ICTNET at Web Track 2011 Ad-hoc Task

Heyuan Li[1,2], Yuanhai Xue[1,2], Xu Chen[1,2], Xiaoming Yu[1], Feng Guan[1], Yue Liu[1], Xueqi Cheng[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
2. Graduate School of Chinese Academy of Sciences, Beijing, 100190

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, ClueWeb09 Dataset [1] was used again as document collection. But the topics developed for this year was less common and ambiguous than before.

The rest of this paper is organized as follows. In Section 2, we discuss the processing of ClueWeb09, derived data and external resources. In Section 3, the BM25 model with term proximity, searching with anchor text, query expansion and promoting authoritative sites are introduced. We report experimental results in Section 4 and conclude our work in Section 5.

2. Data Processing

2.1 Parsing the documents

The ClueWeb09 dataset is consist of 500 million English pages, identified by TREC_ID. We parse these pages and split them into 6 parts, TREC_ID, Title, Keywords, Content, URL and Anchors. The parsed documents are expressed as XML documents for index. During our experiments, it's very common to request a certain content or URL by TRECID. Therefore, we use Tokyo Tyrant to build three big maps, TREC_ID – URL, TREC_ID – Content and URL – TREC_ID. These maps are provided as RPC service and also used by diversity task, session track and entity track.

2.2 System

This year, we use Golaxy [2], a high performance distributed search platform. Golaxy was deployed over eleven servers, one for merge, ten for index and search. Each server has 16 CPU cores, 32GB memory and 2TB hard disk. It takes about ten hours to index all of the 500 million English documents. A C++ search client was developed for TREC experiments. To retrieve N documents, the client send query to merge server. Merge server dispatch query to search servers, merge the results from each of them and return the top N documents to the client.

2.3 Spam filtering

As we found last year, 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 filter out the bottom 50% of the documents, which are most likely to be spam. To speed up this procedure, a map of TREC_ID – Spam-Score was set up using Tokyo Tyrant as we described in Section 2.1.

2.4 Using Open Directory Project

The Open Directory Project (ODP), also known as DMOZ, is a multilingual open content directory of World Wide Web links [5]. Navigational queries are aimed at seeking a single website or web page of a single entity. Traditional retrieval model is good at Informational queries, but may not boost the navigational website. ODP, on the other hand, can help this case by looking up the directory and giving the corresponding URL directly. We use the Search function of ODP and use Python to get the top 1 result for all of the topics this year. Also, the title of the result is checked, the websites which fail to contain all the words in query were dropped and not used.

3. Retrieval Models

3.1 BM25 model with term proximity

Okapi BM25 is a bag-of-words ranking function used by search engines to rank matching documents according to their relevance to a given search query [6]. One shortcoming of BM25 is that it does not take the proximity of query terms into account. This year, we use the BM25 model that we proposed last year [7]. In addition, we use minimal window size which can cover all the query terms as another measure of term proximity. The documents with smaller minimal window size will be boosted.

3.2 Anchor text

Anchor text is ranked highly in search engine, for it tells us what the page is about. Usually, high-quality anchor text leads us directly to the page we want. Dang and Croft [4] created a publicly available Anchor Text Query Log for ClueWeb09 Dataset. We convert the URL into corresponding TREC_ID for further usage. The anchor text collection was separately indexed in the same way as original ClueWeb09 dataset. During the retrieval, documents were sorted by two keys, the BM25 scores and the counts. Firstly, the result was ranked by BM25 score. Then documents with the same BM25 score were sorted by counts.

3.3 Query expansion

The queries need to be expanded before search. This year, we apply “term search” strategy, which is a common tool in search engine such as Google, Yahoo and Bing. Term search take entity in a query as a whole part. For example, ‘ritz calton las vegas’, will be split into 2 terms, ‘ritz calton’ and ‘las vegas’, instead of 4 words. The strategy is implemented by our team members, who participate in Entity Track this year. The expanded queries were used as replacement for the original queries in the corresponding submissions.

3.4 Promoting authoritative sites

As described in Section 2.5, we use Open Directory Project (ODP) as an authoritative index. The topics of this year are less common. Therefore, ODP only works well for part of the queries. We use ODP to promote the result that was based on BM25 model with term proximity. Firstly, we select the authoritative sites that match the topics of this year. The selection was performed by a filter, which accepts the sites that match query well in their title. Secondly, we use the model to generate a baseline and apply the spam filtering. Finally, the authoritative sites were inserted into the baseline with a limitation that five sites from ODP were used at the most.

4. Results and Discussion

We submitted three runs this year for ad-hoc task.The first run, ICTNET11ADR2, uses query expansion. Then we apply BM25 model with term proximity on content field. Also, we search with BM25 model on anchor text and merge the result into previous ones. After that, the score in result set was recalculated and normalized to range [0, 100]. We use this run to confirm whether query expansion and anchor text would help with the performance.The second run, ICTNET11ADR3, use techniques that we use in diversity task. It was found last year that submission for diversity task is better than ad-hoc task. We use the same processing as we did last year in diversity task. We also use this run as a baseline to compare with last year.The third run, ICTNET11ADR4, is used to experiment with the effectiveness of authoritative sites promoting. We use BM25 model with term proximity to generate the base one. Then anchor result was inserted randomly. Finally, we place the authoritative URLs into proper position. We assume that the

authoritative URLs have better performance than previous results; therefore, URLs were inserted before existing result.

Run	Relevant Retrieved	ERR@20	nDCG@20	P@5	P@20	MAP
ICTNET11ADR2	1632	0.14701	0.28622	0.4080	0.3620	0.1747
ICTNET11ADR3	1648	0.15681	0.28261	0.3920	0.3450	0.1746
ICTNET11ADR4	908	0.11532	0.23419	0.3360	0.2840	0.0995

Table 1: Performance of ad-hoc task, TREC 2011

Table 1 summarizes the performance of our ad-hoc submission this year. As shown in the chart, the first run, which use query expansion and anchor text perform better than the last run that without it. The next run, which applies diversity techniques, is nearly the same good as run1, which confirm our guess last year. The authoritative URLs, however, was the worse in boosting the performance. We think that the position of URLs should be seriously considered.

We take a deeper look at the P@20 of run ICTNET11ADR2. Our best run fails to return any relevance results in 8 topics. Most of them were partly related, which means it's relevance to part of the query, but can't meat requirements on the whole point of view. In other words, the query expansion may lead to a wrong direction in some cases and we will do more intensive study in the future.

5. Conclusion

In this paper, we described our experiment in ad-hoc task, TREC 2011. This year, we explore using DMOZ as high-quality external resource. It's not as effective as we thought due to the order problem. We also use query expansion to redefine the quires, it work well on most cases but still need to improve. We use anchor text to promote the result, it can significant improve the result but few topics have sufficient anchor text data. The feasibility of using diversity techniques to boost Ad-hoc result was also practiced this year, it performs well. On the whole, the run using BM25 model with term proximity, query expansion and anchor text performs best. Finally, we analysis the judgment results and point out the problem encountered in our best run.

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. 60903139 and No. 60873243, and by 863 Program of China Grants No. 2010AA012502 and No. 2010AA012503.

References

[1] The ClueWeb09 Dataset - http://boston.lti.cs.cmu.edu/clueweb09

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

[3] Waterloo Spam Rankings - http://plg.uwaterloo.ca/~gvcormac/clueweb09spam

[4] Anchor Text for ClueWeb09 - http://lemurproject.org/clueweb09/anchortext-querylog

[5] Open Directory Project – http://www.dmoz.org

[6] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg, USA, November 1994.

[7] X. Chen, Z. Peng, J. Wang, X. Yu, Y. Liu, X. Cheng. ICTNET at Web Track 2010 Ad-hoc Task. Proceedings of the Nineteenth Text REtrieval Conference (TREC 2010), Gaithersburg, Maryland, November 16-19, 2010.