TREC-8 Automatic Ad-Hoc Experiments at Fondazione Ugo Bordoni

Claudio Carpineto and Giovanni Romano
Fondazione Ugo Bordoni, Rome Italy
{carpinet, romano}@fub.it

Abstract

We present further evidence suggesting the feasibilty of using information theoretic query expansion for improving the retrieval effectiveness of automatic document ranking . Compared to our participation in TREC-7, in which we applied this technique to an ineffective initial ranking, here we show that information theoretic query expansion may be effective even when the quality of the first pass ranking is high. In TREC-8 our system has been ranked among the best systems for both automatic ad hoc and short automatic ad hoc. These results are even more interesting considering that we used single-word indexing and well known weighting schemes. We also investigate the use of term variance to refine the weighting schemes employed by our system to weight documents and queries.

1. Introduction

This is our second participation in the TREC conference. In TREC-7 we experimented with a novel method for automatic query expansion based on an information-theoretic measure (Carpineto and Romano, 1999). The results showed that passing from unexpanded to expanded query yielded a high performance gain (+ 14%), but, on an absolute scale, the figures that we obtained were much lower than those reported by the best TREC systems (e.g., 0.1409 versus 0.3033 for the average precision). This was mainly due to the ineffectiveness of the ranking system on top of which the query expansion stage was added, which used a simple *tf·idf* weighting scheme. Thus, in TREC-8 we have tried to improve the basic ranking system in the hope of increasing not only the retrieval effectiveness of the document ranking produced in response to a query (whether unexpanded or expanded) but also the utility of the query expansion itself. Since the latter usually increases as the quality of the initial retrieval becomes higher, we expected that also the relative performance improvement due to query expansion would benefit from a better first pass ranking. Our results confirmed this hypothesis partially. The second goal of our participation in TREC-8 was to test the effectiveness of using the variance of term occurrence in the collection to refine the weighting schemes used to weight query and documents. The results obtained using term variance were moderately promising. Overall, the average precision of our best run in TREC-8 was 0.3106, which was a great improvement over that of our participation in TREC-7. Moreover, these results were excellent also on an absolute scale. Our system was ranked as the fourth best system for automatic short ad hoc, and as the eighth best system for automatic ad hoc. What makes these results even more interesting is that they were obtained using single-keyword indexing.

2. Test collection indexing

1. *Text segmentation*. Our system first identified the individual terms occurring in the test collection, ignoring punctuation and case.

2. *Word stemming*. To extract word-stem forms, we used a very large *trie*-structured morphological lexicon for English (Karp et al, 1992), that contains the standard inflections for nouns (singular, plural, singular genitive, plural genitive), verbs (infinitive, third person singular, past tense, past participle, progressive form), adjectives (base, comparative, superlative).

3. *Stop wording*. We used a stop list, contained in the CACM dataset, to delete from the texts common function words. In addition, for efficiency reasons, we removed the terms that appeared in more than 75000 and less than 5 documents.

All the test collection indexing was of the single-keyword type. In particular, we used no manually-

predefined multiword phrases to conflate groups of related words into single concepts.

3. First pass ranking

We used Okapi formula (Robertson et al., 1999) for matching queries and documents in the first pass ranking:

$$sim(q,d) = \sum_{t \in q \wedge d} w_{d,t} \cdot w_{q,t} \quad (1)$$

with $w_{d,t}$ given by

$$\frac{(k_1 + 1) \cdot f_{d,t}}{k_1 \cdot \left[(1-b) + b \cdot \frac{W_d}{avr\_W_d}\right] + f_{d,t}} \quad (2)$$

and $w_{q,t}$ given by

$$\frac{(k_3 + 1) \cdot f_{q,t}}{k_3 + f_{q,t}} \cdot \log \frac{N - f_t + 0.5}{f_t + 0.5} \quad (3)$$

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.

4. Information-theoretic query expansion

To automatically expand the query we used the Rocchio formula (Rocchio, 1971) coupled with an information-theoretic term scoring function, similar to the approach described in (Carpineto et al., 1999). The Rocchio formula for pseudo-relevance feedback is:

$$Q_{new} = \alpha \cdot Q_{orig} + \frac{\beta}{|R|} \sum_{r \in R} r \quad (4)$$

where Q_{new} is a weighted term vector for the expanded query, Q_{orig} is a weighted term vector for the original unexpanded query, R is a set of top retrieved documents (assumed to be relevant), and r is a weighted term vector extracted from R.

Using basic Rocchio, the weights of the terms contained in Q_{orig} and r are determined considering their primary weights, i.e., the weight of each term as determined by the weighting scheme used to produce the first pass ranking. In our approach, Q_{orig} was indeed the weighted query vector used in the first pass ranking, i.e., as determined by expression (3), but then we used a different method than expression (2) to weight each expansion term contained in r. Our weighting was based on the Kullback-Lieber distance between the distribution of the term in R and the distribution of the term in the whole collection. More precisely, each expansion term was assigned a score given by:

$$score(t) = [p_R(t) - p_C(t)] \cdot \log [p_R(t) / p_C(t)] \quad 5)$$

where $p_X(t)$ is the probability of occurrence of term t in the set of documents X, R indicates the pseudo-relevant set, C indicates the whole collection. The main rationale of using a term-scoring function based on distribution analysis to reweight expansion terms is that in this way a term that is good for a given query can receive a high score even when its weight in the collection is low, whereas with basic Rocchio there may be a mismatching between the relevance of a term to a given query and the weight actually assigned to it (Carpineto and Romano, 1999).

From a practical point of view, we considered as expansion candidates all terms contained in R, and then selected those with the highest score. To estimate $p_X(t)$, we used the ratio between the number of occurrences of t in X, treated as a long string, and the total number of terms in X. The other parameters involved in this method were chosen as follows. The constant α and β in expression (1) were set to 1 and 1.5, respectively; we used 12 pseudo-relevant documents and 50 expansion terms were considered for inclusion in the expanded query. The choice for the values of these parameters was based on earlier results obtained in past TREC conferences and on some experiments that we performed on the TREC-7 data.

5. Refining query and document weighting with term variance

Classical weighting scheme usually do not take into account the variance (σ^2) of term occurrence in the collecction of documents. This may be a useful information to identify terms that truly differentiate and relate subsets of documents. Generally speaking, the lower the variance of a term, the less likely it is that the term is a good one. Common function words, for

instance, are likely to exhibit a low variance because they tend to occur in similar proportions in all documents. In order to compute the variance of a term, we considered the whole set of documents (including those that did not contain that term) and used two alternative methods to compute the term occurrence in a document: term frequency (tf), and normalized term frequency wrt document length (norm-tf).

We used the term variance to modify both the document weights and the scores used to weight the expansion terms. To change the document weights, we simply multiplied expression (2) by $\log\sigma^2$. The modification of the expansion terms weights required more caution, because the introduction of σ^2 may interact with the information-theoretic measure used to weight the expansion terms in the first place. Recall that, using expression (5), a term is assigned a high score if it is much more concentrated in the top documents than the whole collection. This indication about the goodness of a term may be inversely related to the term variance, in the sense that a small variance may further imply that the term appears only in the top documents while a high variance may suggest that the behavior captured by expression (5) will be more likely attributed to chance. This implies that the terms to be favored to refine query weighting are those with small variance, rather than high variance. In fact, experimenting with TREC-7, we noticed that using σ^2 as a factor hurt performance, while using its inverse improved performance. Thus, we modified the weights of expansion terms by the inverse of $\log\sigma^2$.

It is useful to examine the relation between σ^2 and the *idf* factor (logN/df), because the latter is a relative frequency measure of the same kind as σ^2, with a similar goal: identifying terms that help distinguish the documents to which they are assigned from the remainder of the collection. One might think that σ^2 is directly related to *idf* (e.g., the higher *idf*, the higher σ^2), but it turns out that this is not the case. A high value of *idf* will usually produce a low value of σ^2, while a low value of *idf* may be associated to a high as well as to a low value of σ^2, depending on the distribution of terms in the documents in which they are contained. On the other hand, if we considered only the documents that contain the given term, the value of σ^2 would be totally independent of the value of *idf*, irrespective of how we chose to compute the term occurrence. In fact, σ^2 and *idf* are based on different variables (number of occurrences versus binary value of occurrence/non-occurrence) and perform different statistical operations on those variables (variance versus mean). Therefore they seem to capture different patterns of data regularities and can be used together in a comprehensive weighting scheme.

6. Results

In Table 1 we show the performance of four different document rankings with expanded query. The four runs were characterized by the following parameters: title + description, without variance; title + description + narrative, without variance; title + description + narrative, with σ^2_{tf}; title + description + narrative, with $\sigma^2_{norm\text{-}tf}$. In Table 1 we also show the results of document ranking with unexpanded query used as a baseline. The results are reported using the standard TREC performance evaluation measures.

Table 1 shows that the four rankings with expanded query had better results than unexpanded query for virtually all evaluation measures, including precision for the first retrieved documents. Of the four expanded runs, those using T+D+N fared consistently better than the one with only T+D, with small, but not negligible, differences. The introduction of variance in the method employing the full topic description was beneficial, especially as far as average precision was concerned. In particular, we obtained the best average precision result – i.e., 0.3106 - using $\sigma^2_{norm\text{-}tf}$.

As mentioned before, we wanted to test the hypothesis that when using a better baseline retrieval the shift from unexpanded query to expanded query would increase the relative performance improvement. While the performance of ranking with unexpanded query in TREC-8 was higher than double what it was in TREC-7, the relative performance improvement after query expansion in TREC-8 was the same as TREC-7 (+14%). Thus, the relative performance improvement due to expansion was not as high as expected. One possible explanation for this is that when the initial retrieval is really good, it can be hardly improved further upon. The reported results were obtained by averaging over the whole set of topics; a topic by topic analysis might help better understand when and why the relative performance variations are different. The second main goal of our experiments was to test the effectiveness of the use of term variance in the weighting scheme. Our results provide some evidence that it may be a promising directions, but of course this issue needs to be investigated more carefully.

It is also useful to compare the overall performance of our system with that of the other official runs in the Ad-hoc category.

Considering average precision as the measure for performance comparison, our best runs for automatic ad hoc (fub99tt) and for automatic short ad hoc (fub99td) were ranked as the eight and the fourth best system, respectively. A query by query analysis revealed that we achieved better than median performance for 37 topics in automatic ad hoc and for 40 topics in automatic short

ad hoc. In automatic ad hoc we obtained the best performance results for two topics, in automatic short ad hoc for five topics. Finally, it should be noted that the documents retrieved by our best runs could not be included in the document pool used to produce the topic relevance file by the TREC's assessors. Thus, their performance might have been better than actually reported.

Table 1. Comparative performance of ranking with and without query expansion

	unexp. T+D+N	exp. T+D	exp. T+D+N	exp. T+D+N, σ^2_{tf}	exp. T+D+N, $\sigma^2_{norm\text{-}tf}$
Run tag		fub99td	fub99a	fub99tf	fub99tt
Ret&Rel	2938	3298	3262	3281	3299
AV Prec	0.2718	0.3064	0.3068	0.3099	0.3106
11 Point Prec	0.2978	0.3229	0.3245	0.3281	0.3285
R-Prec	0.3168	0.3366	0.3364	0.3398	0.3354
Prec at 5	0.5960	0.5760	0.6080	0.6040	0.6160
Prec at 10	0.4920	0.5100	0.5300	0.5260	0.5300

7. Conclusion

Our experiments show that information-theoretic query expansion may produce excellent results, both on a relative and on an absolute scale. In addition, they seem to imply that the relative performance improvement due to query expansion does not grow monotonically as the quality of the initial baseline retrieval improves. Finally, they also suggest that it may be useful to investigate the use of term variance to refine the weighting schemes employed to weight documents and queries.

Acknowledgments

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

References

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), pp.363-369*, NIST Special Publication 500-242.

Carpineto, C., and Romano, G. (1999). Towards better techniques for automatic query expansion. *Proceedings of the 3th European Conference on Digital Libraries (ECDL'99)*, Paris, France, pp. 126-141, Lecture Notes in Computer Science 1696, Springer.

Karp, D., Schabes, Y., Zaidel, and M., Egedi, D. (1992). A freely available wide coverage morphological analyzer for English. *Proceedings of the 14th International Conference on Computational Linguistics (COLING '92)*, Nantes, France.

Robertson, S. E., Walker, S., and Beaulieu, M. (1999) Okapi at TREC-7: Automatic ad hoc, filtering, VLC, and interactive track. *Proceedings of the seventh Text REtrieval Conference (TREC-7*), pp. 253-264, NIST Special Publication 500-242.

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.