ICTNET at Web Track TREC2014

Yuanhai Xue[123], Xiaoming Yu[2], Feng Guan[123], Xipeng Li[123], Man Du[123], Yue Liu[12], Xueqi Cheng[12]

1)Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190

2)Key Laboratory of Web Data Science and Technology, CAS, 100190

3) University of Chinese Academy of Sciences, Beijing, 100190

{xueyuanhai, yuxiaoming, guanfeng, lixipeng, duman}@software.ict.ac.cn; {liuyue, cxq}@ict.ac.cn

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, the ClueWeb12[1] dataset are used. 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. Two baselines from different IR systems are supplied this year in order to understand the nature of risk-reward tradeoffs achievable by a system that can adapt to different baselines.

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, the diversification method and the results fusion strategy are introduced. We report experimental results and the corresponding re-ranking strategy in Section 4. Finally, our work is concluded in Section 5.

2. Data Processing

2.1 Search Engine

This year, we continue use the Golaxy Search Engine(GSE)[3], a high performance distributed search platform. The GSE is deployed over ten servers, each of which has 16 CPU cores, 32GB memory and 16TB hard disk.

2.2 Parsing the documents

The ClueWeb12 dataset is consist of over 733 million different pages, identified by TREC_ID. As the same as last year, we parse the pages and split them into 4 parts, TREC_ID, TITLE, CONTENT and URL. The parsed documents are expressed as XML documents for index. In order to speed up the index/search procedure, only the high-quality pages are used in experiments. The Fusion score of Waterloo Spam Rankings[2] is used as spam filter this year. Those pages whose percentile-score are greater than 70 are treated as high-quality ones. High-quality anchor text leads user directly to the page they want. Fortunately, Djoerd Hiemstra shares their anchor text[4] extracted from the TREC ClueWeb12 collection. The anchor texts are used as the fifth part ANCHOR.

2.3 Entity recognition

Some entities such as "orcas island", “african american music” and "windsor knot" 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 provided by Google was used to recognize entity in last year’s experiments. However, we found that a lot of noise was brought in at the same time. We choose the Wikipedia Dump to help extract the entities in the topics this year. What’s more, there exists some redirection information in the Wikipedia pages. They are extracted to treat as the synonyms of the corresponding entity.

3. Experiments

2.1 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. All the entities are treated as one word and the corresponding synonyms are used to do query expansion.

2.2 Diversification

In order to perform well on the multi-facet topics, we diversify the search results and re-rank them. For each topic, we firstly remove all the words in the topic from the search results. Then GibbsLDA++[7] is used to get the subtopics. At last, greedy algorithm are used to re-rank the results according to the document-subtopic distribution.

2.3 Result fusion

This year, we tried the result fusion to achieve risk-sensitive retrieval. For each document in the baseline or our result, its score is defined as:

$$Score_{run_j}(doc_i) = \begin{cases} \dfrac{1}{\sqrt{Rank_{run_j}(doc_i) + \theta}}, & Rank_{run_j}(doc_i) \le 1000 \\ 0, & Rank_{run_j}(doc_i) > 1000 \end{cases}.$$

$$Score(doc_i) = \beta \cdot Score_{baseline}(doc_i) + Score_{ours}(doc_i)$$

The baseline runs and our runs over TREC Web Track 2013 topics and collections are used to train the parameter θ and β.

4. Results

This year, we submitted three runs for ad-hoc task and three ones for risk-sensitive task. Firstly, we apply BM25PF to get the run named ICTNET14ADR3. Then two different greedy algorithms are used to diversify ICTNET14ADR3 to obtain ICTNET14ADR1 and ICTNET14ADR2.

As mentioned in Section 3, the three risk-sensitive runs are all generated using results fusion with $\theta = 5$. ICTNET14RSR1 use the official Indri 2014 baseline run with $\beta = 1.08$; ICTNET14RSR2 use the official Terrier 2014 baseline run with $\beta = 2.85$; ICTNET14RSR3 use ICTNET14ADR1 as baseline with $\beta = 1.00$. The performances of these runs are shown in table 1.

Table 1: Performance of Web track, TREC 2014

Run	ERR-IA@20	Indri, a = 0	Terrier, a = 0	Indri, a = 5	Terrier, a = 5
ICTNET14ADR1	0.566524	/	/	/	/
ICTNET14ADR2	0.564756	/	/	/	/
ICTNET14ADR3	0.579731	/	/	/	/
ICTNET14RSR1	0.566214	0.053179	0.023867	-0.007927	-0.252000
ICTNET14RSR2	0.536450	0.023415	-0.005897	-0.524757	-0.469410
ICTNET14RSR3	0.578743	0.065708	0.036396	-0.365485	-0.349912

As shown in the table, all the risk-sensitive runs fail to control the retrieval losses. We need do more intensive study in the future.

5. Conclusion

In this paper, we described our experiment in Web track, TREC 2014. This year, we explore using

Wikipedia as high-quality external resource to recognize entities in topics. Our diversification methods do not achieve the desired improvements in ERR-IA@20. We tried the results fusion to achieve risk-sensitive retrieval. Unfortunately, all the risk-sensitive runs fail to control the retrieval losses. 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 973 Program of China Grants *No.2012CB316303&No.2013CB329602*,863 program of China Grants *No.2012AA011003&No.2013AA01A213*,NSF of China Grants *No.61232010&No.61173008*, and by the National Key Technology R&D Program Grants *No.2012BAH39B02&No.2012BAH46B04* .

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.

[7] GibbsLDA++. http://gibbslda.sourceforge.net/