Two selfless contributions to web search evaluation

U. Twente at TREC 2014

Djoerd Hiemstra and Robin Aly
University of Twente, The Netherlands
{d.hiemstra, r.aly}@utwente.nl

Abstract

We present our results for the Web Search track and the Federated Web Search track for the 23rd Text Retrieval Conference TREC.

1 Introduction

In this paper we describe the contributions from the University of Twente to the 23rd Text Retrieval Conference. We participate in the Web Track and the Federated Web Track. Our experiments are run by MIREX [2] (MapReduce Information Retrieval Experiments), a library of MapReduce programs to extract data and sequentially scan document representations. Built on Hadoop, sequential scanning becomes a viable approach. MIREX allows researchers to easily experiment with different retrieval models, because the framework is easy to extend. The paper is structured as follows: Section 2 describes our participation in the web track and Section 3 describes our participation in the Federated Web track.

2 Web Track Participation

The Web track models a general web search scenario, with ad hoc queries on the ClueWeb12 web collection, i.e., we investigate the performance of approaches that search a static set of documents using previously-unseen topics.

Table 1 shows the precision at 5, 10 and 20 results of the three official Web Track runs. The runs tagged utbase and utexact use anchor texts as document representations, which we made available for download. The first run, tagged utexact, matches the exact query string to the anchors, and ranks the documents by the number of exact matches found. The run finds exact matches for 40 out of 50 queries. We appended the results from the second

1 http://mirex.sourceforge.net
2 http://www.cs.utwente.nl/~hiemstra/2013/anchor-text-for-clueweb12.html
Report Documentation Page

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE
   NOV 2014

2. REPORT TYPE

3. DATES COVERED
   00-00-2014 to 00-00-2014

4. TITLE AND SUBTITLE
   Two Selfless Contributions to Web Search Evaluation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
   University of Twente, The Netherlands,

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
    presented in the proceedings of the Twenty-Third Text RETrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT
    We present our results for the Web Search track and the Federated Web Search track for the 23rd Text Retrieval Conference TREC.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:
    a. REPORT
       unclassified
    b. ABSTRACT
       unclassified
    c. THIS PAGE
       unclassified

17. LIMITATION OF ABSTRACT
    Same as Report (SAR)

18. NUMBER OF PAGES
    4

19a. NAME OF RESPONSIBLE PERSON

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18
run, i.e. those documents that were not already found by exact matches, to the run as the final result. The second run, tagged utbase, uses a simple
unigram language model with linear interpolation smoothing, $\lambda = 0.95$, and a document length prior. A run without smoothing (or $\lambda = 1$) retrieves the exact same
top 10 documents for 45 out of 50 queries, and therefore also achieves the same precision at 5 and 10 documents. The third run, tagged utold, uses the same
ranking as the second run on the full text of the web pages.

The experimental results show that utexact, exact matching of the full query string on the anchors, outperforms the other runs for precision at 5 documents retrieved, whereas utold, the language model on the full text, performs best at precision at 10. Interestingly, we expected the full text run to perform much worse, because the anchor runs use a document length prior (the more anchor text, the better), that has a similar effect as an inlinks prior (also similar to PageRank).

Table 1: Precision at 5, 10 and 20 (50 queries)

<table>
<thead>
<tr>
<th>Run</th>
<th>P@5</th>
<th>P@10</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>utexact</td>
<td>0.456</td>
<td>0.412</td>
<td>0.375</td>
</tr>
<tr>
<td>utbase</td>
<td>0.440</td>
<td>0.422</td>
<td>0.371</td>
</tr>
<tr>
<td>utold</td>
<td>0.448</td>
<td>0.428</td>
<td>0.420</td>
</tr>
</tbody>
</table>

Table 2 shows the ability of the systems to retrieve documents judged as highly relevant, key, or navigational. So, documents judged as relevant were not considered in this evaluation. The results show that clearly, the exact query string matching favours highly relevant documents for 5 documents retrieved.

Table 2: High Relevance Precision (49 queries)

<table>
<thead>
<tr>
<th>Run</th>
<th>P@5</th>
<th>P@10</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>utbase</td>
<td>0.188</td>
<td>0.186</td>
<td>0.163</td>
</tr>
<tr>
<td>utexact</td>
<td>0.216</td>
<td>0.184</td>
<td>0.161</td>
</tr>
<tr>
<td>utold</td>
<td>0.200</td>
<td>0.190</td>
<td>0.171</td>
</tr>
</tbody>
</table>

All runs were pooled to depth 20 this year, an improvement over last year’s TREC when for many topics, less than 20 documents per topic were judged,
because the pool depth of 20 would have resulted in too many documents to be judged in the allotted amount of assessing time. Table 3 shows the effects of the pool depth of the fraction of judged documents for each run. The run utold was not part of the pool that was judged. Of the runs that contributed to the pool, at 30 documents retrieved, about 10% of the documents is not judged. Last year, at 30 documents retrieved, judged fractions dropped between 22% and 25%. So, TREC seems to have done a better job judging documents this year, at least for our runs.

<table>
<thead>
<tr>
<th>Qrels</th>
<th>total in 2013</th>
<th>per query in 2013</th>
<th>total in 2014</th>
<th>per query in 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>all judged</td>
<td>14474</td>
<td>290</td>
<td>14432</td>
<td>289</td>
</tr>
<tr>
<td>highly rel. (&gt; 1)</td>
<td>1106</td>
<td>22</td>
<td>1877</td>
<td>38</td>
</tr>
<tr>
<td>relevant (&gt; 0)</td>
<td>4150</td>
<td>83</td>
<td>5665</td>
<td>113</td>
</tr>
<tr>
<td>irrelevant (= 0)</td>
<td>10090</td>
<td>202</td>
<td>8210</td>
<td>164</td>
</tr>
<tr>
<td>spam (= -2)</td>
<td>234</td>
<td>5</td>
<td>1614</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 4: Number of documents judged

Table 4 shows general statistics of the TREC 2013 and TREC 2014 Web Track collections. Of the total number of documents that are judged, more than 39% were judged relevant, so it is likely that many more relevant documents would have been found if more resources would have been available for judging.

## 3 FedWeb Track Participation

The Federated Web Track models a distributed search scenario where users send requests to a broker which forwards the requests to a set of search engines that are likely to produce relevant results. We participated in the resource selection task, which requires selecting resources based on resource descriptions, given a search request. The track provides sample texts and snippets form documents sampled from each search engine. Prior to resource selection, these documents have to be transformed into a resource description. Currently, resource descriptions based directly on the sampled documents in a central sample index are the most popular. However, this approach also requires substantial storage space and administrative overhead when selecting resources.

<table>
<thead>
<tr>
<th>Run</th>
<th>ndcg@20</th>
<th>nP@1</th>
<th>nP@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTTailG2000</td>
<td>0.178</td>
<td>0.142</td>
<td>0.223</td>
</tr>
<tr>
<td>Median Performance</td>
<td>0.182</td>
<td>0.202</td>
<td>0.250</td>
</tr>
<tr>
<td>Maximum Performance</td>
<td>0.460</td>
<td>0.534</td>
<td>0.601</td>
</tr>
</tbody>
</table>

Table 5: Official FedWeb Resource Selection results. Median and maximum performance shown for comparison.
Table 5 shows the official overall evaluation score of the run UTaillyG2000. Our run uses Taily [1], an approach for shard selection which showed good performance in centralized search, and adapt it for federated web search. Instead of using a centralized sample index, Taily uses vocabulary-based resource descriptions based on statistics of term related features in each shard that are used in ranking functions. Compared to this centralized setting, the full collection is only represented by a sample and and the ranking function of each individual search engine is unknown. Taily assumes a gamma distribution for scores of a query, which is inferred from the feature statistics.

4 Conclusion

We tried simple, out-of-the-box approaches to this year’s Web track and Federated Web track, contributing documents to the evaluation pools, and evaluation results to the TREC report. We love TREC, and hope to be there next year with some more novel approaches to text search.

Acknowledgements

We are grateful to the European Union Project AXES (FP7-269980) and the Dutch national program COMMIT for funding our work.

References
