ICTNET at Web Track 2013

Yuanhai Xue[1, 2], Xiaoming Yu[1], Feng Guan[1, 2], Yue Liu[1], Xueqi Cheng[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
2. University of Chinese Academy of Sciences, Beijing, 100190

Abstract
In this paper, we report our TREC experiments with both ad-hoc task and risk-sensitive task of Web Track 2013. The ClueWeb12 dataset and its derived data were used this year. We use BM25 model with term proximity, entity recognition and external resource to rank the final results. We submitted six runs which were not optimized for any particular metric.

1. Introduction
An ad-hoc task in TREC investigates the performance of systems that search a static set of documents using previously-unseen topics. This year, a new dataset named ClueWeb12 [1] was used as document collection. The overall goal of the risk-sensitive task is to explore algorithms and evaluation methods for systems that try to jointly maximize an average effectiveness measure across queries, while minimizing effectiveness losses with respect to a provided baseline.

The rest of this paper is organized as follows. In Section 2, we discuss the processing of ClueWeb12, derived data and external resources. In Section 3, the BM25 model with term proximity is introduced. We report experimental results and the corresponding search strategy in Section 4. Finally, our work is concluded in Section 5.

2. Data Processing
2.1 Parsing the documents
The ClueWeb12 dataset is consist of over 733 million pages, identified by TREC_ID. We parse these pages and split them into 4 parts, TREC_ID, TITLE, CONTENT and URL. The parsed documents are expressed as XML documents for index.

2.2 System
This year, we use Golaxy Search Engine [2], a high performance distributed search platform. The GSE was deployed over nine servers, one for merge, eight for index and search. Each server has 16 CPU cores, 32GB memory and 16TB hard disk.

2.3 Spam filtering
As we found in the past years, spam detection and removal was very important for improving the performance of retrieval. We use Waterloo Spam Rankings [3] as spam filter this year. The Fusion score was used. We label those pages whose percentile-score are less than 70 to be spam, and the rest non-spam. To speed up the search procedure, only the non-spam pages were indexed in experiments.

2.4 Anchor text
Anchor text usually directly indicates what a page is about, so more and more attention has been paid to it in web search engine. Actually, high-quality anchor text leads us directly to the page we want. Djoerd Hiemstra shares their anchor text [4] extracted from the TREC ClueWeb12 collection. The data contains anchor text for 0.5 billion pages, about 64% of the total number of pages in ClueWeb12. The data consists of a tab-separated text files consisting of (TREC-ID, URL, ANCHOR TEXT). For each TREC-ID, we

combine all the unique ANCHOR TEXT into one. The processed anchor text was inserted into the parsed document as the fifth part ANCHOR.

2.5 Entity recognition

Some entities such as "Nicolas Cage" and "mad cow disease" consist of more than one word. It is very useful to treat them as one word in the bag-of-words retrieval models. The Freebase Dump which is included in the ClueWeb12 dataset is an RDF version of Freebase that was provided by Google, Inc. We use the Freebase Dump to help recognize the entities in the queries.

3. BM25 model with term proximity

Okapi BM25 [5] is one of the traditional bag-of-words ranking function which is widely used by web search engines. It assumes full independence between terms, so it does not take the proximity of query terms into account. This year, we use the proximity-enhanced retrieval model named BM25PF [6] that combine the phrase frequency information with the basic BM25 model to rank the documents.

4. Results and Discussion

This year, we generate six runs which were not optimized for any particular metric. Three of them were submitted for ad-hoc task and the rest were submitted for risk-sensitive task.

The first run, ICTNET13ADR1, requires that query perfectly appears in ANCHOR part of the document. Then we apply BM25PF model on CONTENT part of the document. The second run, ICTNET13ADR2, in addition to the requirement of ICTNET13ADR1, requires that all the query terms appear in TITLE part. Then we apply BM25 model on TITLE part of the document. The third run, ICTNET13ADR3, requires that all the query terms appear in CONTENT part. Then we apply BM25 model on CONTENT part of the document.

The fourth run, ICTNET13RSR1, requires that query perfectly appears in ANCHOR part of the document. Then we remove the stop words and use entity recognition to optimize the query. BM25 model is applied on TITLE part and BM25PF model is applied on CONTENT part at the same time. The fifth run, ICTNET13RSR2, requires that all the query terms appear in ANCHOR part. Then BM25 model is applied on ANCHOR part and BM25PF model is applied on TITLE part at the same time. The sixth run, ICTNET13RSR3, requires that query perfectly appears in both ANCHOR and TITLE part. Then BM25PF model is applied on CONTENT part.

The performances of these runs are shown in table 1.

Run	ERR-IA@20	a = 0	a = 1	a = 5	a = 10
ICTNET13ADR1	0.485355	0.133034	0.069142	-0.186424	-0.505881
ICTNET13ADR2	0.503956	0.151636	0.095345	-0.129816	-0.411268
ICTNET13ADR3	0.449881	0.097560	0.031378	-0.233349	-0.564259
ICTNET13RSR1	0.528270	0.175949	0.103900	-0.184297	-0.544544
ICTNET13RSR2	0.498541	0.146220	0.045010	-0.359829	-0.865879
ICTNET13RSR3	0.551106	0.198785	0.133525	-0.127516	-0.453818

Table 1: Performance of Web track, TREC 2013

As shown in the chart, using anchor text can significantly boost the performance. However, all the runs fail to control the retrieval losses well when improving the effectiveness. We will do more intensive study in the future.

5. Conclusion

In this paper, we described our experiment in Web track, TREC 2013. This year, we explore using Freebase as high-quality external resource, but how to use it well still needs more study. We also use anchor text to filter the documents, which work well on most cases but still need to improve. On the whole, the run considering anchor text and title at the same time performs best. It is a pity that we do not find a good tradeoff between effectiveness and losses this year. We will continue to explore it in the future.

6. Acknowledgements

We would like to thank all organizers and assessors of TREC and NIST. This work is sponsored by NSF of China Grants No. 61202058 , and by the National Key Technology R&D Program (2012BAH39B04).

References

[1] The ClueWeb12 Dataset. http://www.lemurproject.org/clueweb12.php

[2] Golaxy Search Engine. http://www.golaxy.cn

[3] Waterloo Spam Rankings. http://www.mansci.uwaterloo.ca/~msmucker/cw12spam

[4] Anchor Text. http://wwwhome.ewi.utwente.nl/~hiemstra/2013/anchor-text-for-clueweb12.html

[5] Robertson S E, Walker S, Jones S, et al. Okapi at TREC-3. NIST SPECIAL PUBLICATION SP, 1995: 109-109.

[6] Zhu Y, Xue Y, Guo J, et al. Exploring and Exploiting Proximity Statistic for Information Retrieval Model. Information Retrieval Technology. Springer Berlin Heidelberg, 2012: 1-13.