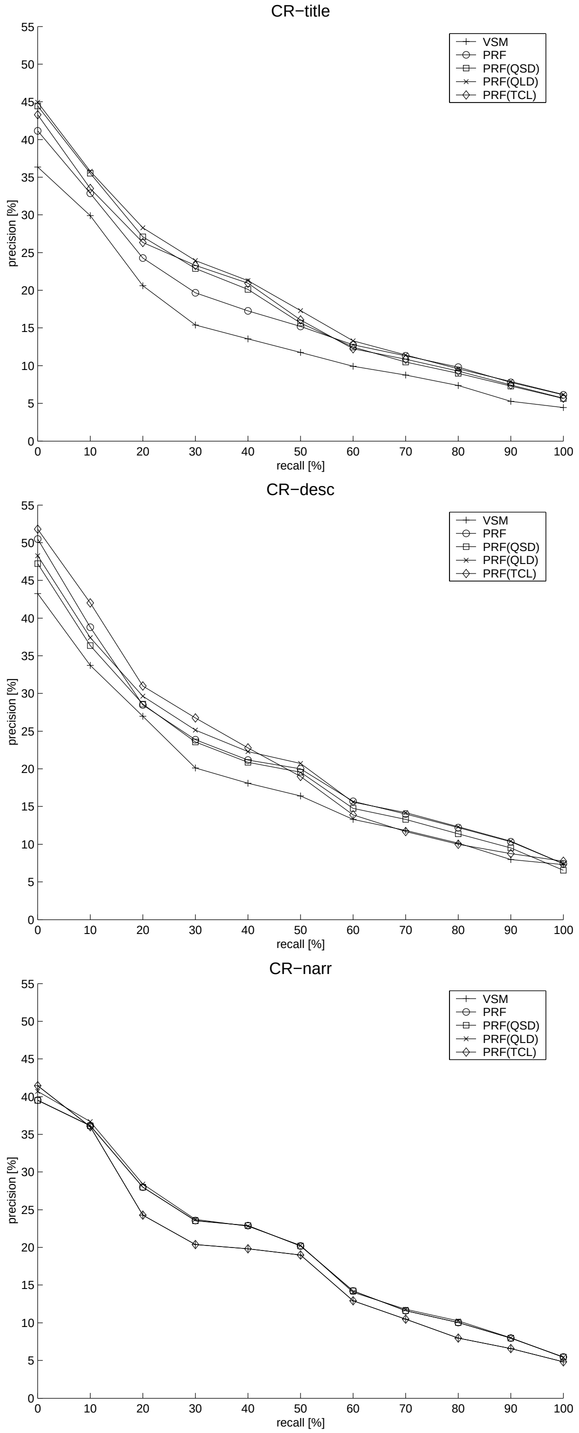

Fig. 1.3. Recall/precision graphs for CR.

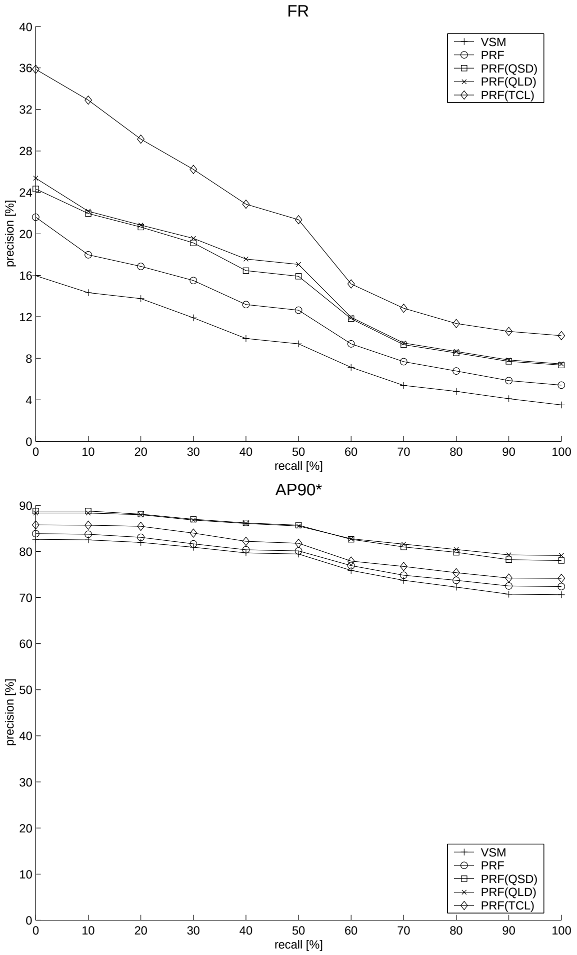

Fig. 1.4. Recall/precision graphs for FR and AP90*.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley Publishing Company, 1999.
2. C. Buckley, G. Salton, and J. Allen. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300, Dublin, Ireland, 1994.
3. H. Cui, J. Wen, J. Nie, and W. Ma. Probabilistic Query Expansion Using Query Logs. In *Proceedings of the Eleventh International World Wide Web Conference (WWW2002)*, Honolulu, Hawaii, USA, 2002.
4. D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, Pittsburgh, PA, USA, 1993.
5. K. Kise, M. Junker, A. Dengel, and K. 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, Washington, USA, 2001.
6. C. Manning and H. Schütze. *Foundations of Natural Language Processing.* MIT Press, 1999.
7. Y. Qiu and H.-P. Frei. Concept-based query expansion. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, PA, USA, 1993.
8. `ftp://ftp.cs.cornell.edu/pub/smart`.
9. J.J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System - Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Englewood Cliffs, N.J., 1971.
10. `http://trec.nist.gov`.
11. E. M. Voorhees and D. Harman. Overview of the ninth text retrieval conference (TREC-9). In *Proceedings of the Ninth Text Retrieval Conference*, pages 1–13, Gaithersburg, Maryland, USA, 2000.
12. Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, University of California, Berkeley, USA, 1999.