# RMIT University at TREC 2006: Terabyte Track

Steven Garcia Nicholas Lester
Falk Scholer Milad Shokouhi

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

## 1 Introduction

The TREC 2006 terabyte track consisted of three tasks: informational (or *ad hoc*) search, named page finding, and efficient retrieval. This paper outlines RMIT University's participation in these tasks.

## 2 The Zettair 0.9 Search Engine

`Zettair` is a publicly available retrieval engine developed by the Search Engine Group at RMIT University, available under a BSD license from `http://www.seg.rmit.edu.au/zettair`. Since TREC 2005, Zettair has been significantly revised. Major changes between the latest (0.9) and the previous (0.6) versions are:

- The ability use either a Porter's stemmer (Porter, 1997), a "light" stemmer which achieves slightly worse effectiveness than Porter's with far better efficiency, or no stemming.
- Transparent indexing of gzip files.
- Similarity metric configuration: an enhanced set of similarity metrics, including Dirichlet-smoothed language modelling (Zhai and Lafferty, 2004), Okapi BM25 (Sparck Jones et al., 2000), cosine and pivoted cosine (Singhal et al., 1996) metrics. In addition, Zettair now defines a simple language for implementing new metrics.
- Impact ordered operation: Zettair can now generate and use impact-ordered (Anh and Moffat, 2002) inverted lists.
- Options to control memory usage, resulting in greatly reduced memory consumption, especially during indexing.
- Increased indexing speed.
- Integrated `trec_eval` effectiveness evaluation.

Zettair 0.9 was used for all experiments submitted to TREC 2006 by RMIT, with the exception of the Terabyte track comparative runs.

## 3 *Ad Hoc* Task

This year 50 new topics were created for *ad hoc* searching over the `.GOV2` collection, a 426GB crawl of the `.gov` domain of the World Wide Web. The RMIT *ad hoc* runs can be grouped into three classes: first, representative probabilistic and language modelling retrieval runs, as implemented in our search engine, `Zettair`; second, two data-fusion runs that combine evidence from the output of two search engines, `Indri` and `Zettair`; and third, a manual run.

| Run | Retrieval approach | MAP | P@10 | P@20 | bpref | R-P |
|---|---|---|---|---|---|---|
| Mean | | 0.284 | 0.548 | – | – | – |
| ZETABM | Okapi BM25 | 0.241 | 0.498 | 0.451 | 0.328 | 0.304 |
| ZETADIR | Dirichlet | 0.305 | 0.532 | 0.480 | 0.372 | 0.357 |
| ZETAMERG | Round-robin | 0.308 | 0.502 | 0.488 | 0.363 | 0.353 |
| ZETAMERG2 | CombSum | 0.299 | 0.548 | 0.501 | 0.355 | 0.350 |
| ZETAMAN | Manual | 0.287 | 0.592 | 0.529 | 0.397 | 0.327 |

Table 1: *Effectiveness of terabyte* ad hoc *task runs based on mean average precision (MAP), precision at 10 and 20 documents returned (P@10, P@20), preference relation based on binary relevance judgements (bpref), and precision at the number of relevant documents for each query (R-P). Metrics are calculated over TREC topics 801–850, with relevance judgments constructed using a pool of depth 50.*

## 3.1 Probabilistic and Language Model Runs

Two runs were submitted to establish baseline performance for representative probabilistic and language model retrieval runs, as implemented in the new version of `Zettair`. Run ZETABM used the Okapi BM25 similarity function (Robertson and Walker, 1999), while ZETADIR used a language model with Dirichlet smoothing (Zhai and Lafferty, 2004); the parameters in both models were set to recommended published values.

## 3.2 Data-fusion Runs

Our first data-fusion run, ZETAMERG, is based on a simple round-robin method. The results returned by two search engines — `Zettair` and `Indri` — are sequentially merged according to their ranks. The final merged list consists of the first-rank documents returned by both search engines, followed by the second-rank documents returned by them and so forth. Duplicates are removed, so that documents that are already merged into the final list from one search engine are not added twice.

`Zettair` results used the Okapi BM25 (Robertson and Walker, 1999) similarity measure. `Indri`, available from `www.lemurproject.org/indri`, uses a retrieval model based on a combination of language modelling (Ponte and Croft, 1998) and inference networks (Turtle and Croft, 1991).

The second data-fusion run, ZETAMERG2, merges documents returned by search engines according to their ranks. It can be considered as a variation of the CombSum data-fusion algorithm (Fox and Shaw, 1993) that calculates the final score of documents as follows:

$$S_d = (1000 - R_d(Zettair)) + (1000 - R_d(Indri)) \quad (1)$$

where $R_d(Zettair)$ and $R_d(Indri)$ represent the rank of a given document $d$ in the results returned by `Zettair` and `Indri`, respectively. Each system returns a maximum of 1000 documents per query. If a document appears in one result list but not in the other, its rank in the second list is set to 1000. Based on a voting principle, documents that are ranked highly by both search engines are more likely to be relevant. This data-fusion technique places such documents towards the top of the final merged list.

## 3.3 Manual

For the 2006 Terabyte track, it was requested that participants submit a manual run. For the run ZETAMAN, topics were created by starting with the <`title`> field of the TREC topics, and manually selecting promising keywords from the <`description`> and <`narrative`> fields. In particular, words in the latter two fields that seemed to form commonly-used phrases were treated as phrase queries, while keywords in the topics that described features that were described as *not* being relevant were not included for the manual queries.

| Characteristic | Zettair | All runs submitted | | |
|---|---|---|---|---|
| | | Maximum | Median | Minimum |
| Percentage of document collection indexed | 100 | 100 | 100 | 100 |
| Indexing time in minutes | 569 | 4700 | 369 | 127 |
| Average time to return top 20 documents in seconds | 2.2 | 4.6 | 0.152 | 0.0125 |
| Total wall-clock time for all topics in seconds | 220210 | 463000 | 15240 | 539 |
| Total number of CPUs in system | 2 | 16 | 1 | 1 |
| Total amount of RAM in system in GB | 4 | 16 | 2 | 1 |
| Size of on-disk file structures in GB | 15.8 | 800 | 15 | 1.1 |
| Year of system purchase | 2005 | 2006 | 2005 | 2003 |
| Estimated hardware cost in US dollars | 4000 | 10000 | 2000 | 500 |

Table 2: *Details of the `Zettair` run submitted for the efficiency task.*

### 3.4 Results

Results for our *ad hoc* runs are shown in Table 1. Metrics are calculated over the 50 new topics (801–850), using relevance judgements constructed using a pool depth of 50. The row labelled Mean shows the average of the median MAP and P10 results of all participants.

The relatively poor performance of Okapi BM25 (ZETABM) is surprising. The Dirichlet smoothed language modelling approach (ZETADIR) and data-fusion techniques produce the highest MAP values. The manual run (ZETAMAN) resulted in the highest P10.

## 4 Efficiency Task

We submitted a single run for the efficiency task this year. To reduce the size of the index we removed word offsets from the postings lists. This reduced the index size from 43.0 Gb with offsets to 15.8 Gb without offsets. No threading or parallel processing was used, as Zettair processes queries serially in a single stream. Results were ranked using Dirichlet smoothed language modelling, and both stopping and light stemming were used.

### Results

Table 2 gives a summary of our submitted `Zettair` run, as well as an overview of the range of systems that were used by other participants.

## 5 Named Page Finding Task

This year we investigated the combination of different features for named page finding: the full text of documents, the text in document titles, and inlink anchor text. Each of these three features were indexed separately using the Zettair search engine. A named page query was then run against each index separately, and the results were merged using a simple liner interpolation approach to give a final ranked list. To combine evidence from two features (for example full text and anchor text):

$$Sim = \alpha Sim_A + (1 - \alpha) Sim_B$$

where $0 \leq \alpha \leq 1$. To add evidence from a third feature:

$$Sim = \alpha Sim_A + \beta Sim_B + (1 - (\alpha + \beta)) Sim_C$$

where $0 \leq \alpha, \beta, \alpha + \beta \leq 1$. The values for the interpolation parameters $\alpha$ and $\beta$ were set empirically, using last year's named page finding topics and relevance judgements (topics NP601–872) to determine regions of good performance.

We submitted four runs for the 2006 task:

| Run | Features used | MRR | Rank 1 | Top 10 | Not found |
|---|---|---|---|---|---|
| Mean | | 0.371 | – | – | – |
| ZETNPBM | Document text | 0.326 | 25.4% | 46.4% | 19.3% |
| ZETNPFA | Document text, anchor text | 0.319 | 21.5% | 54.7% | 19.3% |
| ZETNPFT | Document text, titles | 0.389 | 30.4% | 54.7% | 19.3% |
| ZETNPFTA | Document text, titles, anchor text | 0.388 | 28.7% | 59.7% | 19.3% |

Table 3: *Effectiveness of named page finding runs, showing mean reciprocal rank (MRR), the percentage of queries for which an answer was found at rank 1 and within the top 10 documents, and the percentage of queries for which no answer was found. The mean MRR result is calculated over the median MRR of all queries submitted by all participants.*

- zetnpbm: a baseline run using full text only
- zetnpfa: full text and inlink anchor text ($\alpha = 0.8$)
- zetnpft: full text and title text ($\alpha = 0.85$)
- zetnpfta: full text, title text and inlink anchor text ($\alpha = 0.8, \beta = 0.1$)

The similarity function used for each run was Okapi BM25.

### Results

The results for the named page finding task are shown in Table 3. The row labelled Mean shows the average of the median MRR results of all participants in this year's named page finding task.

Our baseline run (ZETNPBM) used full text only (and performed worse than the mean of all runs submitted for this task). Adding inlink anchor text (ZETNPFA) caused a slight decrease in MRR, but the difference is not statistically significant based on a bootstrap test statistic for paired data (Sakai, 2006). Using full text, combined with title text as a separate feature (ZETNPFT), leads to a relative increase in MRR of 19.3% ($p < 0.01$). Combining all three features (ZETNPFTA) leads to an equivalent increase in MRR ($p < 0.01$) over the full text baseline. Combining all three features leads to the highest proportion of queries for which the named resource was found in the top 10 answers, a 28.7% relative increase over the baseline. We therefore conclude that it is beneficial to include topic data as a separate source of evidence.

### Acknowledgements

This work is supported by the Australian Research Council. Hardware for some experiments was provided with the support of an RMIT University VRII grant.

## References

Anh, V. and Moffat, A. (2002), Impact transformation: effective and efficient web retrieval, *in* M. Beaulieu, R. Baeza-Yates, S. Myaeng and K. Jävelin, eds, "Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval", Tampere, Finland, pp. 3–10.

Fox, E. and Shaw, J. (1993), Combination of multiple searches, *in* "Proceedings of the second Text REtrieval Conference", NIST Special Publication 500-215, Gaithersburg, MD, pp. 243–252.

Ponte, J. M. and Croft, W. B. (1998), A language modelling approach to information retrieval, *in* W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson and J. Zobel, eds, "Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval", Melbourne, Australia, pp. 275–281.

Porter, M. (1997), "An algorithm for suffix stripping", pp. 313–316.

Robertson, S. E. and Walker, S. (1999), Okapi/Keenbow at TREC-8, *in* E. M. Voorhees and D. K. Harman, eds, "The Eighth Text REtrieval Conference (TREC-8)", National Institute of Standards and Technology Special Publication 500-246, Gaithersburg, MD, pp. 151–162.

Sakai, T. (2006), Evaluating evaluation metrics based on the bootstrap, *in* "Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval", Seattle, WA, pp. 525–532.

Singhal, A., Buckley, C. and Mitra, M. (1996), Pivoted document length normalization, *in* "Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval", pp. 21–29.

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

Turtle, H. R. and Croft, W. B. (1991), "Evaluation of an inference network-based retrieval model", *ACM Transactions on Information Systems* **9**(3), 187–222.

Zhai, C. and Lafferty, J. (2004), "A study of smoothing methods for language models applied to information retrieval", **22**(2), 179–214.