# FUB at TREC 2008 Relevance Feedback Track: Extending Rocchio with Distributional Term Analysis

Andrea Bernardini, Claudio Carpineto
*Fondazione Ugo Bordoni*
*{a.berna, carpinet}@fub.it*

## 1. Introduction

The main goals of our participation in the Relevance feedback track at TREC 2008 were the following.

- Test the effectiveness of using a combination of Rocchio and distributional term analysis on a relevance feedback task; so far, this approach has usually been used (with good results) in a pseudo-relevance setting.

- Test whether and when negative relevance feedback is useful; e.g., is negative relevance feedback most effective when the distribution of terms in the negative documents is different than the distribution in the positive documents?

- Study how the performance of relevance feedback varies as the size of the set of feedback documents grows.

- Check if /how the performance of relevance feedback is influenced by the size of the expanded query.

- Compare relevance feedback to pseudo-relevance feedback; e.g, is relevance feedback not only more effective but also more robust than pseudo-relevance feedback?

## 2. Our approach: combining Rocchio with term-ranking scores

The starting point is the improved version [Salton and Buckley 1990] of the original Rocchio's formula [Rocchio 1971]:

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

Qnew is a weighted term vector for the expanded query, Qorig 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 usually computed by a weighting scheme applied to the whole collection.

This approach is simple and computationally efficient, but it has the disadvantage that each

term weight may reflect more the usefulness of that term with respect to the entire collection rather than its importance with respect to the user query. This issue can be addressed by studying the difference of term distribution between the subsets of relevant (or non-relevant) documents and the whole collection. It is expected that terms with little informative content will have the same distribution in any subset of the collection, whereas the terms that are most closely related to the query will have a comparatively higher probability of occurrence in the relevant (or non-relevant) documents.

The term-ranking scores can then be used not only to select the best expansion terms but also to reweight the expanded query using equation (1). This approach was early suggested in (Carpineto et al. 2001) and it has been adopted in several subsequent studies; e.g., Wong et al. 2008, Perez-Aguera and Araujo 2008.

In this paper we use the Bo1 model in the Bose-Einstein statistics to assign a score to each candidate expansion term. Bo1 evaluate the importance of a term by calculating the divergence of its distribution in a pseudo-relevance document set from a random distribution (Amati 2003). Bo1 estimates the score of term t as follows:

$$score(t) = f_R \log_2[(1+nf_C) / nf_C)] + \log_2(1+nf_C)$$

where $f_R$ is the frequency of the term in the relevant documents, and $nf_C$ is given by the frequency of the term in the collection divided by the number of documents in the collection.

As the document-based weights used for the unexpanded query and the Bo1 scores used for the expansion terms had different scales, they were normalized by the maximum corresponding weight.

## 3. Dealing with negative relevance feedback

A straightforward utilization of negative relevance feedback in equation (1) is possible but it is probably not the most effective choice. If we want to take full advantage of the information about nonrelevant documents, more selective policies for choosing the negative expansion terms are necessary. In particular, we do not want to downweight good positive terms which also happen to occur in negative documents. Our approach was to choose those terms in the negative documents that most contributed to the Bose Einstein divergence of the negative documents from the positive ones. For computational reasons, however, we approximated this criterion by using the divergence of the two set of feedback documents *from the collection,* because these measures can be computed more easily.

We defined two distinct methods, which will be referred to as method 1 and method 2 (in both methods the number of positive terms was set to 100). In method 1, we chose the the first 30 terms in the negative documents that most contributed to divergence from the collection, provided that they did *not* appear in the positive terms. In method 2, we chose the 100 terms in the negative documents with the greatest difference between the divergence

of the negative documents from the collection and the divergence of the positive documents from the collection, regardless of whether or not such terms appeared also in the positive terms.
Interestingly, we found that method 1 yielded nearly the same results as those that would be produced by method 2 with only 30 negative terms. The main difference between method 1 and method 2 was thus the size of the set of negative terms used for expanding the query, i.e., 30 and 100, respectively.

## 4. Experiments

To verify our hypotheses, we extended the relevance feedback module of the software Terrier [Ounis et al. 2006], used as underlying information retrieval system. The basic configuration of Terrier was the following: Pl2 DFR model (parameter c=1.0) for document ranking, and Bo1 for Terrier's default query expansion.

We did not use all the input relevance data because we noticed that the information contained in "relevant" documents (score 1) was not always very accurate. We concentrated only on "not relevant" documents (score 0) and "highly relevant" documents (score 2); "relevant documents" (score 1) were thus discarded (except for those topics with no "highly relevant" documents).

For the parameters in the Rocchio formula (equation 1), we chose a uniform set of values: $\alpha=\beta=\gamma=1$. This choice may have adversely affected the results about the utility of negative relevance feedback because negative terms are usually weighted with lower values than positive terms.

In Table 1 we show the performance, averaged over the set of 31 evaluation topics, of the nine runs submitted by FUB (the first half for method 1, the second for method 2). As a general remark, we would like to note that these values were quite good on an absolute scale, as our best run was ranked in the first positions of the official results of the track. The most important findings shown in Table 1 are that E1 consistently achieved the best results across all evaluation measures, and that the retrieval effectiveness of E1 was roughly twice as high as the baseline (i.e., A1), with a record improvement for P10.

A topic by topic analysis reveals that an increase in retrieval effectiveness was obtained by each run for nearly all topics. In Figure 1 we show the performance of the two runs with most feedback relevance documents (i. e., E1 and E2) on individual topics. For instance, considering E1, its application was detrimental to the retrieval effectiveness for only three topics.

Table 1. Mean performance of submitted runs.

| Run | Map | R-prec | P10 |
|---|---|---|---|
| FubRF08.A1 | 0,1091 | 0,1483 | 0,19 |
| FubRF08.B1 | 0,1803 | 0,2016 | 0,32 |
| FubRF08.C1 | 0,1873 | 0,2169 | 0,32 |
| FubRF08.D1 | 0,2017 | 0,2298 | 0,35 |
| FubRF08.E1 | **0,214** | **0,2463** | **0,44** |
| FubRF08.A2 | 0,1091 | 0,1483 | 0,19 |
| FubRF08.C2 | 0,1752 | 0,2008 | 0,31 |
| FubRF08.D2 | 0,1857 | 0,2174 | 0,34 |
| FubRF08.E2 | 0,173 | 0,2105 | 0,35 |

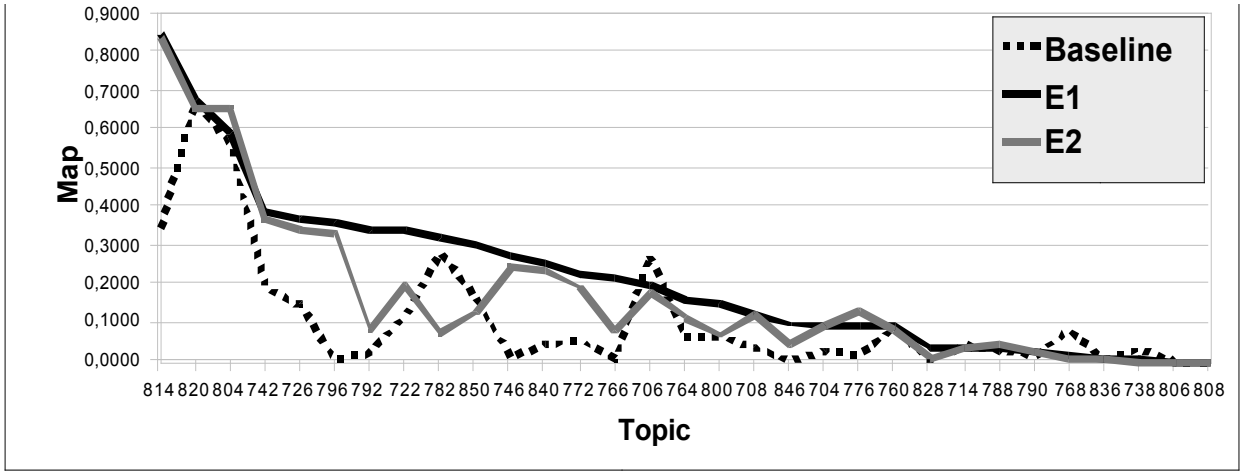


Figure 1. Comparison of E2, E1, and baseline on individual topics (ordered by decreasing value of E1 performance)

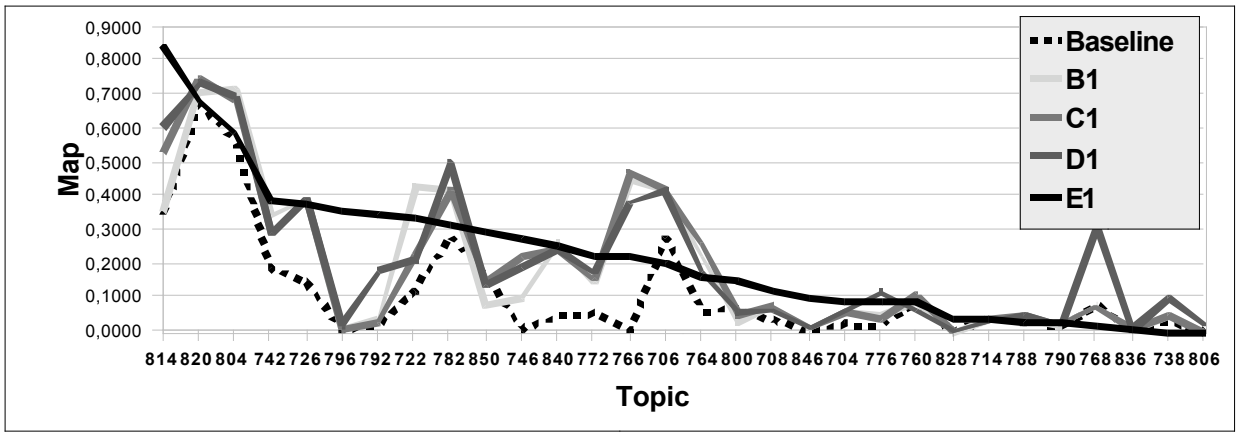


Figure 2. Performance variation of method 1 (with varying amounts of feedback information) on individual topics (ordered by decreasing value of E1 performance).

Recalling that E1 and E2 make use of 30 and 100 negative terms, respectively, and that E1 produced the same terms as those which would be produced by E2 if it were restricted to 30 terms, Figure 1 also shows that increasing the number of negative terms did not result in a performance improvement. In fact, for the overwhelming majority of the topics (28 out of 31), the method with fewer negative terms was better.

In Figure 2 we show how the retrieval effectiveness of method 1 varied as the size of the feedback documents grew. While E1 was, on average, the best method (see Table 1), Figure 2 shows that the relative performance on individual topics differed considerably. For instance, on topic 768 ("women in state legislatures"), E1 performed worse than the baseline while the other runs did better than the baseline, sometimes by a large amount.

A related (somewhat unexpected) observation is that the methods with smaller amounts of feedback seem more robust than the method with the highest amount of feedback, in the sense that the application of the former methods almost never resulted in a decrease of retrieval performance over the baseline.

To gain more insights into the relationship between retrieval performance and amount of feedback information, we ran a more controlled experiment in which we considered how the retrieval performance varies as the number of (highly) *positive* feedback documents increases.

We set the maximum number of feedback documents to 50, because several topics did not have a large number of highly relevant feedback documents. We also let the number of positive expansion terms vary in the range between 0 and 100 (while the number of negative expansion terms was kept fixed at 30). The results are shown in Table 3.

In general, these results show that an increase in the number of documents and/or in the number of terms positively affected the retrieval effectiveness. In particular, the best results were obtained for the largest values of both parameters.

To get a more informative view of this behavior, in Figure 3 we plot the MAP values taken from Table 2 as a function of the number of documents, keeping the number of terms constant.
Dually, in Figure 4 we plot the MAP values as a function of the number of terms, keeping the number of documents constant.

Table 2. Retrieval performance of relevance feedback for various combinations of the highly relevant documents and the number of expansion terms.

| | documents | | | | |
|---|---|---|---|---|---|
| **terms** | **1** | **3** | **10** | **25** | **50** |
| **0** | 0,1091 | 0,1091 | 0,1091 | 0,1091 | 0,1091 |
| **1** | 0,1084 | 0,1088 | 0,1185 | 0,1145 | 0,1132 |
| **5** | 0,1293 | 0,1396 | 0,1535 | 0,1507 | 0,1531 |
| **10** | 0,1330 | 0,1539 | 0,1610 | 0,1697 | 0,1711 |
| **15** | 0,1401 | 0,1567 | 0,1676 | 0,1745 | 0,1811 |
| **20** | 0,1452 | 0,1621 | 0,1722 | 0,1853 | 0,1894 |
| **25** | 0,1513 | 0,1688 | 0,1782 | 0,1924 | 0,1973 |
| **30** | 0,1522 | 0,1684 | 0,1780 | 0,1940 | 0,1996 |
| **35** | 0,1532 | 0,1711 | 0,1801 | 0,1962 | 0,2047 |
| **40** | 0,1570 | 0,1732 | 0,1838 | 0,1991 | 0,2093 |
| **45** | 0,1568 | 0,1765 | 0,1841 | 0,2001 | 0,2101 |
| **50** | 0,1571 | 0,1762 | 0,1849 | 0,2024 | 0,2124 |
| **55** | 0,1577 | 0,1779 | 0,1862 | 0,2025 | 0,2133 |
| **60** | 0,1580 | 0,1800 | 0,1866 | 0,2038 | 0,2125 |
| **65** | 0,1583 | 0,1795 | 0,1868 | 0,2062 | 0,2135 |
| **70** | 0,1579 | 0,1798 | 0,1865 | 0,2060 | 0,2137 |
| **75** | 0,1568 | 0,1788 | 0,1845 | 0,2045 | 0,2121 |
| **80** | 0,1564 | 0,1787 | 0,1853 | 0,2028 | 0,2129 |
| **85** | 0,1562 | 0,1791 | 0,1845 | 0,2026 | 0,2132 |
| **90** | 0,1561 | 0,1791 | 0,1843 | 0,2022 | 0,2136 |
| **95** | 0,1558 | 0,1791 | 0,1842 | 0,2023 | 0,2138 |
| **100** | 0,1553 | 0,1788 | 0,1856 | 0,2025 | 0,2138 |

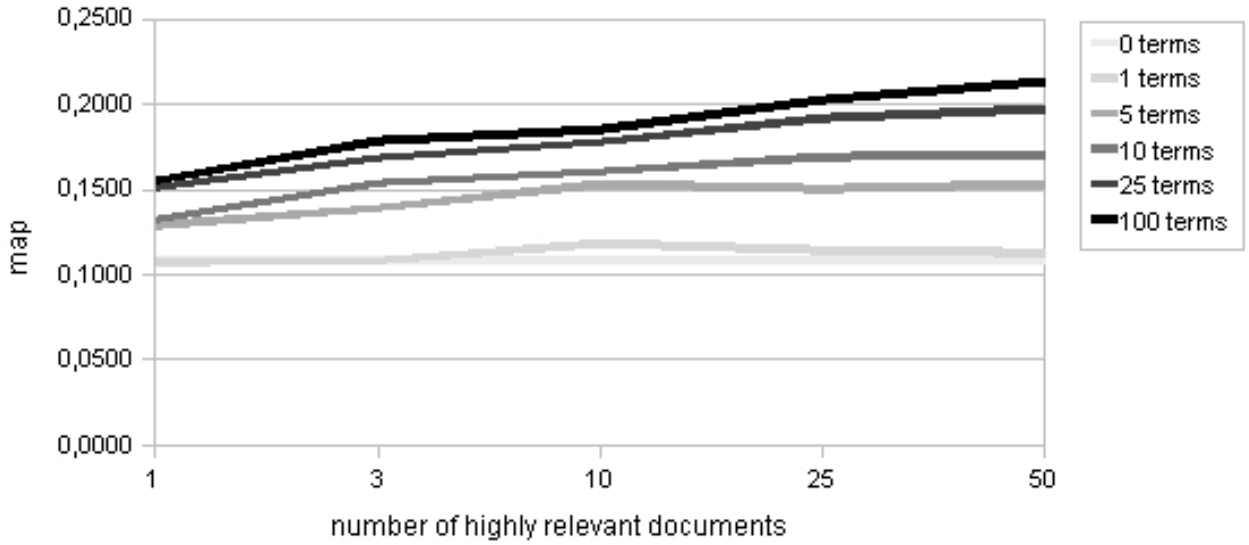


Figure 3. Retrieval performance of relevance feedback as a function of the number of feedback documents, using the number of expansion terms as a parameter.

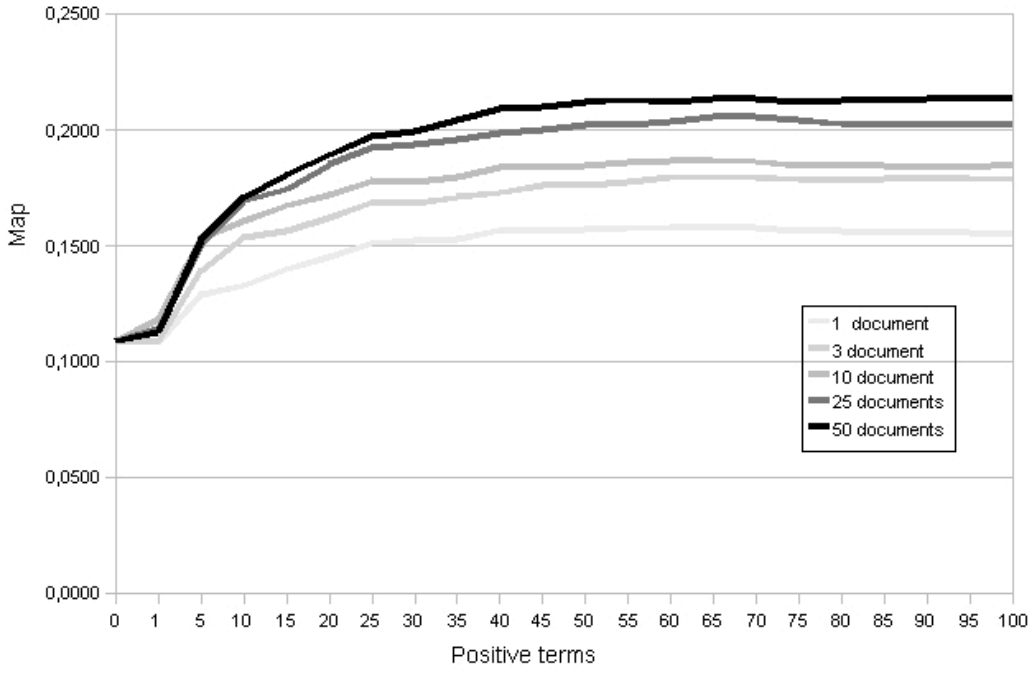


Figure 4. Retrieval performance of relevance feedback as a function of the number of (positive) expansion terms, using the number of feedback documents as a parameter.

Figure 3 and Figure 4 confirm that the retrieval performance grew almost monotonically with respect to each parameter. These results differ from those reported in earlier studies about the low effect of the main relevance feedback parameters on retrieval performance (e.g., Salton and Buckley 1990), while they seem more consistent with some recent findings (Wong et al. 2008).

We then evaluated how relevance feedback compares to pseudo-relevance feedback. In Table 3 we report the MAP values of the two methods, averaged over the 31 topics, for several values of the number of relevant or pseudo-relevant documents (we set the number of expansion terms to 100). Clearly, the benefits of using truly relevant documents (instead of top retrieved documents) are more tangible as their number becomes large, although a marked difference is observable even when we consider only one or three documents.

Table 3. Retrieval performance of relevance feedback versus pseudo-relevance feedback, averaged over the set of topics)

| Documents | Pseudo-Rel Feedback | Relevance Feedback |
|---|---|---|
| 1 | 0.1358 | 0.1553 |
| 3 | 0.1387 | 0.1788 |
| 10 | 0.1680 | 0.1856 |
| 25 | 0.1672 | 0.2025 |
| 50 | 0,1718 | 0.2138 |

Aside from mean retrieval effectiveness, it is interesting to see if the use of truly relevant documents help improve the robustness of query expansion. To this aim, we performed a query-by-query analysis of the relative performance of the two methods, choosing a set of experimental conditions which were more favourable to pseudo-relevance feedback; i.e., 3 documents, 10 expansion terms (in fact, this is a typical parameter setting for automatic query expansion methods). The results are shown in Figure 5.

Apparently, the performance of relevance feedback was not always better than pseudo-relevance feedback. Indeed, there were a few topics where the latter method seemed more effective. Note however that due to the skewed distribution of highly relevant documents in the feedback data (from 0 to 181 documents for a topic), we had to turn to moderately relevant documents when there were not enough highly relevant documents. We checked that the topics with the worst performance of the relevance feedback method were exactly those for which there was a paucity of highly relevant documents. Furthermore, these results were obtained for three documents and ten terms, which was one of the least effective parameter choices for the relevance feedback method (see Table 2).

Finally, we have made some preliminary experiments to evaluate the utility of negative information. We observed a very limited improvement over using just positive

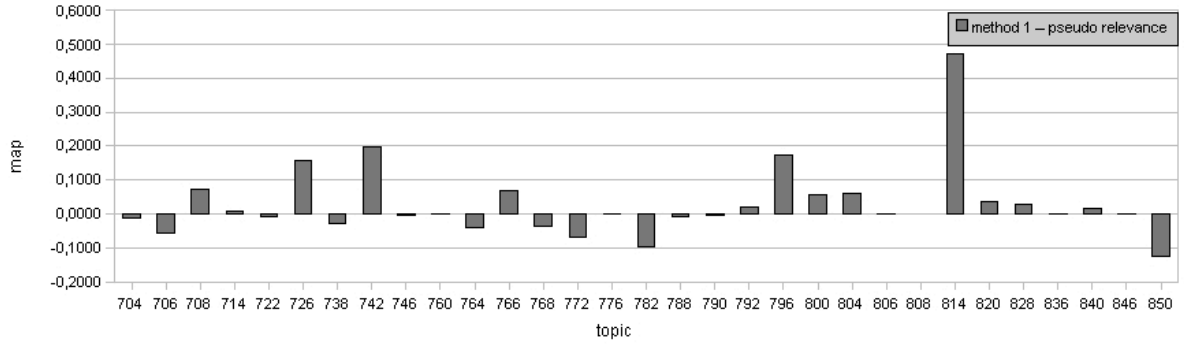


Figure 5. Improvement of relevance feedback (method 1) 1 over pseudo relevance feedback on individual topics (3 relevant documents and 10 expansion terms).

expansion terms, but this may be due to our specific experimental conditions (e.g., ineffective values of $\beta$ and $\gamma$in equation 1, small proportion of negative terms in comparison to the positive ones, etc.). This issue needs more work.

## 5. Conclusions

The main conclusions that can be drawn from our experiments are the following.

1) The use of distribution-based scores within the Rocchio's formula was an effective relevance feedback method.

2) The performance of relevance feedback in general increased as the number of feedback documents and the number of expansion terms grew, even when the two parameters were taken in combination.

3) Other conditions being equal, the use of truly relevant documents resulted in a clear performance improvement over using pseudo-relevance feedback, both in terms of mean retrieval effectiveness and robustness.

**Acknowledgments.**

We would like to thank Gianni Amati for helping with Terrier.

## References


CARPINETO, C., DE MORI, R., ROMANO, G., AND BIGI, B. 2001. An information theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, *19*, 1, 1-27.

AMATI, G. 2003. Probabilistic models for information retrieval based on divergence from randomness. *PhD thesis, University of Glasgow.*

OUNIS, I., AMATI, G., , PLACHOURAS, V., HE, B., MACDONALD, C., AND LIOMA, C. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform._In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*. Seattle, Washington, USA.

PEREZ-AGUERA, J. R., ARAUJO, L. (2008). Comparing and Combining Methods for Automatic Query Expansion. *Advances in Natural Language Processing and Applications. Research in Computing Science,* 33, 177-188.for Automatic Query Expansion. *Advances in Natural Language Processing and Applications. Research in Computing Science,* 33, 177-188.

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

SALTON, G. AND BUCKLEY, C. 1990. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Sciences*, *41*, 288-297.

WONG, W. S., LUK, R. W., LEONG, H. V., HO, K. S., LEE, D. L. (2008). Re-examining the effects of adding relevance information in a relevance feedback environment. *Information Processing and Management*, 44 (3), 1086-1116.