Towards more effective techniques for automatic query expansion

Claudio Carpineto and Giovanni Romano
Fondazione Ugo Bordoni, Via B. Castiglione 59, I-00142, Rome, Italy
{carpinet, romano}@fub.it

Abstract

Techniques for automatic query expansion from top retrieved documents have recently shown promise for improving retrieval effectiveness on large collections but there is still a lack of systematic evaluation and comparative studies. In this paper we focus on term-scoring methods based on the differences between the distribution of terms in (pseudo-)relevant documents and the distribution of terms in all documents, seen as a complement or an alternative to more conventional techniques. We show that when such distributional methods are used to select expansion terms within Rocchio's classical reweighting scheme, the overall performance is not likely to improve. However, we also show that when the same distributional methods are used to both select and weight expansion terms the retrieval effectiveness may considerably improve. We then argue, based on their variation in performance on individual queries, that the set of ranked terms suggested by individual distributional methods can be combined to further improve mean performance, by analogy with ensembling classifiers, and present experimental evidence supporting this view. Taken together, our experiments show that with automatic query expansion it is possible to achieve performance gains as high as 21.34% over non-expanded query (for non-interpolated average precision). We also discuss the effect that the main parameters involved in automatic query expansion, such as query difficulty, number of selected documents, and number of selected terms, have on retrieval effectiveness.

1. Introduction

Experience with operational search systems reveals a significant mismatch between their theoretical assumptions and the actual user behavior. While these systems are designed to take advantage of the presence of many query terms to describe a user's information need, the average number of user-supplied query terms is usually very small, often less than 2. The paucity of query terms exacerbates well known inherent limitations of information retrieval systems, such as the difficulty of recovering from word mismatch between queries and documents, and it may represent a fundamental practical limitation for effective retrieval from large databases. Much of the current research in information retrieval attempts to solve this problem by focusing on methods for the creation of a query "context" by using such diverse knowledge sources as user's relevance feedback (Harman, 1992), thesauri (Cooper and Byrd, 1997),

and conceptual clustering of documents and terms (Carpineto and Romano, 1998), rather than concentrating on better ways of matching queries against documents.

One well known, automatic approach to adding contextual information to user queries is based on the extraction of useful terms from the top retrieved documents, which is also referred to as retrieval feedback or pseudo-relevance feedback. While this technique did not, historically, work well, due to losses in precision being higher than gains in recall, it has recently received renewed attention for its successful application to large scale collections (e.g., Buckley *et al*., 1995; Xu and Croft, 1996; Fitzpatrick and Dent, 1997, Mitra *et al*., 1998). In the TREC environment, for instance, more recently almost all groups have been using variations on expanding queries using information from the top retrieved documents, but the benefits of different query expansion techniques have been usually evaluated with respect to using non-expanded query and not by cross-system comparisons. The growing interest in pseudo-relevance feedback calls for a more careful and systematic evaluation of competing approaches and for a better understanding of their relative strengths and weaknesses.

In this paper we focus on term-scoring functions that are based on the differences between the distribution of terms in (pseudo-)relevant documents and the distribution of terms in all documents. We consider several instances of this general "distributional" approach, including Robertson Selection Value (Robertson, 1990) as well as statistical and information-theoretic functions. We study how to use these distributional functions to improve effectiveness of automatic query expansion. We first analyze 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 negative. We then use the same distributional methods to select and weight expansion terms, this time showing considerable performance improvement over Rocchio's formula, with or without term selection based on distribution analysis.

The results of the latter experiment encourages a deeper query-by-query analysis. We learn that while the distributional methods may achieve comparable mean performance, they may also present large variations on individual queries both on the ranked set of suggested terms and on the retrieval performance. This observation suggests using combination strategies, by analogy with ensembling classifiers in the machine learning field. We present a simple approach to combining the results of multiple distributional methods and show that the combined method may perform better than the individual methods, thus further increasing the performance improvement of expanded query over non-expanded query (up to 21.34% for non-interpolated average precision). We finally study how the retrieval performance varies as a function of the main parameters involved in automatic query expansion, including query difficulty, number of selected documents, and number of selected terms, showing interesting relationships.

The rest of the paper is organized as follows. Section 2 characterizes the main phases of the automatic query expansion process and discusses the rationale of using term-ranking methods based on distribution

analysis. Section 3 precisely introduces the distributional methods tested in the experiments and evaluates their use to select expansion terms within Rocchio's classical reweighting scheme in contrast with basic Rocchio. Section 4 evaluates the performance of the distributional methods when they are used to both select and weight expansion terms. Section 5 analyzes the performance variations of distributional methods on individual queries. Section 6 describes a method to combine the results of multiple distributional query expansion methods and evaluate its performance. Section 7 discusses the role played by the main parameters involved in automatic query expansion in determining the overall effectiveness, and Section 8 provides some conclusions and directions for future work.

2. Approaches to automatic query expansion

To better represent the user information need we can extract useful terms from the results of an initial retrieval run. The idea, not new (Attar and Fraenkel, 1977; Croft and Harper, 1979), is to consider the top few documents retrieved as being relevant, in the absence of any real relevance judgements. Working from this assumption, the process which leads to a query with modified weights and terms typically goes through three main phases: expansion term location, expansion term ranking, and weighting of expanded query.

2.1 Expansion term location

The typical source of evidence for expanding a given query is constituted by all the terms in the first *r* documents retrieved in response to the query from the collection at hand, although more sophisticated schemes for locating the candidate expansion terms have been proposed, such as using passages (Xu and Croft, 1996, Hawking *et al*., 1998), or using the result of past similar queries (Fitzpatrick and Dent, 1997), or running the initial pass on a much larger collection than the target collection (Singhal *et al*., 1999).

2.2 Expansion term selection

The selection of expansion terms is usually performed by ranking candidate terms first, and then choosing the highest ranked terms. For ranking expansion terms, a number of different methods have been proposed, following two main conceptually distinct approaches. One straightforward solution is to rank the candidate expansion terms using the (primary) term weights *w(t)* computed for document ranking (Srinivasan, 1996; Mitra *et al*., 1998; Singhal *et al*., 1999). Usually, the score used for inclusion in the expanded query is given by $\sum_{k=1}^{r} w(t)_{Doc_k}$, where the summation index ranges over the first *r* retrieved documents. This approach is simple and computationally efficient, but it has the disadvantage that each term weight reflects the usefulness of that term with respect to the entire collection. In order to discriminate between good expansion terms and poor expansion terms it seems more convenient to

consider occurrence in relevant documents in comparison to occurrence in all documents. In other words, one may assume that the differences between the distribution of terms in the overall document collection and the distribution of the same terms in a set of relevant documents are related to semantic factors. It is expected, in particular, that good terms will occur with a higher frequency in relevant documents than in the whole collection, and poor terms will occur with the same frequency (randomly) in both. An early example of this approach is in (Doszkocs, 1978), where a comparative statistical analysis of term occurrences – via a chi-square variant - is used to suggest potentially relevant terms for interactive query expansion. A more general theoretical argument that supports the use of the differences in term distribution to select the terms to be included in the expanded query was provided by Robertson (1990). He showed that the inclusion of the term *t* in the expanded query will, under certain strong assumptions, increase the retrieval effectiveness by $w_t (p_t - q_t)$, where w_t is the primary weight of the term *t*, and p_t and q_t are the probabilities that a relevant and a non-relevant document, respectively, contain the term *t*. In fact, variants of Robertson's ranking scheme for expansion terms have subsequently been used by various systems, with different weighting functions and different methods for estimating p_t and q_t (Buckley *et al.*, 1995; Robertson et *al.*, 1995, Hawking *et al.*, 1998). An alternative, more recent, approach to using the differences in term distribution for selecting expansion term relies on the relative entropy, or Kullback-Lieber distance, between the two distributions, from which a computationally simple and theoretically justified method to assign scores to candidate expansion terms can be derived (Carpineto *et al.*, 1999).

2.3 Reweighting of expanded query

Most systems that perform retrieval feedback rely on Rocchio's formula (Rocchio, 1971), as improved by Salton and Buckley (1990), to expand and reweight the query terms (Srinivasan, 1996; Singhal *et al.*, 1999). In the retrieval feedback setting, it is usually assumed that the relevant documents are the *r* top documents retrieved by the systems and that the information about the number of non-relevant documents is absent. The simplified formula becomes:

$$w(\mathrm{t})_{Q\,\mathrm{exp}} = \alpha \; w(\mathrm{t})_{Q\mathrm{non-exp}} + \frac{\beta}{\mathrm{r}} \sum_{k=1}^{r} w(t)_{Doc_k}. \qquad (1)$$

It should be noted that the simple method for ranking expansion terms illustrated above is based on their proposed Rocchio weights. We should also emphasize that some modified versions of Rocchio's formula have recently been proposed that showed better performance than basic Rocchio on tasks involving proper relevance feedback (Buckley and Salton, 1995; Shapire *et al.*, 1998). We did not investigate such extensions in our experiments.

3. Using distributional term selection within Rocchio's weighting scheme

Having introduced different term-scoring methods, the first goal of our experiments was to evaluate the relative performance of these methods on selecting expansion terms. This comparison requires caution,

because the overall retrieval effectiveness may be a compound effect that masks the variables under study. In order to ensure a controlled experiment, we varied only the method used for selecting expansion terms while keeping the other factors involved in the query expansion process constant. Most important, to reweight the query after selection of expansion terms we uniformly used Rocchio's formula reported in expression (1), with α=1, β=1. We used as test collection the TREC-7 collection (TREC disks 4 and 5, containing approximately 2 Gigabytes of data) and query set (topics 351-400). The underlying basic ranking system used in the experiments by all four methods was developed in the context of our participation in TREC-7, and thus its data structures were specifically designed and implemented to efficiently handle the large TREC test collection. The system uses a vector space model with cosine normalization; documents and queries are weighted with the classical *tf·idf* scheme, after word stopping and stemming. The same test collection and basic ranking system were also used in subsequent experiments.

The five term-ranking functions tested in the experiment were the following (R indicates the pseudo-relevant set, C the whole collection, and w(t) is the weight of term t in the collection):

- Rocchio's weights: $score(t) = \sum_{k=1}^{r} w(t)_{Doc_k}$

- Robertson Selection Value (RSV):[1] $score(t) = \sum_{k=1}^{r} w(t)_{Doc_k} \cdot p_R(t)$

- CHI-square (CHI2): $score(t) = [p_R(t) - p_C(t)]^2 / p_C(t)]$

- Doszkocs' variant of CHI-square (CHI1): $score(t) = [p_R(t) - p_C(t)] / p_C(t)]$

- Kullback-Lieber distance (KD): $score(t) = [p_R(t) - p_C(t)] \cdot \log [p_R(t) / p_C(t)]$

We considered as candidate expansion terms those contained in R. To estimate $p_R(t)$, we used the ratio between the frequency of t in R, treated as a long string, and the number of terms in R; analogously, to estimate $p_C(t)$, we used the ratio between the frequency of t in C and the number of terms in C. The estimation of probabilities is an important issue because it might affect performance results. Although we have not fully worked out this aspect, we tried also different estimation functions such as the number of pseudo-relevant documents that contain the term (Buckley *et al*., 1995; Robertson *et al*., 1995), which however seemed to produce worse retrieval effectiveness. Finally, all term-ranking methods tested in the experiment required two values for practical implementation: the number of pseudo-relevant training documents and the number of expansion terms considered for inclusion in the expanded query.

[1] We assumed, as also done in (Robertson et al., 1995), that the probability that a non-relevant document contains the term *t* is negligible.

Consistently with many TREC's researchers, in our experiment the values of the two thresholds were set at 5 and 30, respectively.

For each query, we ran the complete ranking system five times, one for each possible selection of the technique for selecting expansion terms. In Table 1 we report the retrieval performance of each method, averaged over the query set, and show the performance improvement over non-expanded query, used as a baseline. Performance was measured with the TREC's standard evaluation measures. In Table 1, the distributional methods are labeled with 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 one-tailed paired *t* test with a confidence level in excess of 95%.

The results shows that expanded queries worked better than non-expanded queries for all expansion techniques and for all evaluation measures, with the main exception of "Prec-at-10", although the differences usually were not statistically significant. Somewhat unexpectedly, the five expansion methods (Rocchio, RSV_R, $CHI\text{-}1_R$, $CHI\text{-}2_R$, and KD_R) obtained very similar average performance improvement over non-expanded query for all evaluation measures. Indeed, one of the most interesting 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 expansion terms than Rocchio's itself does not seem to produce, on average, any performance improvement. These results confirm and extend to a slightly different setting and a larger database earlier findings about the low importance of selection metrics in the performance of relevance feedback systems.

Table 1. Comparison of mean retrieval performance

	Non-expanded	ROCCHIO	RSV_R	$CHI\text{-}2_R$	$CHI\text{-}1_R$	KD_R
RET&REL	38.56	40.62 +5.34%*	40.54 +5.13%*	40.28 +4.46%*	40.10 +3.99%	40.60 +5.29%*
AV-PREC	0.1231	0.1280 +3.98%	0.1277 +3.74%	0.1262 +2.52%	0.1312 +6.51%	0.1279 +3.90%
11-PT-PREC	0.1502	0.1529 +1.84%	0.1526 +1.64%	0.1518 +1.07%	0.1567 +4.33%	0.1531 +1.93%
R-PREC	0.1694	0.1773 +4.69%*	0.1766 +4.25%	0.1776 +4.84%*	0.1824 +7.66%	0.1765 +4.19%
PREC-AT-5	0.3880	0.3920 +1.03%	0.3920 +1.03%	0.3880 0.00%	0.4040 +4.12%	0.3920 +1.03%
PREC-AT-10	0.3380	0.3360 -0.59%	0.3340 -1.18%	0.3300 -2.37%	0.3380 0.00%	0.3340 -1.18%

4. Comparing distributional reweighting schemes to Rocchio

The five term-scoring functions introduced above can be used not only to select the expansion terms but also to weight them in expression (1), instead of Rocchio's weights. The overall reweighting function becomes:

$$w(\mathrm{t})_{Q\exp} = \alpha \times w(\mathrm{t})_{Q\mathrm{non-exp}} + \beta \times score(t) \qquad (2)$$

We compared the effectiveness of the five reweighting methods derived from equation (2) to Rocchio's scheme (expression 1). The values of the several parameters needed to implement the four methods were chosen as in the earlier experiment (i.e., 5 pseudo-relevant documents, 30 expansion terms, α=1, β=1).

In Table 2 we report the retrieval performance of each method, averaged over the query set, and again show the performance improvement over ranking with non-expanded query, used as a baseline. Table 2 shows that the performance of RSV was, on the whole, slightly inferior to Rocchio, while the other threee distributional methods clearly outperformed Rocchio (and RSV). Compared to the baseline, the performance of the best three distributional methods was still comparable when we considered a very limited number of retrieved documents (i.e., for "Prec-at-5" and "Prec-at-10"), but it dramatically improved for all other evaluation measures, with statistically significant differences.

Thus, the main result of this experiment is that when a distributional method for term selection is also used for query reweighting, the overall retrieval effectiveness may considerably improve. Although this finding should not be over-generalized, because it was obtained for a specific combination of the parameters involved in the weighting schemes, it suggests that that if we have a good method for ranking expansion terms we should try to use it also for assigning weights to terms in the expanded query. The main rationale for this is that if one expansion term, for a given query, is correctly ranked ahead of another then it should receive a proportionally higher weight in the expanded query, while if we use for query reweighting a weighting scheme that computes an absolute value of term goodness ignoring the specific information associated with the query at hand, like Rocchio's formula, then the better term might receive a lower weight than the worse term. The low performance of RSV is consistent with this observation, because the RSV score is more of a variant of Rocchio than a distinct reweighting function based on the differences in term distribution. In the rest of the paper we concentrate on the three fully-distributional methods (i.e., CHI2, CHI1, and KD).

Table 2 shows also the three fully-distributional methods achieved more comparable retrieval performance with respect to one another. As these methods use different mathematical functions, we hypothesized that despite their similar mean effectiveness they would present considerable variation on individual queries. Therefore we decided to test this hypothesis through a query by query analysis.

Table 2. Comparison of mean retrieval performance

	Non-expanded	ROCCHIO	RSV	CHI-2	CHI-1	KD
RET&REL	38.56	40.62	41.56	43.38	42.50	43.16

		+5.34%*	+7.78%*	+12.50%*	+10.22%*	+11.93%*
AV-PREC	0.1231	0.1280	0.1243	0.1466	0.1471	0.1409
		+3.93%	+0.94%	+19.05%*	+19.46%*	+14.39%*
11-PT-PREC	0.1502	0.1529	0.1469	0.1695	0.1720	0.1644
		+1.84%	-2.16%	+12.87%*	+14.53%*	+9.46%
R-PREC	0.1694	0.1773	0.1683	0.1912	0.1970	0.1840
		+4.69%*	-0.61%	+12.87%*	+16.30%*	+8.63%*
PREC-AT-5	0.3880	0.3920	0.3640	0.3800	0.4000	0.3840
		+1.03%	-6.19%*	-2.06%	+3.09%	-1.03%
PREC-AT-10	0.3380	0.3360	0.3320	0.3520	0.3620	0.3400
		-0.59%	-1.78%	+4.14%	+7.10%*	+0.59%

5. Performance variation of distributional methods on individual queries

Xu and Croft (1996) used the overlap between the sets of suggested terms to compare the performance of different query expansion methods on single queries. We observed that in our case the use of such a simple evaluation measure would not help much disclosing the different behavior of the three methods due to their relatively high overlap. Thus, we used a more powerful measure related not also to which terms are suggested by each method but also to how those terms are ranked. We also measured the relative retrieval effectiveness of the terms suggested by each method on single queries.

Variations on term ranking. For each query and for each pair of methods we computed a measure of the difference between term rankings, considering only the first 30 terms suggested by each method. In particular, for each term in the ranked term list produced by one of the two methods, we computed the distance between the position of that term in the ranked lists produced by the two methods; if the term was not contained in the second list we assumed that it was ranked right after the last-ranked term (i.e., as 31st). We then averaged over the set of terms suggested for each query and over the set of queries. It should be noted that this measure is asymmetrical, because the results depends on which is the first selected method. The results are shown in Table 3; the first method to which each pairwise comparison refers is shown on columns. Considering that we used only the first 30 terms and that we assumed that all other terms were equally ranked as 31st, the most important finding is that the variation was substantially high for any pair of methods. In addition, the results show that the variation between KD and each of the other two methods was larger than that between CHI2 and CHI1, which are in nature more similar.

Table 3. Mean term distance in pairwise term-ranking comparison (restricted to the first 30 terms).

	CHI2	CHI1	KD
CHI2	0.00	6.29	11.58
CHI1	5.21	0.00	13.85
KD	11.00	13.77	0.00

Variation on retrieval effectiveness. For each query and for each expansion method, we measured the difference between the average precision obtained with expanded query and that obtained with non-expanded query. In Figure 1 we show for each query the minimum and maximum of such differences; thus, the length of each bar depicts the range of performance variations attainable by the three methods on each query. For most queries, the variations with respect to non-expanded query (x axis) were either all positive or all negative, as might be expected, although there were also a significant number of exceptions; most important, despite showing similar mean performance over the query set (see Table 2) but consistently with the term distance analysis, the inter-method variations on single queries were ample, with a mean value of 50.3%. Thus, methods which generated better terms on some queries produced poorer terms on others. The fact that individual methods disagreed with one another on individual queries while predicting, on average, equally good terms suggests trying combination strategies with the aim of retaining, on average, the most informative terms. This issue is discussed below.

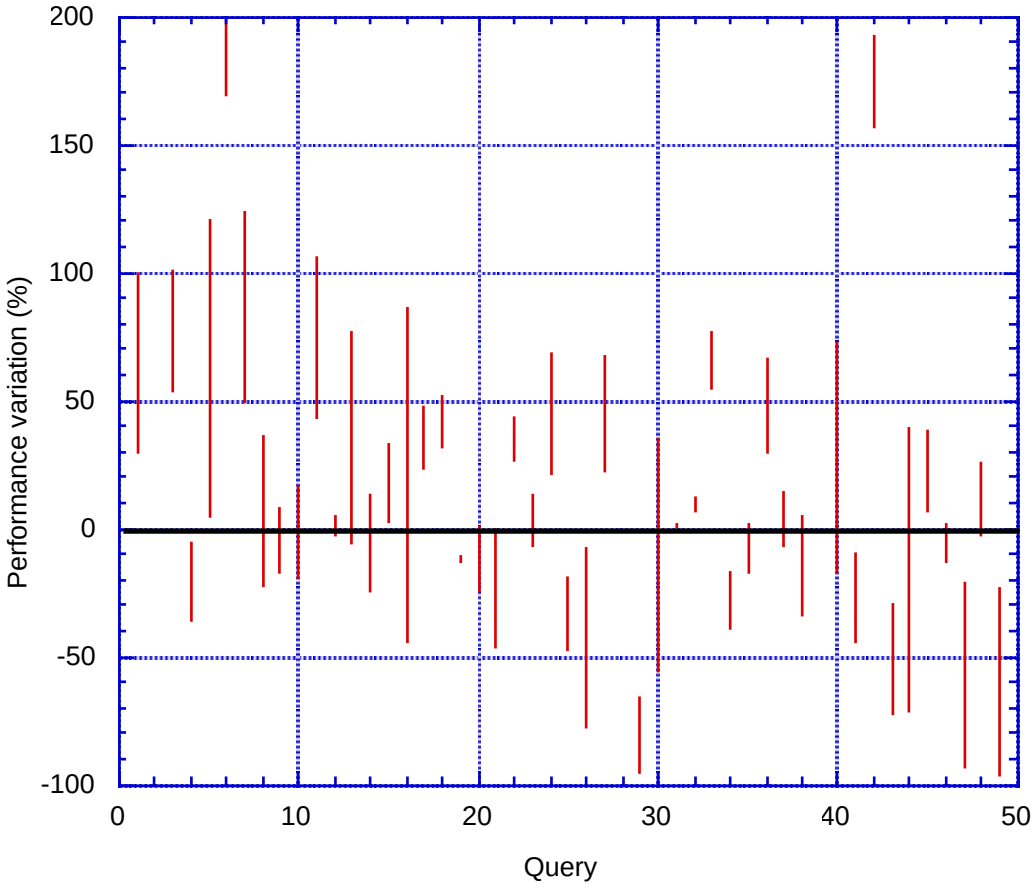

Figure 1. Performance variation of query expansion methods on individual queries.

6. Combining multiple distributional query expansion methods

Recent research in machine learning and information retrieval has shown that ensembling multiple classifiers, whether produced by single or different learning algorithms, may be a viable technique for improving classification accuracy (Larkey and Croft, 1996; Breiman, 1996; Dietterich, 1997). Two keys to success are that the individual classifiers must disagree with one another and that their average accuracies must be comparable. In this case one can try to guess the right prediction by taking a majority vote, in the hope that the single classifiers make uncorrelated errors.

In the retrieval feedback setting, the output of each method is represented by a ranked list of new terms instead of a sharp yes/no procedure as in concept classification. By analogy with ensembling classifiers, we can hypothesize that the individual methods make uncorrelated errors in suggesting new terms, i.e., when a term erroneusly gets a high rank in one method the same term gets a low rank in the other methods, so that a majority procedure can correctly rank the term. As described above, we successfully checked for mean performance and diversity of the individual retrieval feedback techniques, so the next issue is how to ensemble the results of the individual methods. One simple approach is to compute an average score for each term from the scores assigned to that term by individual methods and then use these new scores. This approach however would require that the scores produced by each method will have similar absolute values, otherwise the average scores will be dominated by the method with high score. This condition was not met by the retrieval feedback feedback methods tested in our experiment, because the KD scores were comparatively higher than other methods' scores. Therefore we took an alternative approach.

As the individual methods presented quite large variations on the order in which terms were ranked, we decided to focus on the differences between the relative position of each term in the three rankings, ignoring the term scores. Thus, the ranks of the terms were averaged and the mean was used to rerank them. Once the ranks have been merged, the relevance score of the terms can be computed by using some inverse function of their final position. We used as a new term-scoring function the simple ratio between 1 and the position of the term; i.e., 1 for the first term, 1/2 for the second term, 1/3 for the third term, etc. The scores obtained this way were used in equation (2) to assign weights to the new terms, and the resulting combined reweighting method was tested for performance using the same parameter setting as previuos experiments with individual methods.

The results are shown in Table 4, again with improvement over ranking with non-expanded query used as a baseline. A comparison between Table 4 and Table 2 shows that the combined method had better performance than any individual method for almost any evaluation measure, thus further improving the performance over non-expanded query. In particular, the performance improvement of average precision is especially notable (+21.34%) for this is the most informative evaluation measure of ranking performance. The results shown in Table 4 and Table 2 also indicate that the performance scores of the combined method represented, in general, a small improvement over the scores obtained by the best individual method. However, as combination strategies work best when the results being combined are

generated independently (Hull *et al.*, 1996), there are reasons to believe that such an improvement could be higher if we weakened some experimental parameters that are likely to increase the correlation between the term-relevance estimates of the individual methods (e.g., varying the document representation and the set of training data). Furthermore, the merging of several term ranks can be performed using more sophisticated techniques involving linear combination of individual ranks and parameter optimization, similar to work on combining multiple ranked document lists (Bartell *et al.*, 1994).

The results of this experiment should be taken with caution and cannot be easily generalized without further evidence, because they were obtained for specific ensembling methods and parameter combinations; nonetheless, since we used very simple and untuned functions, they represent an indication that this approach is feasible.

Table 4. Mean retrieval performance of combined method

REL-RET DOCS	AV-PREC	11-PT-PREC	R-PREC	PREC-AT-5	PREC-AT-10
45.72	0.1494	0.1733	0.1930	0.3920	0.3560
18.57%*	+21.34%*	+15.38%*	+13.96%*	+1.03%	+5.33%

7. Effect of method parameters on performance

As most approaches to automatic query expansion, including ensembling methods, rely on a number of parameters, it is important to study how these parameters affect performance. One of the key factor to success 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 pseudo-relevance 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 combined 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 Figure2. 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 non-expanded 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 negative, as expected, and then it starts climbing until it reaches a maximum (initial precision of 20-30%), after which it declines and may drop again below zero. In fact,

our experiment supports the view 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 might develop selective policies for query expansion that focus on queries that are neither too difficult nor too easy.

Two other main parameters of automatic query expansion systems are the number of pseudo-relevant documents used to collect expansion terms and the number of terms selected for query expansion. We performed some experiments to see how the retrieval performance varied as a function of these two parameters. Let us consider first the number of documents. Based on the ground 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. The retrieval performance 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 considering that the percentage of truly relevant documents in the pseudo-relevant documents is not the only factor affecting performance here. If we select a very small number of pseudo-relevant 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. Thus, the optimal choice should represent a compromise between the maximization of the percentage of relevant documents and the presence of at least some relevant document. Consistently with the results reported above, we found that these two parameters were best balanced when the size of the training set ranged from 4 to 12; for smaller sizes the number of queries with no relevant documents was proportionally higher, for larger sizes the percentage of nonrelevant documents grew large.

The results concerning the variation of the retrieval performance with the number of expansion term were more predictable. We found that the performance improvement initially increased as more terms were selected, at least as long as we selected truly new terms (consider that the first suggested terms usually coincide with the original query terms), and then it gradually decreased as more and more less-informative terms were chosen.

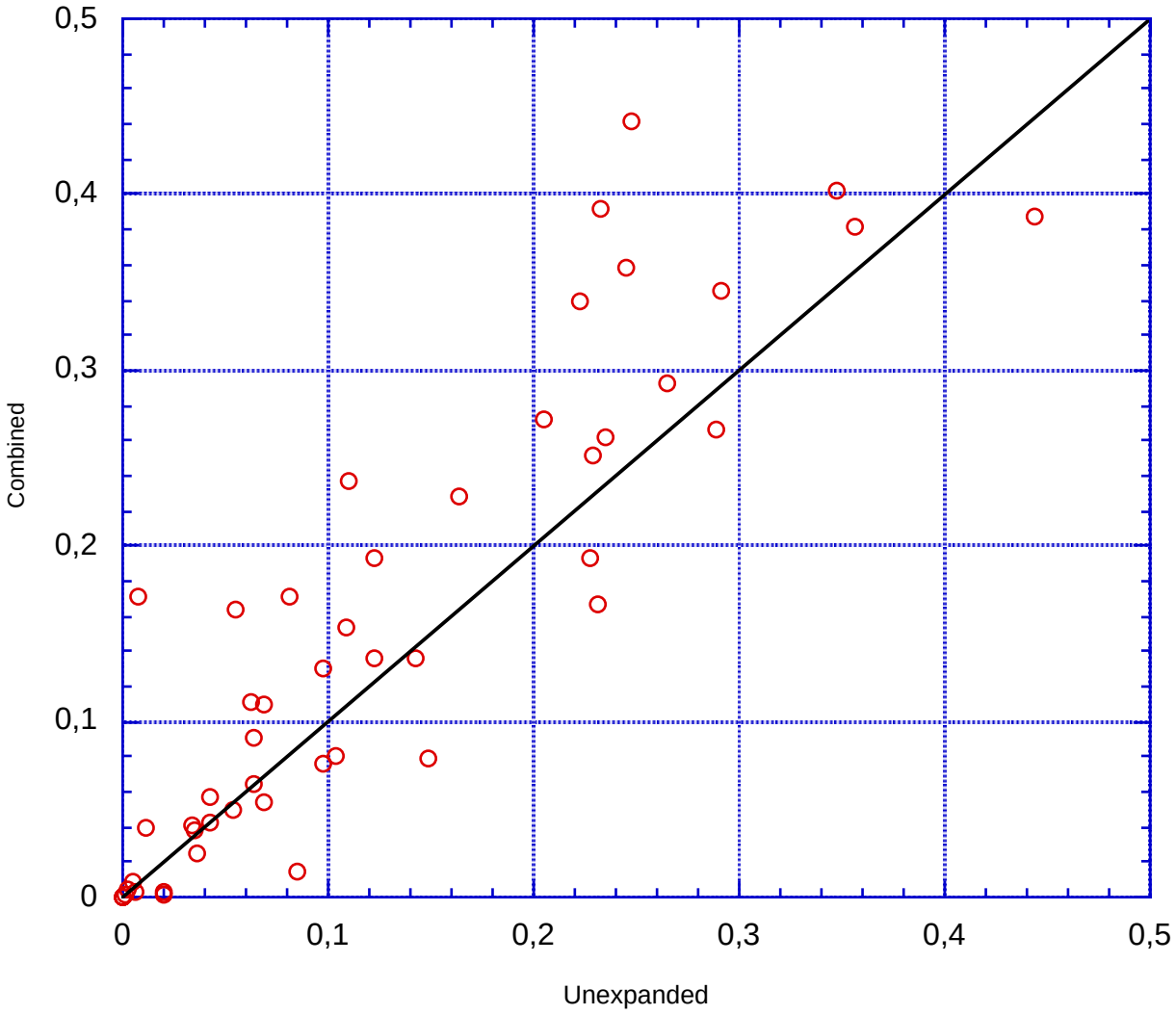

Figure 2. Improvement versus initial query difficulty

8. Conclusions

This paper extended earlier results about the effectiveness of automatic query expansion techniques in several directions. In particular, from our experimental evaluation, three main conclusion can be drawn.

- Term-scoring methods based on distribution analysis are not likely to improve performance when they are used only to select expansion terms, but the same methods may produce a considerable performance improvement when they are used to both select and reweight the expansion terms.
- The combination of the set of expansion terms produced by different distributional methods may perform better than the individual methods.
- The retrieval performance of automatic query expansion usually increases as the query difficult decreases, but it may decrease as the query becomes very easy. Similarly, the optimal number of

pseudo-relevant documents and expansion terms should represent a compromise between using little new information and much new information.

While we mainly focused on term selection and term reweighting, there are also other aspects of the proposed approach to query expansion that need be evaluated more carefully such as robustness of probability estimation and combination of multiple results. Aside from experimental investigations, we need a better theoretical understanding of the relative strengths and weaknesses of the individual query expansion techniques and of why their combination may work well. Also, having ascertained the importance of term reweighting over term selection in the good performance of distributional methods in a pseudo-relevance feedback task, it is tempting to see if these methods can be used as primary term-weighting schemes, in a proper relevance feedback environment. Finally, our approach could be used to generate good search terms not only for automatic query expansion but also in interactive searches, with the aim of help users to expand or refine a query based on the actual content of the collection. We are currently investigating these issues.

Acknowledgments

This work has been carried out within the framework of an agreement between the Italian PT Administration and the Fondazione Ugo Bordoni.

References

Attar, R., and Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery*, 2493), 397-417.

Bartell, B., Cottrell, G., and Belew, R. (1994). Automatic combination of multiple ranked retrieval systems. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pp. 173-181, Dublin.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2), pp. 123-140.

Buckley, C., and Salton, G. (1995). Optimization of relevance feedback weights. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, pp. 351-357, Seattle.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1995). Automatic query expansion using SMART: TREC3. *Proceedings of the third Text REtrieval Conference (TREC-3)*.

Carpineto, C., and Romano, G. (1998). Effective reformulation of Boolean queries with concept lattices. *Proceedings of the 3rd International Conference on Flexible Query-Answering Systems (FQAS'98)*, Lecture Notes in Artificial Intelligence, Springer Verlag, pp. 83-94.

Carpineto, C., De Mori, R., and Romano, G. (1999). Informative term selection for automatic query expansion. *Proceedings of the Seventh Text Retrieval Conference (TREC-7).*

Cooper, J., & Byrd, R. (1997). Lexical navigation: visually prompted query expansion and refinement. *Proceedings of the 2nd ACM Digital Library Conference*, pp. 237-246.

Croft, B., and Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35, 285-295.

Dietterich, T. (1997). Machine-learning research: four current directions. *AI Magazine*, Winter 1997, pp. 97-135.

Doszcocks, T.E. (1978). AID: an associative interactive dictionary for online searching. *Online Review* 2(2), pp. 163-174.

Fitzpatrick, L., and Dent, M. (1997). Automatic feedback using past queries: social searching? *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pp. 306-313, Philadelpia.

Hawking, D., Thistlewaite, P., and Craswell, N. (1998). ANU/ACSys TREC-6 Experiments. In D. K. Harman, editor, Proceedings of the Sixth Text Retrieval Conference (TREC-6).

Harman, D. (1992). Relevance feedback and other query modification techniques. In Information Retrieval - Data Structures and Algorithms, Frakes, W.B., and Baeza-Yates, R. (Eds.), pp. 241-263, Prentice Hall, Englewood Cliffs, NJ.

Hull, D., Pedersen, J., and Schutze, H. (1996). Method combination for document filtering. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pp. 279-287, Zurich.

Larkey, L., and Croft, B. (1996). Combining classifiers in text categorization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pp. 289-297, Zurich.

Mitra, M., Singhal, A., and Buckley, C. (1998). Improving automatic query expansion. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pp. 206-214, Melbourne.

Robertson, S.E. (1990). On term selection for query expansion. *Journal of Documentation*, 46(4), pp. 359-364.

Robertson, S.E., Walker, S., Jones, G.J.F., Hancock-Beaulieu, and Gatford, M. (1995). Okapi at TREC-3. *Proceedings of the third Text REtrieval Conference (TREC-3).*

Rocchio, J. (1971). Relevance feedback in information retrieval. In Salton, G. (ed.), *The SMART retrieval system - experiments in automatic document processing*, chapter 14, Prentice Hall, Englewood Cliffs.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288-297.

Shapire, R, Singer, Y., and Singhal, A. (1998). Boosting and Rocchio applied to text filtering. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pp. 215-223, Melbourne.

Singhal, A., Choi, J., Hindle, D., Lewis, D., and Pereira, F. (1999). AT&T at TREC-7. *Proceedings of the Seventh Text Retrieval Conference (TREC-7).*

Srinavasan, P. (1996). Query expansion and MEDLINE. *Information Processing & Management,* 32 (4), pp. 431-443.

Xu, J., and Croft, B. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96),* pp. 4-11, Zurich.