# THUIR at TREC2008: Relevance Feedback Track[1]

Bo Zhou, Qi Fang, Rongwei Cen, Min Zhang, Yiqun Liu, Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China
Tsinghua National Laboratory for Information Science and Technology, Tsinghua University
b-zhou07@mails.tsinghua.edu.cn

**Abstract.** Our group has participated into Relevance Feedback (RF) Track in TREC2008. In our experiments, two kinds of techniques, query expansion and search result re-ranking based on document relevance model, are adopted to improve the retrieval performance. The TMiner search engine, from IR group of Tsinghua University, is used as our text retrieval system.

## 1. Introduction

Tsinghua University Information Retrieval Group (THUIR) has participated into the first Relevance Feedback Track of TREC2008. The TMiner search engine has been used as our text retrieval system, because the processing capability and flexibility of this system on large text data has been testified during many years' Web Track and Terabyte Track. In the track, we studied two approaches: 1) query expansion, 2) search result re-ranking based on document relevance model.

***Query Expansion*:** Terms in the annotated documents (feedback) are used to expand the original query; the new born queries are sent to the search engine for further information retrieval; users get the documents retrieved by the expanded queries.

***Search Result Re-ranking*:** The relevance between the annotated documents and other documents are used to influence the search results; users finally get the re-ranked document list. In detail, we have experimented two different methods on which search result re-ranking based: a) Clustering; b) Documents Relevance Model.

The rest of this paper is structured as follows. After the introduction of Query Expansion approach in Section 2, Search Result Re-ranking is discussed in Section 3. The evaluation results of the submitted runs are illustrated in Section 4. The last section contains summaries and outlines promising future work.

## 2. Query Expansion

In Query Expansion approach, Terms in the annotated documents (feedback) are used to expand the original query; the new born queries are sent to the search engine for further information retrieval. Two-phase process has been adopted: a) Term selection with feedback documents; b) Search Result Integration.

### 2.1 Term Selection

We selected expansion terms using local context analysis method. Assume *the query to be expanded is Q*, *the query terms in Q are* $q_1, q_2...q_m$, *the collection being searched is C* and *the set of relevant documents is S.* Our approach prefers terms of higher co-occurrence with query terms. Specifically, we will derive a function $f(t, Q)$ which measures how good a term $t$ is for expanding based on t's co-occurrence with $q_i$ in $S$. All terms are ranked by $f$, and the best $k$ terms are selected. We have adopted the following formula [1] to measure the degree of co-occurrence of t with $q_i$:

[1] Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60621062, 60503064, 60736044) and National 863 High Technology Project (2006AA01Z141.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **NOV 2008** | | **00-00-2008 to 00-00-2008** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **THUIR at TREC2008: Relevance Feedback Track** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Tsinghua University,State Key Laboratory of Intelligent Technology and Systems,Beijing 100084, China,** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES
**Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).**

14. ABSTRACT
**see report**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT **unclassified** | b. ABSTRACT **unclassified** | c. THIS PAGE **unclassified** | **Same as Report (SAR)** | **5** | |

$$co\_degree(t, q_i) = \frac{1}{|S|}\sum_{d \in S} \log\,(tf(t,d)+1) \times \log\,(tf(q_i,d)+1)$$

where $tf(t,d)$and $tf(q_i,d)$are the frequencies of t and $q_i$ in document $d$, respectively.
To combine the degrees of co-occurrence with all query terms, the following function is used:

$$f(t,Q) = \prod_{q_i \in Q} (\delta + co\_degreee(t,q_i))^{idf\,(t)idf\,(q_i)}$$

where $idf(t)$ is the inverse document frequency of t in the whole collection C. so is $idf(q_i)$.

### 2.2 Search Result Integration

Intuitively, terms selected from the relevant documents should be treated separately with documents selected from irrelevant documents. We first tried the following formula, in which terms from different relevant documents are treated as if they are from one document. This is because we suppose relevant documents are prone to be concept focused, and on the contrary the irrelevant documents may share many unique concepts respectively.

$$sim(Q,D) = sim(Q,D) + \alpha * sim\left(\sum_i wi * Q^i_{pos}\,, D\right) - \beta * \sum_i sim(D^i_{neg}\,, D)$$

However, through our training experiments, the irrelevant part $-\beta * \sum_i sim(D^i_{neg}\,, D)$ doesn't help, so it was finally abandoned. As a result, the actual formula is adopted.

$$\mathrm{sim(Q,D)} = \mathrm{sim(Q,D)} + \alpha * \mathrm{sim}\left(\sum_{\mathrm{i}} \mathrm{wi} * \mathrm{Q^i_{pos}}\,, \mathrm{D}\right)$$

Figure 1 shows the MAP changes with the number of the selected terms. Different curves are the results with different combination parameter $\alpha$.

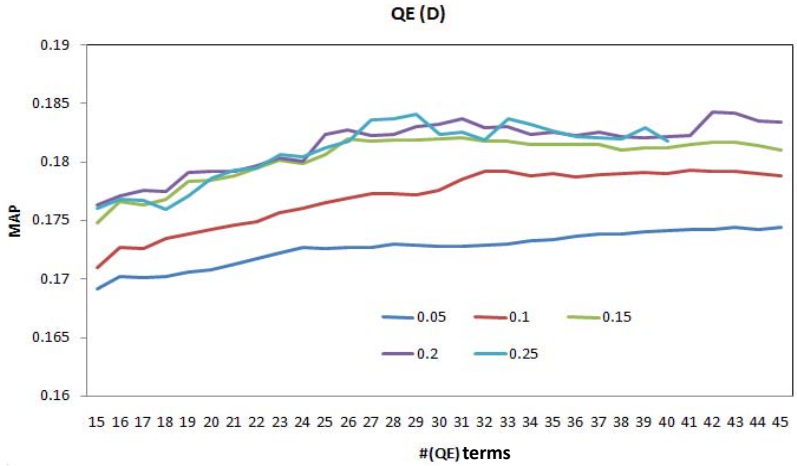


Figure 1: MAP with the increase of the selected terms under different combining parameter: α

## 3. Search Result Re-ranking

We have studied two re-ranking methods to improve the retrieval performance using relevance feedback information: clustering and document relevance model.

### 3.1 Clustering

Under the assumption that relevant documents share several concepts (needs behind the query), our intuitions are: by appropriate clustering algorithms, 1) relevant and irrelevant documents

would be clustered into several different clusters (R-cluster and NR-cluster) separately; 2) There would be more NR-clusters than R-clusters. If several feedback relevant documents are in one cluster, there is high probability that a) it is a R-cluster, b) other documents in the cluster are also relevant. The known irrelevant documents, on the contrary, can be used to determine NR-clusters. Unfortunately, the experimental result is disappointing. Although we used all clustering methods of CLUTO [2], relevant documents cannot be clustered with high purity. Figure 2 shows the clustering results of bagglo-wclink-10[2] parameter.

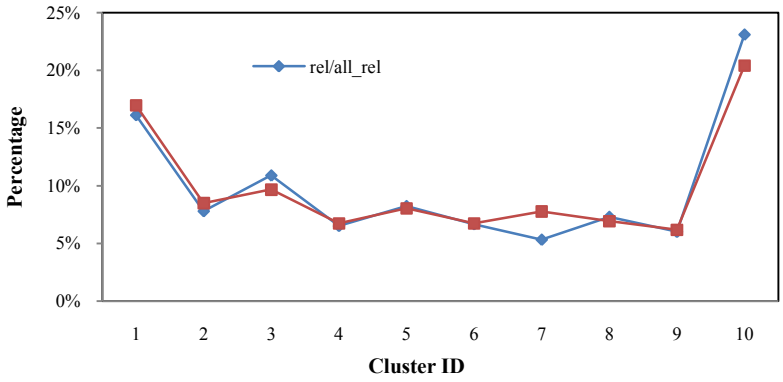


Figure 2: The clustering result using CLUTO

The "rel/all_rel" series means the number of relevant docs divide the number of all relevant docs in the cluster; the "nrel/all_nrel" series means the number of irrelevant docs divide the number of all irrelevant docs in each cluster. We can discover that it's not easy to cluster relevant and irrelevant documents into separate clusters by direct using clustering algorithms.

### 3.2 Document Relevance Model

Under the assumption that relevant documents are more relevant to the annotated relevant documents and less relevant to the annotated irrelevant documents, we suppose that the relevance measure between annotated documents and other documents is useful for determining the relevance between query and documents. In this approach, the original search result is reranked based on the relevance measures between the annotated documents (feedback) and the other documents in the search result list.

The score function for one document is as follows.

$$Score(d_i) = \alpha * \boldsymbol{norm}\{\boldsymbol{max}[\boldsymbol{dis}(d_i, D_{rel})] - \boldsymbol{max}[\boldsymbol{dis}(d_i, D_{nrel})]\} - (1-\alpha) * \boldsymbol{norm}(similarity),\ d_i \epsilon D_{search} Score(d_i)$$

$D_{search}$ is the set of originally retrieved documents; $D_{rel}$ is the feedback relevant documents set; $D_{nrel}$ is the feedback irrelevant documents set; *dis*(…) represents the cosine distance between documents; *similarity* means the original score of one document; *norm*(…) means normalization. The training results (on non- submission topics) are shown in the following table.

Table 1: MAP improvement on training set

| RF Set | RF doc count | Baseline MAP | α | Improved MAP | Imp |
|---|---|---|---|---|---|
| B | 1 | 0.2940 | 0.35 | 0.3268 | +11.16% |
| C | 6 | 0.2860 | 0.32 | 0.3247 | +13.53% |
| D | 10 | 0.2805 | 0.33 | 0.3307 | +17.90% |
| E | 246 | 0.2122 | 0.23 | 0.2889 | +36.15% |

From Table 1, we can see that the MAP is steadily increasing with the increase of the number of

RF documents. However, this isn't discovered in the evaluation results of our submitted runs, which is illustrated in Section 5.

## 4. Evaluation Results of Submitted Runs

Table 2 illustrates the evaluation results of Query Expansion runs. Consistent improvement has been shown with the increase of RF information from RF set B to D. However, the improvement decreases for RF set E. This is possibly caused by the irrelevant documents in RF-set E. Further analysis will be made when the *qrels* is given.

Table 2: Evaluation results of Query Expansion runs

| runs | map | Imp+ | bpref | Imp+ | R-prec | Imp+ | Mtc | Imp+ |
|---|---|---|---|---|---|---|---|---|
| A | 0.1357 | 0 | 0.1970 | 0 | 0.1571 | 0 | 0.0483 | 0 |
| B1 | 0.1498 | 10.4% | 0.2108 | 7.0% | 0.1703 | 8.4% | 0.0573 | 18.5% |
| C1 | 0.1568 | 15.5% | 0.2233 | 13.4% | 0.1802 | 14.7% | 0.0595 | 23.1% |
| D1 | **0.1646** | **21.3%** | **0.2319** | **17.8%** | **0.1939** | **23.4%** | 0.0596 | 23.4% |
| E1 | 0.1568 | 15.5% | 0.2309 | 17.2% | 0.1899 | 20.9% | **0.0606** | **25.5%** |

Table 3 shows the evaluation of Search Result Re-ranking runs. Best MAP improvement is achieved On RF-set E. Further investigation will be given when *qrels* is given.

Table 3: Evaluation results of Search Result Re-ranking runs

| runs | map | Imp+ | bpref | Imp+ | R-prec | Imp+ | mtc | Imp+ |
|---|---|---|---|---|---|---|---|---|
| A | 0.1357 | 0 | 0.1970 | 0 | 0.1571 | 0 | 0.0483 | 0 |
| B2 | 0.1684 | 24.1% | 0.2262 | 14.8% | 0.1893 | 20.5% | **0.0642** | **32.9%** |
| C2 | 0.1525 | 12.4% | 0.2169 | 10.1% | 0.1724 | 9.8% | 0.0595 | 23.2% |
| D2 | 0.1607 | 18.4% | 0.2262 | 14.8% | 0.1859 | 18.4% | 0.0598 | 23.9% |
| E2 | **0.1965** | **44.8%** | **0.2866** | **45.5%** | **0.2348** | **49.5%** | 0.0614 | 27.1% |

Table 4 shows the evaluation results of all runs.

Table 4: Evaluation results of all runs

| runs | map | Imp+ | R-prec | Imp+ | mtc | Imp+ |
|---|---|---|---|---|---|---|
| A1 | 0.1357 | 0 | 0.1571 | 0 | 0.0483 | 0 |
| B1 | 0.1498 | 10.4% | 0.1703 | 8.4% | 0.0573 | 18.5% |
| B2 | 0.1684 | 24.1% | 0.1893 | 20.5% | **0.0642** | **32.9%** |
| C1 | 0.1568 | 15.5% | 0.1802 | 14.7% | 0.0595 | 23.1% |
| C2 | 0.1525 | 12.4% | 0.1724 | 9.8% | 0.0595 | 23.2% |
| D1 | **0.1646** | **21.3%** | **0.1939** | **23.4%** | 0.0596 | 23.4% |
| D2 | 0.1607 | 18.4% | 0.1859 | 18.4% | 0.0598 | 23.9% |
| E1 | 0.1568 | 15.5% | 0.1899 | 20.9% | **0.0606** | **25.5%** |
| E2 | **0.1965** | **44.8%** | **0.2348** | **49.5%** | 0.0614 | 27.1% |
| Best | 0.4215 | | 0.4530 | | 0.0868 | |
| Median | 0.1427 | | 0.1801 | | 0.0564 | |
| Worst | 0.0008 | | 0.0040 | | 0.0057 | |

Figure 3 gives the per-topic analysis on the greatest MAP improvement of Query Expansion approach (D1-A1) and Search Result Re-ranking approach (E2-A1). Both approaches are effective on majority of topics. And Search Result Re-ranking approach is slightly better than the Query

Expansion approach.

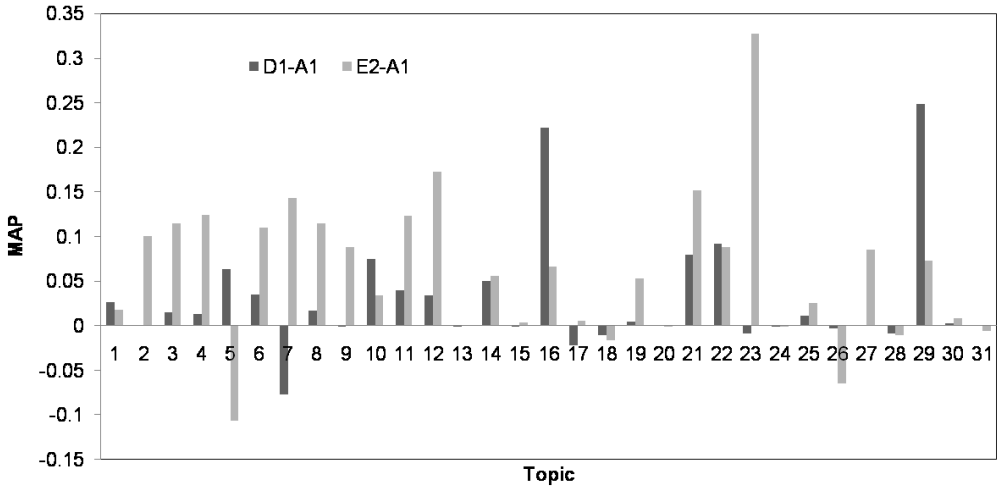


Figure 3: Improvement of D1 vs. E2 on every topic

## 5. Discussion and Future Work

In the Relevance Feedback Track, we studied different approaches to improving retrieval performance using feedback information. Both the Query Expansion approach and the Search Result Re-ranking approach are effective for MAP improvement on majority of topics.

In the future, further investigation and comparative studied will be made on irrelevant information in different feedback sets.

## Reference

[1] J. Xu, W.B. Croft, "Improving the Effectiveness of Information Retrieval with Local Context Analysis", ACM Transactions on Information Systems, 2000, 18(1):79-112.

[2] http://glaros.dtc.umn.edu/gkhome/views/cluto