An Information-Theoretic Approach to Automatic Query Expansion

CLAUDIO CARPINETO
Fondazione Ugo Bordoni
RENATO DE MORI
University of Avignon
GIOVANNI ROMANO
Fondazione Ugo Bordoni
and
BRIGITTE BIGI
University of Avignon

Techniques for automatic query expansion from top retrieved documents have shown promise for improving retrieval effectiveness on large collections; however, they often rely on an empirical ground, and there is a shortage of cross-system comparisons. Using ideas from Information Theory, we present a computationally simple and theoretically justified method for assigning scores to candidate expansion terms. Such scores are used to select and weight expansion terms within Rocchio's framework for query reweighting. We compare ranking with information-theoretic query expansion versus ranking with other query expansion techniques, showing that the former achieves better retrieval effectiveness on several performance measures. We also discuss the effect on retrieval effectiveness of the main parameters involved in automatic query expansion, such as data sparseness, query difficulty, number of selected documents, and number of selected terms, pointing out interesting relationships.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; *Relevance feedback*; *Query formulation*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms: Algorithms, Design, Experimentation, Theory

Additional Key Words and Phrases: Information retrieval, automatic query expansion, pseudorelevance feedback, information theory

Authors' addresses: C. Carpineto, Fondazione Ugo Bordoni, Via B. Castiglione, 59, Rome, 00142, Italy; email: carpinet@fub.it; R. De Mori, Laboratoire d'Informatique, University of Avignon, BP 1228, Avignon, Cedex 9, 84911, France; email: renato.demori@lia.univ-avignon.fr; G. Romano, Fondazione Ugo Bordoni, Via B. Castiglione, 59, Rome, 00142, Italy; email: romano@fub.it; B. Bigi, Laboratoire d'Informatique, University of Avignon, BP 1228, Avignon, Cedex 9, 84911, France; email: brigitte.bigi@lia.univ-avignon.fr.

1. INTRODUCTION

Current information retrieval systems are limited by many factors reflecting the difficulty to satisfy user requirements expressed by short queries for identifying documents which often are long and have complex structures. Moreover, the user may express the conceptual content of the required information with query terms which do not match the terms appearing in the relevant documents.

This *vocabulary problem*, discussed, for example, by Furnas et al. [1987], is more severe when the user queries are short or when the database to be searched is large, as in the case of Web-based retrieval.

The importance of the vocabulary problem for short queries has recently motivated a research effort on methods for augmenting query contexts given an underlying retrieval model. A popular solution for multiple-query searches is to use manually [Brajnik et al. 1996] or automatically built [Grefenstette 1994; Cooper and Byrd 1997; Carpineto and Romano 1998] thesauri for guiding the user to reformulate queries in an interactive process. A thesaurus can also be used for automatic query expansion on single-query searches, but doubts have been expressed about the benefits for the retrieval effectiveness in this case [Voorhees 1994].

A different, more elaborate, approach to automatic query expansion is to exploit the content relationships between the documents in a collection. Approaches followed for this purpose rely on Boolean term decomposition [Wong et al. 1987], statistical factor analysis [Deerwester et al. 1990], and formal concept analysis [Carpineto and Romano 2000a]. Systems using these methods explicitly handle word mismatch. Unfortunately, they are computationally expensive and have not shown so far tangible advantages with respect to systems based on the best match of the original query.

A third well-known, and perhaps more simple and effective, approach to alleviate the vocabulary problem is the automatic extraction of useful terms from the top retrieved documents. This is sometimes referred to as *retrieval feedback* or *pseudorelevance feedback*. Past experiments reported in the literature, reviewed by Salton and Buckley [1988] and Harman [1992], have shown that the application of this technique often resulted in a loss in precision higher than the corresponding gain in recall. Nevertheless, recently, some success is reported in applications to large-scale collections (e.g., Buckley et al. [1995], Xu and Croft [1996], Vélez et al. [1997], and Xu and Croft [2000]). In fact, almost all the groups involved in the TREC evaluations have reported improved results by expanding queries using information from the top retrieved documents [Voorhees and Harman 1998; 1999]. The growing interest for this technique makes evident the need to develop well-founded methodologies for ranking and weighting expansion terms and to perform experimental studies for evaluating and contrasting merits and drawbacks of currently used methods for query expansion, rather than just making comparisons with the use of nonexpanded queries.

This paper introduces a new term-scoring function that is based on the differences between the distribution of terms in (pseudo) relevant documents

and the distribution of terms in all documents. A computationally simple and sound method is proposed to assign scores to candidate expansion terms. The method is based on the Kullback-Leibler divergence measure well known in Information Theory [Losee 1990; Cover and Thomas 1991].

The second contribution of this paper is a set of experimental results for assessing the effectiveness of automatic query expansion using the scores of the proposed method compared to the use of other distributional functions (e.g., based on chi-squared). The purpose of a first experiment is to evaluate whether distributional functions can effectively complement more conventional reweighting methods such as Rocchio's formula by selecting the terms to be used for query expansion. The results are, in general, negative. When the same distributional methods are used both to select and weight expansion terms, another experiment shows that the performance of most methods, while occasionally improving over the former experiment, remains inferior to Rocchio's for several data points. However, experimental evidence shows that the information-theoretic method provides the best retrieval effectiveness across different data sets and for almost all evaluation measures, with considerable performance improvement over all methods tested in the experiments, including Rocchio.

A third major contribution of the paper is the study of retrieval performance for different levels of data sparseness, query difficulty, number of selected documents, and number of selected terms when query expansion is performed with the proposed information-theoretic method. In particular, it is shown that it is possible to extract expansion terms with negative weights from the set of pseudorelevant documents, although this may at most marginally improve the system's retrieval effectiveness for the first retrieved documents. Furthermore, it appears that performance does not necessarily improve with easy queries.

The rest of the paper is organized as follows. Section 2 characterizes the main steps of the automatic query expansion process and discusses the rationale of using term-ranking methods based on distribution analysis. Section 3 introduces a term-scoring method based on the Kullback-Leibler distance to rank and weight expansion terms. Section 4 is devoted to the evaluation of retrieval effectiveness. The general experimental setting is described, including the baseline ranking system and the other distributional term-scoring functions tested in the experiments. The use of such functions is then evaluated within Rocchio's classical reweighting scheme in two different scenarios, namely, by only allowing ranking of expansion terms, or by allowing term ranking followed by term weighting. Section 5 discusses the role played by the main parameters involved in automatic query expansion in determining the overall retrieval effectiveness. Section 6 offers some conclusions and directions for future work.

2. APPROACHES TO AUTOMATIC QUERY EXPANSION

Attar and Fraenkel [1977] and Croft and Harper [1979] made a conjecture that, in the absence of any other relevance judgment, the top few documents

retrieved on the basis of an initial query are relevant. Using the content of these documents, a new, more detailed query is obtained by a three-step process: search for expansion terms, expansion term ranking, and construction of an expanded query with the new terms.

The typical source of elements for expanding a given query is the set of all the terms in the first documents retrieved in response to the original query from the collection at hand. More sophisticated methods for selecting candidate terms for query expansion have been proposed in recent years. They use information such as passages [Xu and Croft 1996; Hawking et al. 1998; Xu and Croft 2000], the result of past similar queries [Fitzpatrick and Dent 1997], or the documents retrieved in a much larger collection than the target one [Singhal et al. 1999].

In most systems, the following improved version [Salton and Buckley 1990][1] of the original Rocchio's formula [Rocchio 1971] is the starting point for updating the term weights:

$$Q_{new} = \alpha \cdot Q_{orig} + \frac{\beta}{|R|} \sum_{r \in R} r - \frac{\gamma}{|R'|} \sum_{r' \in R'} r' \tag{1}$$

Q_{new} is a weighted term vector for the expanded query; Q_{orig} is a weighted term vector for the original unexpanded query; R and R' are respectively the sets of relevant and nonrelevant documents; r and r' are term weighting vectors extracted from R and R', respectively. The weights in each vector are computed by a weighting scheme applied to the whole collection.

If term ranking and automatic query expansion rely on a set of highly scored retrieved documents, then the (1) reduces to

$$Q_{new} = \alpha \cdot Q_{orig} + \frac{\beta}{|R|} \sum_{r \in R} r \tag{2}$$

where R is the set of top retrieved documents which are assumed to be relevant. The weights obtained with Eq. (2) are typically used to both rank and reweight the query terms [Srinivasan 1996; Mitra et al. 1998; Singhal et al. 1999].

This approach is simple and computationally efficient, but it has the disadvantage that each term weight reflects more the usefulness of that term with respect to the entire collection rather than its importance with respect to the user query.

A conceptually different approach for assessing the appropriateness of a term is based on distribution analysis. In order to discriminate between

[1]Some other versions of Rocchio's formula have recently been proposed. These versions as well as the one used here have led to improvements with respect to the use of the basic formula on tasks involving proper relevance feedback (Buckley and Salton, 1995; Schapire et al. 1998). The experiments described in this paper have been performed using only the improved version described here.

good expansion terms and poor expansion terms it may be more convenient to compare occurrence in relevant documents with occurrence in all documents, for each given query. In other words, one may assume that the difference between the distribution of terms in a set of relevant documents and the distribution of the same terms in the overall document collection is an index of semantic difference. In particular, it is expected that the frequency of appropriate terms will be higher in relevant documents than in the whole collection, while other terms will occur with the same frequency (randomly) in both document sets. This distributional view of discriminating relevant from nonrelevant documents has been extensively discussed, among others, by van Rijsbergen [1977] and Harper and van Rijsbergen [1978].

An early application of this concept can be found in the system developed by Doszkocs [1978], where a comparative statistical analysis of term occurrences—via a chi-squared variant—was made to suggest potentially relevant terms for interactive query expansion. In CUPID [Porter 1982], the first probabilistic search engine, the simple differences in term distribution were used as a term selection statistic. The use of the differences in term distribution to select the terms to be included in the expanded query was then theoretically analyzed by Robertson [1990]. He showed, that under certain strong assumptions, if index terms are to be weighted for retrieval with w_t, then they should be selected with RSV (Robertson Selection Value) = $w_t(p_t - q_t)$, where p_t and q_t are the probabilities that a relevant and a nonrelevant document, respectively, contain the term t.

Variants of the ranking scheme proposed by Robertson [1990] have subsequently been used in various systems with different weighting functions and different methods for estimating p_t and q_t [Buckley et al. 1995; Robertson et al. 1999; Hawking et al. 1998]. Moreover, Efthimiadis [1993] presented an evaluation of the ability to rank potentially relevant terms with several term-scoring functions based on distribution analysis, including BIM [Robertson and Spark Jones 1976], EMIM [van Rijsbergen et al. 1981], and RSV. This evaluation was confined to an interactive retrieval scenario and did not cover automatic query expansion.

3. USING RELATIVE ENTROPY TO EVALUATE EXPANSION TERMS

In the query expansion model proposed in this paper, queries and documents are represented by vectors of weighted terms belonging to a vocabulary V, and it is assumed that a distance is defined in the term vector space. The retrieval problem is that of inferring the set D^* of all the documents relevant to a given query from the vector Q_{orig}, representing the query.

Unfortunately, most of the coordinates of Q_{orig} are zero, because the query usually contains very few terms, while vectors in D^* have, in general, many more nonzero coordinates because the corresponding documents contain many more terms than the query. Thus, the set of vectors having

distance from Q_{orig} less than a given threshold may be a poor approximation of D^*. This is especially true if we assume, as with the cluster hypothesis [van Rijsbergen 1979; Hearst and Pedersen 1996], that the intraset distance between any pair of vectors in D^* is lower than the interset distance between any vector in D^* and any vector outside D^*, because in this case the distance between Q_{orig} and vectors in D* would be much higher than many interset distances.

Nevertheless, a better approximation to D^* could be found by considering the term vectors having limited distance from a suitable vector Q_{new}.

The model for query expansion proposed in the following is based on a new method for determining Q_{new}. The vector Q_{new} is built from an approximation of the unigram probability distribution (i.e., under assumptions of independence) of all the terms in V inferred from the vectors in D^*, which is unknown.

The reasonable starting point for building Q_{new} is the set of documents R retrieved from Q_{orig} by a classical method without query expansion. Let p_R be the unigram probability distribution of all terms t in V inferred from the vectors representing the documents in R. Let p_C be the unigram probability distribution of all terms t in V inferred from the vectors representing the documents of the entire collection C. The vector Q_{new} will have as elements the frequencies of those terms which mostly contribute to make p_R divergent with respect to p_C. A suitable divergence measure should be that which maximizes the *relative entropy* between the two probability distributions. This relative entropy is measured by the *Kullback-Leibler* divergence (*KLD*) measure defined in Information Theory [Losee 1990; Cover and Thomas 1991]:

$$KLD_{(p_R, p_C)} = \sum_t \left\{ p_R(t) \times \log \frac{p_R(t)}{p_C(t)} \right\} \tag{3}$$

KLD has also been used, for example, in natural language and speech processing applications based on statistical language modeling [Dagan et al. 1999], and in information retrieval, for topic identification [Bigi et al. 1997] , for choosing among distributed collections [Xu and Croft 1999], and for modeling term weighting as deviation from randomness [Amati and van Rijsbergen 2000]. A related measure, the expected mutual information, has been used to handle the parameter estimation problems raised by the Robertson-Sparck Jones formula [van Rijsbergen 1979], as well as for term selection [van Rijsbergen et al. 1981].

The terms to be considered for query refinement are those which mostly contribute to the divergence defined in expression (3). In other words, we associate to each term a score given by the corresponding summand in expression (3), then rank the terms according to their scores and choose a fixed number of best ranked terms.

It should be noted that, although the relative entropy is always nonnegative (it is zero if and only if $p_R = p_C$), the individual summands in expression (3) might be negative. This happens whenever the probability of occurrence of a term t in R is smaller than the corresponding one in the entire collection C. The effects of this phenomenon will be discussed later.

4. PERFORMANCE EVALUATION

4.1 Objective of the Experiments

The objective of the experiments reported in this section was to test the hypothesis that term-ranking methods based on distribution analysis, including the information-theoretic one, can be used to improve the retrieval effectiveness of Rocchio's automatic query expansion. Two experiments were executed. The scores produced by the distributional functions were first used to rank the expansion terms within Rocchio's formula. In a second experiment, the same scores were used not only to select but also to weight the selected terms.

4.2 Term-Ranking Functions

In addition to the KLD-based method, we tested four other term-ranking functions. The complete list of functions tested is the following (R indicates the pseudorelevant set, C the whole collection, and $w(t)$ is the weight of term t in the collection):

—Rocchio's weights:

$$\text{score}(t) = \sum_{k=1}^{r} w(t)_{Doc_k} \tag{4}$$

—Robertson Selection Value (RSV):[2]

$$\text{score}(t) = \sum_{k=1}^{r} w(t)_{Doc_k} \cdot p_R(t) \tag{5}$$

—CHI-squared (CHI2):

$$\text{score}(t) = [p_R(t) - p_C(t)]^2 / p_C(t)] \tag{6}$$

—Doszkocs' variant of CHI-squared (CHI1):

$$\text{score}(t) = [p_R(t) - p_C(t)] / p_C(t)] \tag{7}$$

—Kullback-Leibler distance (KLD):

$$\text{score}(t) = [p_R(t)] \cdot \log[p_R(t) / p_C(t)] \tag{8}$$

[2]We should emphasize that *RSV* was specifically intended as a selection function, not as a term weight for retrieval. For its computation, we assumed, as also done in Robertson et al. [1995], that the probability that a nonrelevant document contains the term t is negligible.

In the experiments described below, we considered as candidate expansion terms only the terms contained in the set of documents R; we will show in Section 5.1 how to include also the terms that are not contained in R. The size of R was set to 10, while the number of expansion terms considered for inclusion in the expanded query was set to 40. These choices are consistent with other experimental studies on the TREC collection; no additional parameter tuning was performed.

The estimation of probabilities in the above expressions is an important issue because it might affect performance results. To compute the probability of occurrence of a term t in X (whether the set of documents R or the whole collection C), the maximum likelihood estimate of $p(t)$ was used under the term distribution for X, i.e., the ratio between the raw frequency of t in X, treated as a long sequence of terms, and the number of term tokens in X:

$$p_x(t) = \frac{f_x(\mathrm{t})}{NT_{\mathrm{x}}} \tag{9}$$

Different estimates of probabilities than expression (9) were also tried, including the number of pseudorelevant documents that contain a term [Buckley et al. 1995]. This latter method was found to produce worse retrieval effectiveness for any term-scoring function but RSV. Thus, we chose the document-based probability only for RSV; in fact, this is also the recommended choice for RSV by Robertson et al. [1995].

4.3 Test Collections and Baseline Document Ranking System

The TREC-7 and TREC-8 collections were used for test. They consist of the same set of documents (i.e., TREC disks 4 and 5, containing approximately 2 gigabytes of data) and different query sets (topics 351–400 and topics 401–450, respectively). The full topic statement was considered, including "title," "description," and "narrative." TREC-7 topics were described with an average of 57.6 terms, while the average on TREC-8 topics was 51.8 terms.

The basic system used in all the experiments was developed in the context of our participation in TREC-8 [Carpineto and Romano 2000b]. The system performs word stopping and word stemming, using a very large *trie*-structured morphological lexicon for English [Karp et al. 1992]. Single keyword indexing was performed for both test collections.

In the first-pass ranking, the system used the following Okapi formula [Robertson et al. 1999] for computing a similarity measure between a query q and a document d:

$$sim(q, d) = \sum_{t \in q \wedge d} W_{d,t} \cdot W_{q,t} \tag{10}$$

with

$$W_{d,t} = \frac{(k_1 + 1) \cdot f_{d,t}}{k_1 \cdot \left[(1 - b) + b \cdot \frac{W_d}{avr\_W_d} \right] + f_{d,t}} \tag{11}$$

and

$$W_{q,t} = \frac{(k_3 + 1) \cdot f_{q,t}}{k_3 + f_{q,t}} \cdot \log \frac{N - f_t + 0.5}{f_t + 0.5} \tag{12}$$

where k_1, k_3, and b are constants which were set to 1.2, 1000, and 0.75 respectively. W_d is the length of document d expressed in words, and avr_W_d is the average document length in the entire collection. The value N is the total number of documents in the collection; f_t is the number of documents in which term t occurs; and $f_{x,t}$ is the frequency of term t in either document d or query q.

The whole system was implemented in Common Lisp and runs on a SUN-Ultra workstation. The time spent to index the whole collection (several hours) and to compute the document ranking for each query (several seconds) was relatively large because I/O procedures were not fully optimized. Nonetheless, the time necessary to perform just query expansion was negligible. As the collection frequencies were stored in the inverted file built from the set of documents, the computation of $p_C(t)$ was straightforward. In order to obtain a fast computation of $p_R(t)$, one pass through the first retrieved documents was performed. This makes information-theoretic query expansion applicable even in interactive applications, provided that it is used in conjunction with a more efficient baseline ranking system than ours (TREC systems with best response times take about one second per query).

4.4 Experiment 1: Using KLD within Rocchio to Rank Expansion Terms

The first goal of the experiments was to evaluate and compare the performance of the term-scoring functions introduced above for selecting expansion terms. As the overall retrieval effectiveness may depend on many factors, only the method used for selecting expansion terms was varied, while the other factors involved in the query expansion process were kept constant. Most important, in order to reweight the query after selection of expansion terms, we uniformly used Rocchio's formula reported in expression (2), with $\alpha = 1$, $\beta = 1$.[3] The weights of the expanded query were then used within expression (10) to compute the second-pass ranking.

For each query, the complete ranking system was executed for each of the five methods considered for term expansion. Tables I and II show the retrieval performance of each method on TREC-7 and TREC-8, respectively; they also show the performance improvement over a baseline in

[3]We tried also the pair $\alpha = 1$, $\beta = 2$, obtaining similar results.

Table I. Comparison of Retrieval Performance on TREC-7 Using Distributional Term Ranking

	Unexp.	Rocchio	RSV_R	$CHI\text{-}2_R$	$CHI\text{-}1_R$	KLD_R
RET&REL	2751	3009 +9.38%*	2999 +9.01%*	2982 +8.40%*	2926 +6.36%*	3019 +9.74%*
AV-PREC	0.2291	0.2625 +14.54%*	0.2610 +13.92%*	0.2569 +12.10%*	0.2493 +8.80%*	0.2629 +14.72%*
11-PT-PREC	0.2545	0.2806 +10.25%*	0.2792 +9.71%*	0.2779 +9.21%*	0.2719 +6.82%*	0.2829 +11.14%*
R-PREC	0.2711	0.2972 +9.62%*	0.2980 +9.93%*	0.2912 +7.43%*	0.2835 +4.56%*	0.2953 +8.94%*
PREC-AT-5	0.5480	0.5760 +5.11%	0.5800 +5.84%	0.5920 +8.03%*	0.5920 +8.03%*	0.5840 +6.57%
PREC-AT-10	0.5120	0.5240 +2.34%	0.5240 +2.34%	0.5280 +3.12%	0.5340 +4.30%	0.5380 +5.08%*

which queries were not expanded. Performance was measured with the TREC's standard evaluation measures:

`RET&REL`: number of relevant documents contained in the first thousand retrieved documents,

`AV-PREC`: average precision,

`PT-PREC`: 11-point precision,

`R-PREC`: R-precision,

`PREC-AT-5`: precision at five retrieved documents,

`PREC-AT-10`: precision at 10 retrieved documents.

The first measure is relative to all queries; the other measures are averaged over the query set. In Table I, the distributional methods have an R subscript to indicate that they were coupled with Rocchio's reweighting scheme. Asterisks are used to denote that the difference is statistically significant, using a two-tailed paired *t* test with a confidence level in excess of 95%.

The results show that expanded queries worked markedly better than nonexpanded queries for all expansion techniques and for all evaluation measures, the differences being almost always statistically significant. The performance of query expansion was better even for the first retrieved documents, where it is harder to improve over nonexpanded query. The only exception was PREC-AT-5 on TREC-8, where the unexpanded method was occasionally better than the expanded ones.

The results also show that the five expansion methods (Rocchio, RSV_R, $CHI\text{-}1_R$, $CHI\text{-}2_R$, and KLD_R) obtained similar average performance

Table II. Comparison of Retrieval Performance on TREC-8 Using Distributional Term Ranking

	Unexp.	Rocchio	RSV_R	$CHI\text{-}2_R$	$CHI\text{-}1_R$	KLD_R
RET&REL	2938	3111 +5.89%*	3109 +5.82%*	3119 +6.16%*	3075 +4.66%*	3120 +6.19%*
AV-PREC	0.2718	0.2972 +9.33%*	0.2960 +8.90%*	0.2960 +8.90%*	0.2954 +8.70%*	0.3002 +10.44%*
11-PT-PREC	0.2978	0.3174 +6.57%*	0.3189 +7.07%*	0.3172 +6.52%*	0.3181 +6.81%*	0.3226 +8.33%*
R-PREC	0.3168	0.3377 +6.58%	0.3329 +5.05%	0.3419 +7.89%*	0.3380 +6.69%*	0.3390 +6.98%*
PREC-AT-5	0.5960	0.6040 +1.34%	0.6080 +2.01%	0.5880 −1.34%	0.5800 −2.68%	0.5960 0%
PREC-AT-10	0.4920	0.5160 +4.88%*	0.5140 +4.47%*	0.5180 +5.28%*	0.5260 +6.91%*	0.5220 +6.10%*

improvement over nonexpanded query for all evaluation measures. On close inspection, Rocchio was slightly better than RSV_R, $CHI\text{-}1_R$, and $CHI\text{-}2_R$, and slightly worse than KLD_R, but the differences were negligible. Indeed, one of the main findings of this experiment is that as long as we employ Rocchio's formula for reweighting an expanded query, the use of a more sophisticated method for ranking and selecting expansion terms than Rocchio's itself does not seem to produce, on average, any performance improvement. These results extend to pseudorelevance feedback and to a larger database earlier findings about the low importance of selection metrics in the performance of relevance feedback systems [Salton and Buckley 1990; Harman 1992].

4.5 Experiment 2: Using KLD within Rocchio to Rank and Weight Expansion Terms

The four term-scoring functions introduced above can be used not only to select the expansion terms but also to replace Rocchio's weights in expression (2). More specifically, the weights of the vector r in expression (2) can be computed using the scores of the expressions (5)–(8), rather than using the document weights given by expression (11), as in expression (4). The weights of the vector r and the weights of the original query computed by expression (12) can then be normalized, to ensure consistency, and summed up.

Using the just described procedure to determine the weights of the expanded query, the expression (10) computing the second-pass ranking can be seen as the product of two components, each involving a distinct weighting scheme (i.e., a weighting of index terms with respect to the documents and a weighting of query terms with respect to the query). The

use of a compounded weighting scheme may better reflect the different importance of the same term with respect to the documents in the collection and with respect to the user query. The form of our function agrees with suggestions made by Robertson [1990] and Efthimiadis [1993], and it bears also a similarity with the retrieval with probabilistic indexing model [Fuhr 1989], which combines probabilistic indexing weighting with query-term weighting based on relevance feedback. The contention that the properties of document and query spaces are different and thus must be exploited separately has been also theoretically investigated by Bollmann-Sdorra and Raghavan [1998].

We computed the effectiveness of the four just introduced reweighting methods, using the same experimental conditions as Experiment 1. The normalization of the original query vector and of the expansion term vector was performed by dividing each weight by the maximum weight in the corresponding vector. To ensure fair comparison between the four distributional methods and basic Rocchio, we considered also a normalized version of the latter method, in which the weights of the original query, given by expression (12), and the weights of the expansion terms, given by expression (11), were normalized before being summed up. This is especially useful considering that great care is usually taken to ensure that the query weights and document weights in the Rocchio formula are commensurate, whereas the BM25 query weights and document weights being combined for the basic Rocchio method in our experiments might be very different from each other.

In Tables III and IV we report the retrieval performance of each distributional method and of normalized Rocchio on the test collections TREC-7 and TREC-8. We use a subscript NORM to indicate that such methods used the normalized weights for query expansion. We also report again, for convenience, the retrieval effectiveness of basic unnormalized Rocchio, denoted just Rocchio as in Tables I and II, and, as done in Experiment 1, we show the performance improvement of each method over ranking with nonexpanded query, used as a baseline.

The first interesting finding concerns the performance variation of Experiment 2 over Experiment 1. By comparing the performance values of distributional methods reported in Tables III and IV with those reported in Tables I and II, it is apparent that the variations were, in general, very different depending on the particular method and evaluation measure considered and on the test collection being tested. On the other hand, such a comparison also shows that the KLD method did considerably improve for most evaluation measures and on both test collections. On TREC-7, RET&REL increased by 6.06%, AV-PREC by 9.30%, 11-PT-PREC by 7.96%, R-PREC by 3.66%, PREC-AT-5 by 4.11%, while PREC-AT-10 decreased by 2.60%, with the first three differences being statistically significant. On TREC-8, RET&REL increased by 4.78%, AV-PREC by 1.70%, 11-PT-PREC by 0.10%, and PREC-AT-5 by 0.67%, while R-PREC and PREC-AT-10 decreased, respectively, by 1.08% and 1.92%, with the difference concerning RET&REL being statistically significant.

Table III. Comparison of Retrieval Performance on TREC-7 Using Distributional Term Ranking and Weighting

	Unexp.	Rocchio	$\text{Rocchio}_{\text{NORM}}$	RSV_{NORM}	$\text{CHI-2}_{\text{NORM}}$	$\text{CHI-1}_{\text{NORM}}$	KLD_{NORM}
RET&REL	2751	3009 +9.38%*	2872 +4.40%	3058 +11.16%*	3007 +9.31%*	3063 +11.34%*	3202 +16.39%*
AV-PREC	0.2291	0.2625 +14.54%*	0.2288 −0.15%	0.2670 +16.53%*	0.2555 +11.50%	0.2687 +17.26%*	0.2873 +25.39%*
11-PT-PREC	0.2545	0.2806 +10.25%*	0.2494 −2.02%	0.2858 +12.28%*	0.2747 +7.93%	0.2876 +12.99%*	0.3054 +19.98%*
R-PREC	0.2711	0.2972 +9.62%*	0.2700 −0.39%	0.3077 +13.50%*	0.2903 +7.08%	0.3012 +11.11%*	0.3061 +12.93%*
PREC-AT-5	0.5480	0.5760 +5.11%	0.5360 −2.19%	0.5680 +3.65%	0.5480 0%	0.5560 +1.46%	0.6080 +10.95%*
PREC-AT-10	0.5120	0.5240 +2.34%	0.4900 −4.30%	0.5160 +0.78%	0.5080 −0.78%	0.5120 0%	0.5240 +2.34%

Table IV. Comparison of Retrieval Performance on TREC-8 Using Distributional Term Ranking and Weighting

	Unexp.	Rocchio	$\text{Rocchio}_{\text{NORM}}$	RSV_{NORM}	$\text{CHI-2}_{\text{NORM}}$	$\text{CHI-1}_{\text{NORM}}$	KLD_{NORM}
RET&REL	2938	3111 +5.89%*	2657 −9.56%	2976 +1.29%	3173 +8.00%*	3217 +9.50%*	3269 +11.27%*
AV-PREC	0.2718	0.2972 +9.33%*	0.2345 −13.73%	0.2749 +1.14%	0.2798 +2.94%	0.2918 +7.35%	0.3053 +12.32%*
11-PT-PREC	0.2978	0.3174 +6.57%*	0.2559 −14.08%*	0.2963 −0.52%	0.3007 +0.96%	0.3127 +4.99%	0.3229 +8.44%*
R-PREC	0.3168	0.3369 +6.34%	0.2757 −13.00%	0.3215 +1.47%	0.3071 −3.08%	0.3377 +6.58%	0.3353 +5.82%
PREC-AT-5	0.5960	0.6040 +1.34%	0.5720 −4.03%	0.5800 −2.68%	0.5720 −4.03%	0.5480 −8.05%	0.6000 +0.67%
PREC-AT-10	0.4920	0.5160 +4.88%*	0.5020 +2.03%	0.5320 +8.13%*	0.4800 −2.44%	0.4840 −1.63%	0.5120 +4.07%

The main rationale for explaining a performance improvement when passing from Experiment 1 to Experiment 2 is the following. If one expansion term, for a given query, was correctly ranked ahead of another term, then the former should receive a proportionally higher weight in the expanded query, whereas if we use for query reweighting a weighting scheme that computes an absolute value of term goodness ignoring the specific relevance to the query at hand, like Rocchio's formula, then the better term might receive a lower weight than the worse term. Thus, the

use of distributional term scoring for query reweighting, as in Experiment 2, may better handle the possible mismatch between the relevance of a term to a given query and the relevance of the same term to the collection.

However, as mentioned, the performance improvement of Experiment 2 over Experiment 1 did not hold consistently. Some significant exceptions were found, most notably concerning the inferior performance values of most methods tested in the experiments on the TREC-8 test collection, that suggest that performance variations due to distributional query reweighting may be heavily dependent on the set of topics being used. Under this respect, our results partially contradict recent findings reported in Carpineto and Romano [1999], where the performance of distributional term-scoring functions improved in a more marked and consistent manner when passing from term selection to term weighting, while being more in accordance with other results concerning different query expansion techniques [Harman 1992; Yang et al. 1999].[4] Yang et al. [1999], in particular, discovered that retrieval feedback performed very differently when used in the various subcollections of TREC, whereas our results are relative to the use of the same set of documents searched using different sets of topics.

The other main finding of Experiment 2 was the excellent performance of the KLD method. The superiority of KLD over the other distributional methods tested in the experiments was apparent for almost all evaluation measures and on both test collections, with many differences being statistically significant. Of course this raises the question of why KLD scored better than the other distributional methods. The superiority of KLD over RSV and CHI-1 can be explained by considering that the former method, by definition, implicitly performs a smoothing operation over the range of values, as opposed to the latter ones. Due to the observed presence of a large fraction of suggested terms with very low scores, this might have resulted in a more balanced and representative set of weights. The better performance of KLD over CHI-2 may depend on the fact that CHI-2 may select, besides good terms, terms which are good indicators for nonrelevance [Ballerini et al. 1997].

The KLD method was also, in general, better than basic Rocchio, whereas the other distributional methods obtained worse performance than Rocchio for most data points and did not improve over Rocchio in a statistically significant manner for any evaluation measure. In particular, on TREC-7, KLD fared consistently better than Rocchio (RET&REL: +6.41%; AV-PREC: +9.47%; 11-PT-PREC: +8.83%; R-PREC: +3.01%; PREC-AT-5: +5.56%; PREC-AT-10: 0), and, on TREC-8, KLD had better performance values than Rocchio for RET&REL (+5.08%), AV-PREC (+4.62%), and 11-PT-PREC (+1.75%), and slightly worse values for R-PREC(−0.47%), PREC-AT-5 (−0.66%), and PREC-AT-10 (−0.78%). Of these differences, those concerning RET&REL, AV-PREC, and 11-PT-PREC on TREC-7, and

[4]The results shown in Carpineto and Romano [1999] have a more limited scope than those reported in this paper, because the experiments were performed on a smaller scale and using a simplistic baseline ranking system with very low absolute retrieval performance.

RET&REL on TREC-8, were statistically significant. Furthermore, Table IV shows that the performance of normalized Rocchio was very poor, even lower than the unexpanded case, thus ruling out the possibility that the relatively low performance of basic Rocchio was due to normalization problems.

Compared to the baseline, the performance gain of KLD grew dramatically for almost all evaluation measures and across both test collections, with a peak of 25.39% for average precision on TREC-7. Besides relative performance improvement, we should emphasize that the results obtained by the KLD method are very good results even on an absolute scale. Considering average precision as the measure for performance comparison, the KLD results would be ranked among those of the best systems in both TREC-7 and TREC-8. Indeed, the retrieval performance reported here is consistent with that obtained using a similar approach at TREC-8 [Carpineto and Romano 2000b]. These results are even more remarkable because they were achieved without using sophisticated and computationally expensive indexing techniques (e.g., based on natural language processing).

A further comment about the results obtained by KLD concerns the generation of expansion terms with negative weights. As already remarked, the use of expression (8) may, in principle, produce negative weights: this happens whenever a term is comparatively less frequent in R than in C. Since the number of pseudorelevant documents is usually a small fraction of the collection size, this situation is unlikely. In fact, it was found in the experiments that, for some topics, there were some expansion terms with negative KLD weights. For instance, TREC-8's topic 402 yielded two expansion terms with negative weights: "Tuesday" and "help." However, because we selected only the 40 terms with highest score, terms with negative weights hardly contributed to the final query expansion. Similar considerations hold for CHI-1.

As most approaches to automatic query expansion, including those considered here, rely on a number of parameters, it is important to study how these parameters affect performance. This issue is discussed in the next section.

5. EFFECT OF METHOD PARAMETERS ON PERFORMANCE

5.1 Data Sparseness

The experiments described above were performed assuming that all candidate expansion terms were contained in R. Although some theoretical considerations may justify this choice, it is useful to investigate the effect of expanding the set of candidate terms to include the whole vocabulary V.

Let $v(R) \subset V$ be the vocabulary of the terms which do appear in the documents in R. For the terms not in $v(R)$, it is useful to introduce a back-off probability. The use of a back-off probability to overcome the data sparseness problem has been extensively studied in statistical language modeling. A possible back-off scheme assumes that a nonoccurring term

has the same probability as in the whole collection [Ponte and Croft 1998]. A better scheme is based on discounting the probabilities of the observed terms while the probability mass recuperated in this way is redistributed over the unobserved terms (see Katz [1987], de Mori [1998] for a survey, and Bigi et al. [1997] or Dagan et al. [1999], for recent applications). The following back-off scheme is used here:

$$p_R(t) = \begin{cases} \mu \dfrac{f_R(\mathrm{t})}{NT_R} & if\ t \in v(R) \\ \theta p_C(t) & otherwise \end{cases} \tag{13}$$

with

$$p_C(t) = \frac{f_C(\mathrm{t})}{NT_C}$$

and

$$\sum_{t \in v(R)} f_R(t) = NT_R.$$

In order to ensure that probabilities of all terms sum to 1, the following relation must hold:

$$\mu + \theta \sum_{t \notin v(R)} p_C(t) = \mu + \theta A = 1 \tag{14}$$

where

$$A = \sum_{t \notin v(R)} p_C(t).$$

The back-off probability can be used to compute the KLD score of all terms in V. For the terms contained in R, it is sufficient to choose a value for μ and substitute expression (13) in expression (8). For the terms not contained in R, the following procedure can be used:

(1) choose $0 < \mu < 1$,

(2) find a proper value for θ through expression (14),

(3) use the computed value of θ to estimate $p_R(t)$, $t \notin v(R)$ according to expression (13),

(4) compute score(t), $t \notin v(R)$, with expression (8), using the value of $p_R(t)$ found in step 3.

The KLD score of any term that is not contained in R is always negative, because the value returned by the just described procedure is $\theta \mathrm{p_C(t)} \cdot \log\theta$, with $\theta < 1$ by expression (14). In terms of the relative entropy between the two distributions p_R and p_C, given by expression (3), this means that any

Table V. Comparison of Mean Retrieval Performance

	TREC-7		TREC-8	
	KLD ($\mu = 1$)	KLD ($\mu = 0.9$)	KLD ($\mu = 1$)	KLD ($\mu = 0.9$)
RET&REL	3202	3149	3269	3221
AV-PREC	0.2873	0.2821	0.3053	0.3012
11-PT-PREC	0.3054	0.3000	0.3230	0.3181
R-PREC	0.3061	0.3070	0.3353	0.3319
PREC-AT-5	0.6080	0.6180	0.6000	0.6000
PREC-AT-10	0.5240	0.5240	0.5120	0.5180

such term yields a negative summand, thus making p_R less divergent with respect to p_C. In fact, the presence of those terms in a document will decrease the relevance of that document to the query associated with R.

In our experiments, however, since we use only the 40 terms with the highest score, negatively weighted terms would be hardly selected for query expansion. In order to take practical advantage of negative weights, we used a different experimental procedure.

The score of each term not contained in R was computed, and then the 10 terms with the lowest scores (i.e., highest absolute values) were selected. We took also into account the terms contained in R that obtained negative weights, even though they did not affect the result because they were always ranked behind the 10 best terms not contained in R. Eventually, a query was expanded using both the best 40 positive terms, as in the experiments described above, and the 10 best negative terms. As it turns out that for practical values of μ the absolute value of the best negative terms may be significantly lower than the absolute values of the best positive terms, positive and negative terms were normalized separately, and then added to the original query.

Table V shows, for the test collections TREC-7 and TREC-8, the retrieval performance of basic KLD ($\mu = 1$) and of KLD augmented with negative weights. The latter were obtained using $\mu = 0.9$. It was observed that lower values of μ caused a performance degradation. The experimental setting was the same as Experiment 2, using the KLD scores both to select and weight terms.

Although the use of negative weights led to a small improvement on the system's precision for the first retrieved documents, which is an important performance measure in many applications, it is apparent that the variation in retrieval effectiveness was, in general, very limited. Our results seem to indicate that the use of negative weights did not significantly alter the overall performance, similar to earlier results on relevance feedback [Salton and Buckley 1990; Harman 1992] and more recent findings on the use of pseudononrelevant documents in the TREC environment [Hawking et al. 1998].

In fact, it is unlikely that any straightforward selection from the terms that do not occur in relevant documents or occur in nonrelevant documents would have any significant impact on performance. Selection policies that

exploit richer contextual information seem to be necessary in order to fully take advantage of such terms. One example is to try to identify those negative terms that cooccur with significant positive terms, and are thus likely to occur in nonrelevant documents that would otherwise be retrieved high in the ranking without such negative weights.

Finally, it is worth noticing that although the approach described above produced terms that contributed to the expanded query with negative weights, this should not be confused with extracting information from a set of nonrelevant documents. In fact, we did not use any set of nonrelevant documents; we extracted negatively weighted terms from the same set of relevant documents used to detect positively weighted terms. If we were to take advantage of feedback from nonrelevant documents, we might instead apply the same procedure as positive feedback to extract and weight a set of terms characterizing the nonrelevant documents, this time subtracting the weights of such terms while reweighting the whole query (see the last term of full Rocchio's formula in expression (1)). We did not investigate this approach in our experiments.

5.2 Query Difficulty

One of the key factors to success of automatic query expansion is the quality of the initial retrieval run. In particular, one might expect that query expansion will work well if the top retrieved documents are good and that it will perform badly if they are poor. Xu and Croft [1996], for instance, found that pseudorelevance feedback tends to hurt queries with baseline average precision less than 5%. To test this hypothesis more deeply, we studied how the retrieval effectiveness of the information-theoretic method varied as the difficulty of a query changed, where the latter was characterized by the average precision of the initial run relative to the given query (the lower the average precision, the greater the difficulty).

The results are shown in Figures 1 and 2, one for each test collection. Each circle represents one of the 50 queries; if the circle is above (below) the bisecting line, then the performance increased (decreased) when we passed from nonexpanded to expanded query. The query difficulty decreases as we move away from the origin.

These results are somewhat unexpected, because no clear pattern seems to emerge. The performance improvement does not monotonically grow with easiness of query. Indeed, if we split the x-axis in intervals and compute the average performance of the queries within each interval, then it is easy to see that performance variation is initially very limited and occasionally negative, as expected, and then it starts climbing until it reaches a maximum, after which it declines and may drop below zero.

While this issue needs to be studied more carefully, using other test collections and query sets, our experiment seems to support the hypothesis that queries with low precision do not carry useful information for improvement, while queries with high initial precision can be hardly further improved upon. As an indication to achieve further mean improvement, one

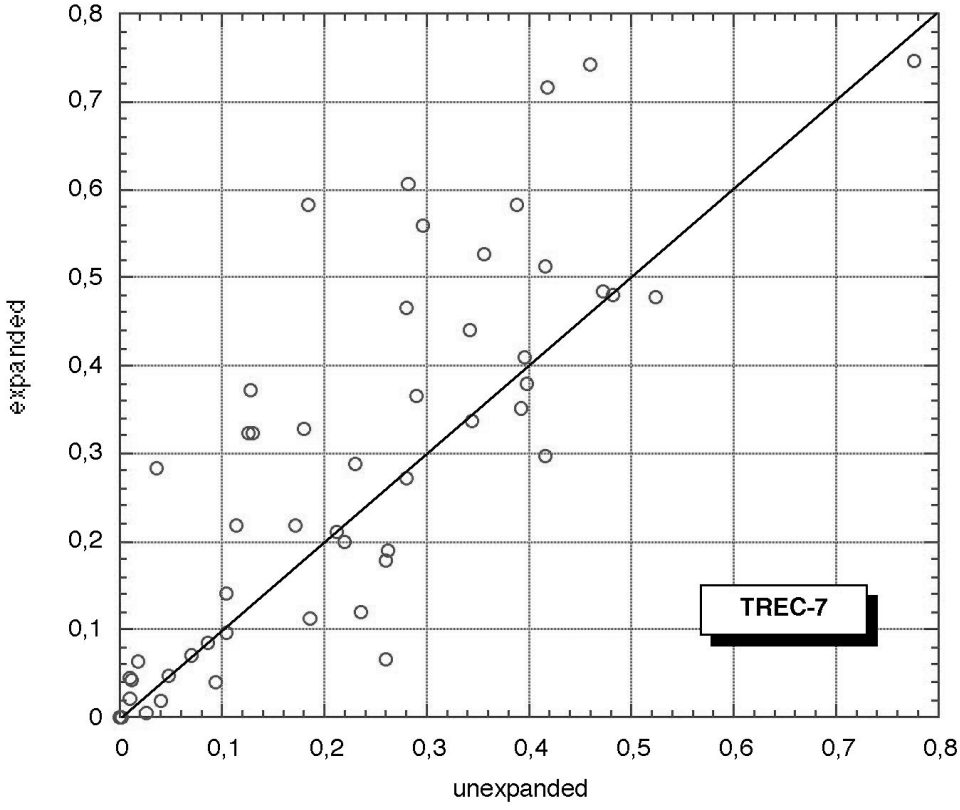

Fig. 1. Performance versus initial query difficulty for KLD on TREC-7.

might try to develop selective policies for query expansion that focus on queries that are neither too difficult nor too easy. In practice, however, the assessment of the difficulty of a given query without retrospective feedback, as in our experiments, may result rather problematic, although it is conceivable that useful hints may be obtained by some kind of preprocessing (e.g., linguistic analysis, similarity to past queries with known difficulty, etc.).

Two other main parameters of automatic query expansion systems are the number of pseudorelevant documents used to collect expansion terms and the number of terms selected for query expansion. In the next two sections, we will consider each of them, in turn.

5.3 Number of Pseudorelevant Documents

Based on the fact that the density of relevant documents is higher for the top-ranked documents, one might think that the fewer the number of documents considered for expansion, the better the retrieval performance. However, this was not the case. As shown in Tables VI and VII, where the maximum value of each measure is displayed in bold, the retrieval performance

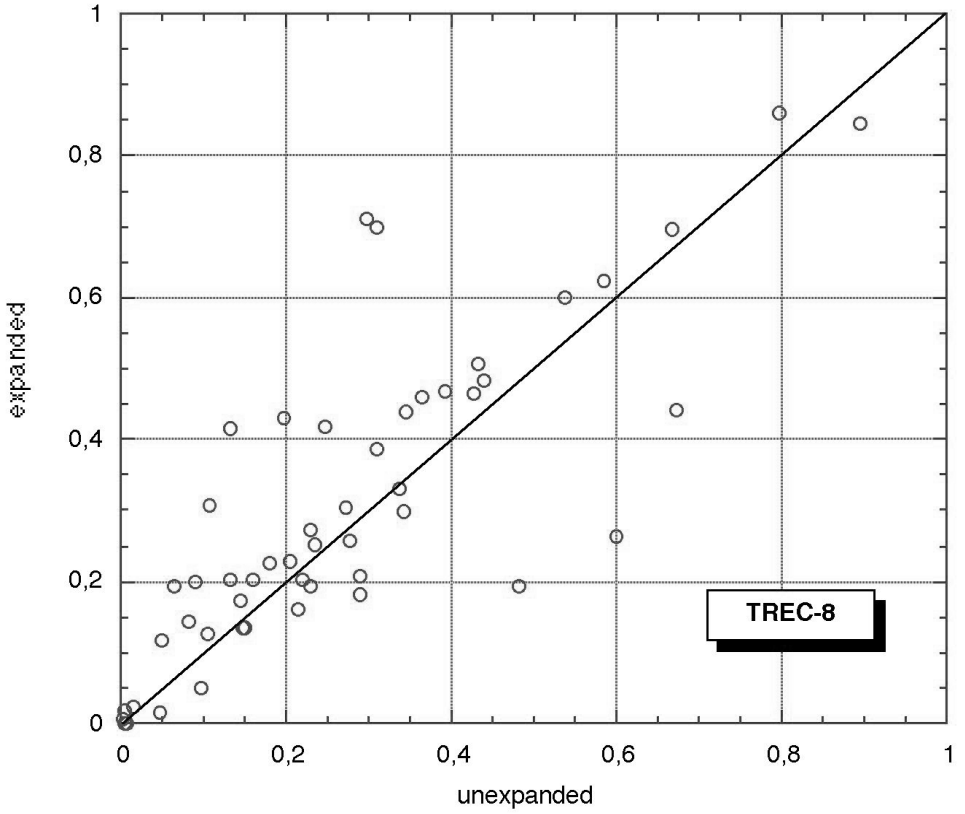

Fig. 2. Performance versus initial query difficulty for KLD on TREC-8.

was found to increase as the number of documents increased, at least for a small number of documents, and then it gradually dropped as more documents were selected.

This behavior can be explained by considering that the percentage of truly relevant documents in the pseudorelevant documents is not the only factor affecting performance here. If we select a very small number of pseudorelevant documents, it is more likely that we will get, for some queries, no relevant document at all, which may produce very bad results on those queries and a mean performance degradation. It is also conceivable that with very few relevant documents the system simply does not have enough information, and is very dependent on the idiosyncrasies of those documents (i.e., a sort of sampling view of the problem).

Thus, the optimal choice should represent a compromise between the maximization of the percentage of relevant documents and the presence of enough relevant documents. Consistently with the results reported in Tables VI and VII, we found that these two parameters were best balanced when the size of the training set ranged from 4 to 10. For smaller sizes, the number of queries with no relevant documents was proportionally higher;

Table VI. Performance versus Number of Pseudorelevant Documents for KLD on TREC-7

	2	4	6	8	10
RET&REL	3026	3121	3194	3186	**3202**
AV-PREC	0.2576	0.2708	0.2862	0.2862	**0.2873**
11-PT-PREC	0.2804	0.2898	0.3035	0.3029	**0.3054**
R-PREC	0.2906	0.3034	**0.3066**	0.3065	0.3061
PREC-AT-5	0.5520	0.5760	0.5800	0.5920	**0.6080**
PREC-AT-10	0.5220	0.5180	**0.5480**	0.5440	0.5240

Table VI. *Continued*

	12	14	16	18	20
RET&REL	3156	3157	3122	3079	3094
AV-PREC	0.2834	0.2843	0.2712	0.2663	0.2657
11-PT-PREC	0.2989	0.3001	0.2870	0.2815	0.2804
R-PREC	0.2958	0.3000	0.2904	0.2832	0.2873
PREC-AT-5	0.5960	0.5760	0.5520	0.5440	0.5520
PREC-AT-10	0.5240	0.5280	0.5000	0.5080	0.5100

for larger sizes, the percentage of relevant documents became proportionally smaller.

The optimal number of pseudorelevant documents for TREC-8 was lower than TREC-7 due to the different query difficulty of the two test collections. As shown in Tables I and II, the difference between PREC-AT-5 and PREC-AT-10 of the first-pass ranking was proportionally higher in TREC-8 than TREC-7. The results about query difficulty are thus consistent with those reported in Tables VI and VII.

Further experiments revealed that the system performance decreased nearly monotonically as the number of documents was increased beyond those shown in Tables VI and VII. The decline in performance was however slow, because the percentage of relevant documents remained substantially high even after a large number of retrieved documents. For instance, for TREC-8, the average precision at 60 documents was 0.2583, at 100 documents was 0.2311.

5.4 Number of Expansion Terms

We performed some experiments to see how the retrieval performance was affected by changes in the number of expansion terms. We let the number of expansion terms vary from 10 to 100 (step = 10), computing for each value the retrieval performance of the system. Tables VIII and IX show that the maximum values of the different performance measures were reached for different choices of the number of selected terms. Most important, the results show that the variations in performance were very limited for all measures and for all selected sets of expansion terms. This behavior held across both test collections.

Our findings are consistent with earlier results reported by Salton and Buckley [1990], suggesting that query expansion using all terms from the

Table VII. Performance versus Number of Pseudorelevant Documents for KLD on TREC-8

	2	4	6	8	10
RET&REL	3057	3281	**3299**	3280	3269
AV-PREC	0.2759	**0.3158**	0.3028	0.3048	0.3053
11-PT-PREC	0.2962	**0.3352**	0.3231	0.3229	0.3230
R-PREC	0.3119	**0.3425**	0.3343	0.3374	0.3353
PREC-AT-5	0.6040	**0.6320**	0.6000	0.6000	0.6000
PREC-AT-10	0.5220	0.5220	**0.5240**	0.5200	0.5120

Table VII. *Continued*

	12	14	16	18	20
RET&REL	3258	3262	3242	3263	3256
AV-PREC	0.3001	0.2964	0.2967	0.2951	0.2978
11-PT-PREC	0.3173	0.3155	0.3151	0.3128	0.3153
R-PREC	0.3318	0.3275	0.3290	0.3295	0.3291
PREC-AT-5	0.5920	0.5880	0.5840	0.5720	0.5840
PREC-AT-10	0.5200	0.5220	0.5040	0.5020	0.5060

retrieved relevant documents may be only slightly better than a selection of those terms. On the other hand, using only a limited number of expansion terms may be important to reduce response time, especially for large collections. The latter is therefore also our recommended choice for automatic query expansion.

6. CONCLUSIONS AND FUTURE WORK

An information-theoretic method for query expansion has been introduced. A comprehensive set of experiments has shown the effectiveness of the method with respect to unexpanded queries as well as to a number of techniques for query expansion.

The following major conclusions seem to emerge from the experimental evaluation:

—The information-theoretic method is more effective when used within Rocchio's framework not only for selecting expansion terms but also for weighting them.

—Of the four term-scoring methods based on distribution analysis, the information-theoretic method is the only one which leads to some significant improvements over basic Rocchio.

—The retrieval effectiveness of the information-theoretic method is not altered by the number of expansion terms selected and is only moderately affected by the generation of expansion terms with negative weights. Furthermore, the performance increases as the query difficult decreases, although very easy queries were observed to hurt performance. Similarly, the performance improves as the number of pseudorelevant

Table VIII. Performance versus Number of Expansion Terms for KLD on TREC-7

	10	20	30	40	50
RET&REL	3103	3172	3184	3202	3212
AV-PREC	0.2822	0.2837	0.2859	0.2873	0.2892
11-PT-PREC	0.3024	0.3025	0.3042	0.3054	0.3060
R-PREC	0.3044	0.3044	0.3031	0.3061	0.3054
PREC-AT-5	0.5800	0.6040	0.6000	**0.6080**	**0.6080**
PREC-AT-10	0.5300	0.5140	0.5160	0.5240	0.5260

Table VIII. *Continued*

	60	70	80	90	100
RET&REL	3221	3220	3219	**3226**	3225
AV-PREC	**0.2919**	0.2916	0.2914	0.2916	**0.2919**
11-PT-PREC	0.3080	0.3077	0.3073	0.3082	**0.3085**
R-PREC	0.3068	**0.3070**	0.3064	0.3060	0.3054
PREC-AT-5	**0.6080**	**0.6080**	0.6040	0.6000	0.6000
PREC-AT-10	0.5260	**0.5280**	0.5260	0.5240	0.5220

Table IX. Performance versus Number of Expansion Terms for KLD on TREC-8

	10	20	30	40	50
RET&REL	3209	3256	3272	3269	3261
AV-PREC	0.2949	0.3032	0.3037	**0.3053**	0.3040
11-PT-PREC	0.3127	0.3221	0.3213	0.3230	0.3216
R-PREC	0.3299	0.3352	0.3334	0.3353	0.3354
PREC-AT-5	0.5800	0.5840	0.5920	**0.6000**	0.5960
PREC-AT-10	**0.5220**	0.5180	0.5160	0.5120	0.5120

Table IX. *Continued*

	60	70	80	90	100
RET&REL	3270	3272	**3278**	3270	3268
AV-PREC	0.3044	0.3043	0.3035	0.3037	0.3037
11-PT-PREC	0.3227	**0.3228**	0.3220	0.3221	0.3220
R-PREC	**0.3362**	0.3351	0.3342	0.3349	0.3345
PREC-AT-5	0.5880	0.5880	0.5960	0.5960	0.5960
PREC-AT-10	0.5120	0.5080	0.5120	0.5100	0.5080

documents considered for expansion gets smaller, unless very few documents are considered.

This work can be extended in several directions. While we mainly focused on theoretical measures for term quality, there are also other aspects of the proposed approach to query expansion that might be evaluated more carefully such as query length and robustness of probability estimation.

The queries used in the experiments were rather long, whereas the advantages for query expansion should be larger with short queries. Some encouraging indications about the effectiveness of the KLD method for

short queries are provided by the recent TREC-8 evaluations, where the performance of a system using information-theoretic query expansion [Carpineto and Romano 2000b], while being on an absolute scale higher for the full-topic run (title + description + narrative), was comparatively better than the other TREC participants for the short-topic run (just title and description). As short queries may better reflect a real situation, this issue deserves more study.

To estimate probabilities, one obvious choice is to use alternative simple functions. In our experiments, as remarked in Section 4.2, we tried the document-based probability in addition to the maximum likelihood probability. This issue might be investigated more thoroughly by considering a larger set of estimation functions, such as those suggested by Robertson et al. [1995], and evaluating how their use affect overall performance. A more complex approach is to model the number of occurrences of each term by using a known distribution rather than a binary value function. For instance, we could use a binomial or the Poisson, or even a multiple parameter distribution such as the N-Poisson [Margulis 1993]. Such a richer model might better capture the heterogeneous structure of documents, while lending itself to being used in our query expansion framework with small changes.

A third topic for future research is to use the power of ensembling [Dietterich 1997]. Similarly to the fact that retrieval effectiveness in the domain of text categorization can be improved by combining different classification methods [Larkey and Croft 1996], one may expect that combining the individual decisions of a set of term-scoring functions may result in a more accurate and effective representation for the expanded query than any individual function. One simple, and perhaps nonoptimal, solution to implement this strategy has been explored in Carpineto and Romano [1999] with promising results.

ACKNOWLEDGMENTS

This work has been carried out within the framework of an agreement between the Italian PT Administration and the Fondazione Ugo Bordoni. We would like to thank Susan Dumais, Stephen Robertson, and three anonymous reviewers for their valuable comments and suggestions. We would also like to thank Giambattista Amati for several useful discussions about term weighting issues.

REFERENCES

AMATI, G. AND VAN RIJSBERGEN, K. 2000. Probabilistic models of information retrieval based on measuring the divergence from randomness.

ATTAR, R. AND FRAENKEL, A. S. 1977. Local feedback in full-text retrieval systems. *J. ACM 24*, 3 (July), 397–417.

BALLERINI, J. P., BUCHEL, M., DOMENIG, R., KNAUS, D., MATEEV, B., MITTENDORF, E., SCHAUBLE, P., SHERIDAN, P., AND WECHSLER, M. 1996. SPIDER Retrieval System at TREC-5. In *Proceedings of the 5th Conference on Text Retrieval* (TREC-5, Gaithersburg, MD, Nov.), E.

M. Voorhees and D. K. Harman, Eds. National Institute of Standards and Technology, Gaithersburg, MD, 217–228.

BIGI, B., DE MORI, R., EL-BÈZE, M., AND SPRIET, T. 1997. Combined models for topic spotting and topic-dependent language modeling. In *Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Press, Piscataway, NJ, 535–542.

BOLLMANN-SDORRA, P. AND RAGHAVAN, V. V. 1998. On the necessity of term dependence in a query space for weighted retrieval. *J. Am. Soc. Inf. Sci. 49*, 13, 1161–1168.

BRAJNIK, G., MIZZARO, S., AND TASSO, C. 1996. Evaluating user interfaces to information retrieval systems: A case study on user support. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '96, Zurich, Switzerland, Aug. 18–22), H.-P. Frei, D. Harman, P. Schaüble, and R. Wilkinson, Chairs. ACM Press, New York, NY, 128–136.

BUCKLEY, C., SALTON, G., ALAN, J., AND SINGHAL, A. 1995. Automatic query expansion using SMART: TREC3. In *Proceedings of the 3rd Conference on Text Retrieval* (TREC-3, Gaithersburg, MD), D. Harman, Ed. National Institute of Standards and Technology, Gaithersburg, MD, 69–80.

CARPINETO, C. AND ROMANO, G. 1998. Effective reformulation of Boolean queries with concept lattices. In *Proceedings of the 3rd International Conference on Flexible Query-Answering Systems* (FQAS 98, Roskilde, Denmark). Springer-Verlag, Heidelberg, Germany, 83–94.

CARPINETO, C. AND ROMANO, G. 1999. Towards better techniques for automatic query expansion. In *Proceedings of the 3rd European Conference on Digital Libraries* (ECDL'99, Paris, France). 126–141.

CARPINETO, C. AND ROMANO, G. 2000a. Order-theoretical ranking. *J. Am. Soc. Inf. Sci. 51*, 7, 587–613.

CARPINETO, C. AND ROMANO, G. 2000b. TREC-8 automatic ad-hoc experiments at FUB. In *Proceedings of the 8th Conference on Text Retrieval* (TREC-8, Gaithersburg, MD). 377–380.

COOPER, J. W. AND BYRD, R. J. 1997. Lexical navigation: Visually prompted query expansion and refinement. In *Proceedings of the 2nd ACM International Conference on Digital Libraries* (DL '97, Philadelphia, PA, July 23–26), R. B. Allen and E. Rasmussen, Chairs. ACM Press, New York, NY, 237–246.

COVER, T. M. AND THOMAS, J. A. 1991. *Elements of Information Theory*. Wiley series in telecommunications. Wiley-Interscience, New York, NY.

CROFT, W. AND HARPER, D. J. 1979. Using probabilistic models of document retrieval without relevance information. *J. Doc. 35*, 285–295.

DAGAN, I., LEE, L., AND PEREIRA, F. 1999. Similarity-based models of word cooccurrence probabilities. *Mach. Learn. 34*, 43–69.

DE MORI, R. 1998. *Spoken Dialogues with Computers*. Academic Press, Inc., New York, NY.

DEERWESTER, S., DUMAI, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci. 41*, 6, 391–407.

DIETTERICH, T. 1997. Machine-learning research: Four current directions. *AI Mag. 18*, 4, 97–135.

DOSZCOCKS, T. E. 1978. AID: An associative interactive dictionary for online searching. *Onl. Rev. 2*, 2, 163–174.

EFTHIMIADIS, E. N. 1993. A user-centred evaluation of ranking algorithms for interactive query expansion. In *Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '93, Pittsburgh, PA, June 27–July), R. Korfhage, E. Rasmussen, and P. Willett, Eds. ACM Press, New York, NY, 146–159.

FITZPATRICK, L. AND DENT, M. 1997. Automatic feedback using past queries: Social searching? In *Proceedings of the 20th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '97, Philadelphia, PA, July 27–31), W. Hersh, F. Can, and E. Voorhees. ACM Press, New York, NY, 306–313.

FUHR, N. 1989. Models for retrieval with probabilistic indexing. *Inf. Process. Manage. 25*, 1, 55–72.

FURNAS, G. W., LANDAUER, T. K., GOMEZ, L. M., AND DUMAIS, S. T. 1987. The vocabulary problem in human-system communication. *Commun. ACM 30*, 11 (Nov.), 964–971.

GREFENSTETTE, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Hingham, MA.

HARMAN, D. 1992. Relevance feedback and other query modification techniques. In *Information Retrieval: Data Structures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 241–263.

HARPER, D. J. AND VAN RIJSBERGEN, C. J. 1978. An evaluation of feedback in document retrieval using co-currence data. *J. Doc. 34*, 3, 189–216.

HAWKING, D., THISTLEWAITE, P., AND CRASWELL, N. 1998. ANU/ACSys TREC-6 experiments. In *Proceedings of the 6th Conference on Text Retrieval* (TREC-6), E. Voorhees, Ed. 275–290. NIST Special Publication 500-240.

HEARST, M. A. AND PEDERSEN, J. O. 1996. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '96, Zurich, Switzerland, Aug. 18–22), H.-P. Frei, D. Harman, P. Schaüble, and R. Wilkinson, Chairs. ACM Press, New York, NY, 76–84.

KARP, D., SCHABES, Y., ZAIDEL, M., AND EGEDI, D. 1992. A freely available wide coverage morphological analyzer for English. In *Proceedings of the 14th International Conference on Computational Linguistics* (COLING '92, Nantes, France). 950–954.

KATZ, S. 1987. Estimation of probabilities from sparses data for language model component of a speech recognizer. *IEEE Trans. Acoust. Speech Signal Process. 35*, 400–401.

LARKEY, L. S. AND CROFT, W. B. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '96, Zurich, Switzerland, Aug. 18–22), H.-P. Frei, D. Harman, P. Schaüble, and R. Wilkinson, Chairs. ACM Press, New York, NY, 289–297.

LOSEE, R. M. 1990. *The science of information: Measurements and applications*. Academic Press Prof., Inc., San Diego, CA.

MARGULIS, E. L. 1993. Modelling documents with multiple Poisson distributions. *Inf. Process. Manage. 29*, 2 (Mar.-Apr.), 215–227.

MITRA, M., SINGHAL, A., AND BUCKLEY, C. 1998. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 206–214.

PONTE, J. M. AND CROFT, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 275–281.

PORTER, M. F. 1982. Implementing a probabilistic information retrieval system. *Inf. Tech. Res. Dev. Appl. 1*, 2, 131–156.

ROBERTSON, S. E. 1991. On term selection for query expansion. *J. Doc. 46*, 4 (Dec. 1990), 359–364.

ROBERTSON, S. E. AND SPARCK JONES, K. 1976. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci. 27* (May), 129–146.

ROBERTSON, S. E., WALKER, S., AND BEAULIEU, M. 1999. Okapi at TREC-7: Automatic ad hoc, filtering, VLC, and interactive track. In *Proceedings of the 7th Conference on Text Retrieval* (TREC-7, Gaithersburg, MD). 253–264.

ROBERTSON, S. E., WALKER, S., JONES, S., HANCOCK-BEAULIEU, M. M., AND GATFORD, M. 1994. Okapi at TREC-3. In *Proceedings of the 3rd Text Retrieval Conference* (TREC-3, Nov.). 109–126.

ROCCHIO, J. 1971. Relevance feedback in information retrieval. In *The Smart Retrieval System: Experiments in Automatic Document Processing*, G. Salton, Ed. Prentice-Hall, Englewood Cliffs, NJ, 313–323.

SALTON, G. AND BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage. 24*, 5, 513–523.

SALTON, G. AND BUCKLEY, C. 1990. Improving retrieval performance by relevance feedback. *J. Am. Soc. Inf. Sci. 41*, 4, 288–297.

SCHAPIRE, R. E., SINGER, Y., AND SINGHAL, A. 1998. Boosting and Rocchio applied to text filtering. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98, Melbourne, Australia, Aug. 24–28), W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Chairs. ACM Press, New York, NY, 215–223.

SINGHAL, A., CHOI, J., HINDLE, D., LEWIS, D., AND PEREIRA, F. 1999. AT&T at TREC-7. In *Proceedings of the 7th Conference on Text Retrieval* (TREC-7, Gaithersburg, MD). 239–252.

SRINIVASAN, P. 1996. Query expansion and MEDLINE. *Inf. Process. Manage. 32*, 4, 431–443.

VAN RIJSBERGEN, C. J. 1977. A theoretical basis for the use of co-occurence data in information retrieval. *J. Doc. 33*, 2 (June), 106–119.

VAN RIJSBERGEN, C. 1979. *Information Retrieval*. 2nd ed. Butterworths, London, UK.

VAN RIJSBERGEN, C. J., HARPER, D. J., AND PORTER, M. F. 1981. The selection of good search items. *Inf. Process. Manage. 17*, 2, 77–91.

VÉLEZ, B., WEISS, R., SHELDON, M., AND GIFFORD, D. 1997. Fast and effective query refinement. In *Proceedings of the 20th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '97, Philadelphia, PA, July 27–31), W. Hersh, F. Can, and E. Voorhees. ACM Press, New York, NY, 6–15.

VOORHEES, E. M. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR '94, Dublin, Ireland, July 3–6), W. B. Croft and C. J. van Rijsbergen, Eds. Springer-Verlag, New York, NY, 61–69.

VOORHEES, E. AND HARMAN, D. 1998. Overview of the sixth text retrieval conference (TREC-6). In *Proceedings of the 6th Conference on Text Retrieval* (TREC-6), E. Voorhees, Ed. 1–24. NIST Special Publication 500-240.

VOORHEES, E. M. AND HARMAN, D. K. 1999. Overview of the seventh text retrieval conference (TREC-7). In *Proceedings of the 7th Conference on Text Retrieval* (TREC-7, Gaithersburg, MD). 1–23.

WONG, S. K., ZIARKO, W., RAGHAVAN, V. V., AND WONG, P. C. 1987. On modeling of information retrieval concepts in vector spaces. *ACM Trans. Database Syst. 12*, 2 (June), 299–321.

XU, J. AND CROFT, W. B. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '96, Zurich, Switzerland, Aug. 18–22), H.-P. Frei, D. Harman, P. Schaüble, and R. Wilkinson, Chairs. ACM Press, New York, NY, 4–11.

XU, J. AND CROFT, B. 1999. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '99, Berkeley, CA). 254–261.

XU, J. AND CROFT, B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst. 18*, 1, 79–112.

YANG, K., MAGLAUGHLIN, K., MEHO, L., AND SUMNER, R. G. JR. 1999. IRIS at TREC-7. In *Proceedings of the 7th Conference on Text Retrieval* (TREC-7, Gaithersburg, MD). 555–566.

Received: March 2000; revised: September 2000 and December 2000; accepted: December 2000