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
RMIT University at TREC 2008: Enterprise 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) **RMIT University,School of Computer Science and IT,GPO Box 2476V,Melbourne 3001, Australia,**	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

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**2**	

RMIT University at TREC 2008: Enterprise Track

Mingfang Wu Falk Scholer Steven Garcia

School of Computer Science and IT
RMIT University, GPO Box 2476V
Melbourne 3001, Australia

1 Introduction

RMIT participated in the 2008 Enterprise Track document search task. Our experiments investigated the use of local outdegree, and whether this can improve the ranking quality of a search result list.

Unlike global outdegree, which counts the number of out-links of a page that point to any other pages in a collection, local outdegree only counts the out-links that point to pages contained in a search result list. Intuitively, restricting the outdegree to the result set of a query transforms this source of evidence from something general into a topically-focused source of information, and may help to reduce the problem of topic shift.

For our experiments, we used the Zettair search engine[1] to index and search the CSIRO collection used for the 2008 Enterprise Track. This collection is a crawl of the the public-facing web of the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO) in 2007 (Bailey et al., 2007). Document weights were calculated using the Okapi BM25 similarity function (Sparck Jones et al., 2000), with query words being terms from the `query` fields of the track topics. During indexing and search, words are stemmed and stopped.[2]

2 Description Of Runs

We submitted four runs to the 2008 Enterprise Track: a baseline, two variants using local outdegree, and a pseudo relevance feedback approach:

- **RmitDocQ**: Baseline run.
- **RmitDQComLO:** The top 1000 retrieved documents from RmitDocQ are re-ranked based on a linear combination of document weight and local outdegree:

$$weight = \alpha \cdot similarity + (1 - \alpha)\, outdegree$$

- **RmitDocQRerank:** The top 100 retrieved documents from RmitDocQ run are re-ranked using local outdegree.

[1] Zettair is available under a BSD License from: `http://www.seg.rmit.edu.au/zettair`
[2] The stoplist used is available from: `http://www.csse.unimelb.edu.au/~jz/resources/stopping.zip`

- **RmitDQExp** For each query, the top 10 retrieved documents from the run RmitDocQ are treated as "relevant" documents. Terms in this selected set of documents are weighted according to the following term selection value:

$$TSV = w^{(1)} \times \frac{r}{R}$$

The weight $w^{(1)}$ is the Robertson/Spark Jones weighting function; r is number of selected documents which contain a term, and R is the number of documents in the set. The top 10 terms are then selected and combined with the original query terms to form a new query, with the original query terms being up-weighted by a factor of 3.

3 Results

As shown below, on average our runs proved unsuccessful. From an initially effective baseline, all techniques reduced the mean inferred average precision and mean inferred NDCG of the run.

Run	Mean infAP	Mean infNDCG
TREC median	0.2670	0.4679
RmitDocQ	0.2975	0.5040
RmitDQCombLO	0.2837	0.4970
RmitDQRerank	0.2644	0.4810
RmitDQExp	0.2640	0.4399

However, when considering individual query performance, the results are varied. Each approach improved some topics, while hindering the performance of others. The RmitDQComLO run, which combined local outdegree with document weight, performed best with a positive effect on the average precision of 26 topics and a negative effect on 31 topics. RmitDQRerank, which placed more emphasis on local outdegree, had a less noticeable effect, with an improvement in average precision for only 12 topics, and a negative outcome for 20 topics. Surprisingly, our query expansion run RmitDQExp resulted in the worst performance, with 43 topics decreasing in average precision, and only 17 increasing.

The results suggest that, like many other proposed query evaluation improvement techniques, there are potential gains to be achieved. However, understanding when to apply local outdegree factors to query evaluation, and how to best utilize the information, remain an open question.

References

Bailey, P., Craswell, N., de Vries, A. P. and Soboroff, I. (2007), Overview of the TREC 2007 enterprise track, *in* 'The Sixteenth Text REtrieval Conference (TREC 2007)', National Institute of Standards and Technology Special Publication 500-274, Gaithersburg, MD.

Sparck Jones, K., Walker, S. and Robertson, S. E. (2000), 'A probabilistic model of information retrieval: development and comparative experiments. Part 1', *Information Processing and Management* **36**(6), 779–808.