# PLIERS AT TREC 2002

A MacFarlane
Centre for Interactive Systems Research, Department of Information Science,
City University, Northampton Square, LONDON EC1V 0HB, UK

**Abstract**: We describe our experiments with the .GOV collection in both the topic distillation and named page tasks at the 2002 TREC web track. We report on our indexing speed, retrieval efficiency results and effectiveness results for both tasks.

## 1. Introduction

We report on our experiments for the TREC 2002 web track for both the topic distillation and named page tasks. We use a very simple method for both tasks which takes the first hit page in the top 10 for a give web site and discards any further pages from that web site (section 2 describes our research aims and objectives in more detail). We also describe indexing results (section 3), give a description of the runs and settings used (section 4), briefly describe our retrieval efficiency results in section 5, and outline our retrieval efficiency results in sections 6 and 7. A conclusion is given in section 8.

## 2. Research aims and objectives

We take a very simple approach both the topic distillation and named page tasks. We want to test the hypothesis "does the best BM25 ranked document from any given web site yield the best web page for users information needs". We want to compare this rather simple technique with other more complex techniques which use link information in order to find the best given web page or pages.

Our retrieval efficiency experiments differ from our previous work [2] which concentrated on using large scale parallelism to speed up the processing of both indexing and search. In these experiments we want to show that we can successfully process large amounts of text with our system using a single machine (even if it does has multiple processors on it).

## 3. Indexing methodology and results

### *3.1 Indexing methodology*

We used a simple and straightforward methodology for indexing: parsing, remove stop words, stemming in the given language. The PLIERS HTML/SGML parser needed to be altered to detect non-ASCII characters such as those with umlauts, accents, circumflexes etc. We also incorporated non-English stemmers into the PLIERS library (these were not used for these experiments). We used a standard stop word list defined by Fox [3]. Apart from this our indexing methodology is much the same a described in [2].

### *3.2 Indexing results*

| Elapsed Time (hrs) | Dictionary file size MB | Postings file size GB | Map file size MB | % of text |
|---|---|---|---|---|
| 10.54 | 110 | 1.17 | 40.4 | 7% |

Table 1 – Indexing results for .GOV collection

Table 1 gives the indexing results for the .GOV collection. PLIERS was able to process the data in a reasonable time (just under 11 hours) and produced an inverted file that was only 7% of the collection size. This compares favourably with our previous web track experiments with WT100g [2], in which indexes were 11% of the collection size. The final merge took only about 10 minutes (a total of 1.5% of total indexing time): this represents a significant improvement on previous single processor experiments. This can be explained by our usage of a significantly faster machine. We regard it as a success to be able to index data of this size: we suspect that the system would not be able to handle a slightly larger collection without failing.

## 4. Run descriptions and settings used

All experiments were conducted on a Pentium 4 machine with 256 MB of memory and 240 GB of disk space. The operating system used was Red Hat Linux 7.2. All search runs were done using the Robertson/Sparck Jones Probabilistic model. All our runs are in the Web track. All queries derived from topics are automatic.

Changes to software in order to conduct these particular experiments were minimal. We used the URL/TREC ID list supplied with the .GOV collection to identify and eliminate documents from the top 10 results which are from the same web site. Only the highest ranked document from a web site is retained. The top 10 results are therefore guaranteed to have unique URL's in them i.e. all documents in the top 10 are from different web sites. We used this technique on both Web track tasks.

The weighting function used for these experiments was BM25 [1]. There are a number of tuning constants for this function with which we have done experiments on before, in order to find the best combination for search [2]. There are two constants: **K1** and **B** [1]. The **K1** constant alters the influence of term frequency in the BM25 function, while the **B** constant alters the influence of normalised average document length. Values of **K1** can range from 0 to infinity, whereas the values of **B** are with the range 1 (document lengths used unaltered) to 0 (document length data not used at all). Table 2 shows the details of our official Web track runs [Note: T = Title only queries, TD=Title and Description, D=Description only].

| Run ID | Description | Query Type | K1 Constant | B Constant |
|---|---|---|---|---|
| pltr02wt1 | Distillation run | T | 1.5 | 0.8 |
| pltr02wt2 | Non-Distillation run | T | 1.5 | 0.8 |
| pltr02wt3 | Distillation run | T | 1.5 | 0.2 |
| pltr02wt4 | Non-Distillation run | T | 1.5 | 0.2 |
| pltr02wt5 | Distillation run | TD | 1.5 | 0.2 |
| pltr02wt6 | Named page run | D | 1.5 | 0.2 |
| pltr02wt7 | Named page run | D | 1.5 | 0.4 |
| pltr02wt8 | Named page run | D | 1.5 | 0.6 |
| pltr02wt8 | Named page run | D | 1.5 | 0.8 |

Table 2 – TREC 2002 Web track run details

We used 1.5 for the **K1** constants for all our runs as this was the best found in our previous Web track experiments for a large collection of web data [2]. For the topic distillation task we varied the **B** constant between 0.2 and 0.8 in order to investigate the effect of document length on this task. We also included some non-distillation runs to allow us to quantify the effectiveness of our distillation runs. Most of our distillation task runs used title only queries (realistic), but we did submit one title/description run. We used description only queries for the named page task (this was the only allowed method). We were able to vary the **B** constant on the named page task a little more as we had less flexibility on those runs: this allowed us to investigate the effect of document length in more detail for this task.

## 5. TREC 2002 retrieval efficiency results

### *5.1 Retrieval efficiency results*

Table 3 gives a sample of the average elapsed time for each of the official runs. The distillation task runs contained 50 queries, whilst the named page runs contained 150 queries. We are very satisfied with our query response times on the .GOV collection. All our runs have met the one to ten second response time criteria specified by Frakes [5], and they are good for a collection of this size. We believe that these response times could be considerably improved by using various query optimisation techniques (currently we do not use any in our query processing).

| Query Type | Distillation runs | Non-Distillation runs | Named page runs |
|---|---|---|---|
| T | 1.24 | 1.29 | - |
| TD | 7.17 | - | - |
| D | - | - | 1.48 |

Table 3 – TREC 2002 average elapsed time for official runs (sample)

## 6. Topic distillation task results

The topic distillation results are shown in Table 4.

| Run ID | Description | Precision @ 10 | Average Precision | Query Type | B |
|---|---|---|---|---|---|
| pltr02wt1 | Distillation run | 0.200 | 0.144 | T | 0.8 |
| pltr02wt2 | Non-Distillation run | 0.241 | 0.190 | T | 0.8 |
| pltr02wt3 | Distillation run | 0.175 | 0.109 | T | 0.2 |
| pltr02wt4 | Non-Distillation run | 0.200 | 0.143 | T | 0.2 |
| pltr02wt5 | Distillation run | 0.088 | 0.044 | TD | 0.2 |

Table 4 – TREC 2002 Topic distillation results

An interesting result from our experiments was that the Non-distillation runs did better than the Distillation runs, and that one of our Non-distillation runs (pltr02wt2) came second overall in this years Web track Topic distillation task [5]. Two significant observations can be made about these experiments. The first is that just using a simple minded URL removal technique to improve topic distillation simply does not work. The second is that for this task, using ordinary BM25 search techniques with no relevant feedback is comparable to those methods which utilize such evidence as document structure, anchor text and link structure. With respect to the BM25 tuning constant parameter it is clear that a lower value of B was better for both our types of runs: runs with **B** set at 0.8 did better than those with **B** set at 0.2 (when comparing like with like e.g. distillation runs).

## 7. Named page task results

The named page results are shown in Table 5.

| Run ID | MRR | % in top 10 | % not found | B |
|---|---|---|---|---|
| pltr02wt6 | 0.334 | 44.7% | 44.0% | 0.2 |
| pltr02wt7 | 0.414 | 53.7% | 41.3% | 0.4 |
| pltr02wt8 | 0.416 | 52.7% | 41.3% | 0.6 |
| pltr02wt8 | 0.418 | 52.7% | 42.0% | 0.8 |

Table 5 – TREC 2002 Named page task results

Overall the results are disappointing: in most runs we are only finding about 50% of the named pages in the top 10, and our experiments do not find up to 40% of the resources at all. Therefore our MRR results are not as good as we would have liked – up to something in the region of 0.72 as found with the top scoring run in this years Named page task [5]. We believe that one important factor may be the cause of reduced effectiveness for this task given the evidence found in topic distillation runs: all experiments used the URL removal technique – and this has obviously had a significant effect on our MRR scores. It would be useful to do Named page experiments without the URL removal procedure in order to quantify the effect of using such a method. We could also make a contribution to the IR community, being able to compare a realistic BM25 technique with those which make use of document/link structures and anchor text. It should be noted that MRR increases with the value of **B**, but the increase is not significant beyond **B**=0.4. The increase from **B**=0.2 to **B**=0.4 is significant however: the percentage increase is 24%. Increases on the other runs with increasing value of **B** are all below the half percent mark.

## 8. Conclusion

The simple minded technique of removing multiple hits from web pages used for the purposes of the experiments described in this paper, do not appear to have work particularly well. We have found, significantly, that a straight BM25 term weighting run with no relevance feedback compares very well indeed with methods which use document/link structures and anchor text in the Topic distillation task. Our Named page runs are disappointing, and we believe that part of the problem relates to removing multiple hits from web pages.

With respect to our hypothesis, we have demonstrated that for the topic distillation task, BM25 appears to work quite well. However we have not been able to demonstrate this for the named page task and further investigation is required. In particular the issues of removing documents from the top 10 when other document from the same web site have already been retrieved needs to be investigated.

The evidence from the experiments described in this paper show that altering the value of the **B** constant in the BM25 model does appear to have an effect: in particular a high value of **B** parameter appears to work well with the .GOV collection for both of this years web track tasks. We have been able to show that our system scales to much larger collections of the .GOV size, and have shown the indexing/search speeds are acceptable for data sets of this size.

## References

[1] Robertson, S.E., and Sparck Jones, K., Simple, proven approaches to text retrieval, University of Cambridge Technical report, May 1997, TR356 , [http://www.cl.cam.ac.uk/Research/Reports/TR356-ksj-approaches-to-text-retrieval.html] – visited 22nd July 2002.

[2] MacFarlane, A., Robertson, S.E.., McCann, J.A., PLIERS AT TREC8, In: Voorhees, E.M., and Harman, D.K., (eds), The Eighth Text Retrieval Conference (TREC-8), NIST Special Publication 500-246, NIST: Gaithersburg, 2000, p241-252.

[3] Fox, C., A stop list for general text, SIGIR FORUM, ACM Press, Vol. 24, No. 4. December 1990.

[4] Frakes, W.B., Introduction to information storage and retrieval systems. In: Frakes, W.B. and Baeza-Yates, R. (eds), Information retrieval; data structures and algorithms, Prentice Hall, 1992, p1-12.

[5] Craswell, N., and Hawking, D., Overview of the TREC-2002 Web Track, [in this volume].