ICTNET at Federated Web Search Track 2013

Feng Guan[1,2],Yuanhai Xue[1,2], Xiaoming Yu[1], 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

{guanfeng,xueyuanhai, yuxiaoming}@software.ict.ac.cn

{liuyue,cxq}@ict.ac.cn

Abstract

This paper is about work done for result merging task of TREC 2013 Federated Web Track. We introduce three methods for calculating score of documents. These methods are based on linear combination with scores of document fields. The distinction is different weight factors. Score of base line method is the combination with score of basic html fields. Page rank score is added in second method. Documents with lower score are filtered during the third method.

1. Introduction

Federated search is the approach of querying multiple search engines simultaneously, and combining their results into one coherent search engine result page. [1] The goal of the Federated Web Search (FedWeb) track is to evaluate approaches to federated search at very large scale in a realistic setting, by combining the search results of existing web search engines. The goal of results merging, which is the second task of federated search, is to combine results selected from the given search engines into a single ranked list.

In contrast to typical TREC tasks, no document collection is provided but training collection FedWeb2012 [2,3,4] which was created to support research on federated search for the web.

Considering BM25 is a classic and effective method for document retrieval, we investigate different linear combination with scores of document fields based on BM25. The paper is organized as below:

Data preprocessing including additional resource and our query methods will be introduced in Section 2. Section 3 describes the result of our submitted runs. The conclusion will be given in Section 4.

2. Query methods

For a given query, up to 10 results of the selected search engines are provided and both title and content are extracted in XML format, but information of some other important parts is missing, such as content with different font. In our opinion, the importance of contents is in proportion to its font-size. Therefore, we extract content with four kinds of font (content with html label h1, h2, h3 and h4).

Besides, considering authority of web page could improve the ranking, we collect all the URLs in sampling data, and then query their page rank score[5].Finally, we build index with fields url, title, content, content with four kinds of font and page rank score.

For a given query q and field F, the score of similarity is calculated as follow:

$$Sq(F) = BM25(q, F)$$

The score of document D is:

$$S_q(D) = \sum_{F_i \in D} W(F_i) * S_q(F_i)$$

$$\sum_{F_i \in D} W(F_i) = 1$$

W(Fi) is the weight of field Fi in document. By adjusting W(Fi), we get three query methods. Fields of url, title, h1 and page rank are considered as important as each other.

In the first method, page rank score is not taken into account. The weights of url, title, h1, h2, h3, h4, content are 0.22, 0.22, 0.22, 0.14, 0.1, 0.05, 0.05.

In the second method, weights of url, title, pr, h1, h2, h3, h4, content are 0.17, 0.17, 0.17, 0.17, 0.16, 0.08, 0.04, 0.04.

Weights used in the 3rd method are the same as in the 2nd method. The difference is that documents with score lower than threshold (threshold=2) will be filtered out.

3. Results

We submitted three results for merging task. The result is given in table 1. For the given 51 queries, there are 11 queries obtained max P@10 in our first method and up to 13 queries in the 2nd and 3rd method. As the NDCG column shows, because of threshold, the query number of reaching max NDCG is much fewer than the former methods.

	queries with max P_10	queries with max NDCG
ICTNETRun1	11	6
ICTNETRun2	13	7
ICTNETRun3	13	1

Table 1: Results of runs

4. Conclusion

This paper reports the methods and the results of our team on Federated Web Track 2013 result merging task. We focus on improving ranking of documents from different search engines with several fields. And the results show that page rank score is one positive factor for merging task.

Although BM25 is an effective method for ad-hoc task, other methods such as classification and language model should taken into consideration for result merging. According to our observation, search engines may have categories, for example video search engine, QA search engine, shopping search engine and so on. And queries could tell the identification of user goals. Thus, in the future, we will devote to combine BM25 with classification of both search engines and queries.

5. Acknowledgement

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).

6 Reference

[1] https://sites.google.com/site/trecfedweb/

[2] Dong Nguyen, Thomas Demeester, Dolf Trieschnigg, and Djoerd Hiemstra. "Federated Search in the Wild: The Combined Power of over a Hundred Search Engines". In Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM), 2012

[3] Thomas Demeester, Dong Nguyen, Dolf Trieschnigg, and Djoerd Hiemstra (2012). "What Snippets

Say About Pages in Federated Web Search". In Proceedings of the 8th Asia Information Retrieval Society Conference (AIRS), 2012

[4] Thomas Demeester, Dong Nguyen, Dolf Trieschnigg, Chris Develder and Djoerd Hiemstra (2013). "Snippet-based Relevance Predictions for Federated Web Search". In Proceedings of the 35th European Conference on Information Retrieval (ECIR), 2013

[5] http://prchecker.info/