Evaluating the Effectiveness of Axiomatic Approaches in Web Track

Peilin Yang
Department of Electrical and Computer Engineering
University of Delaware
franklyn@udel.edu

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware
hfang@udel.edu

ABSTRACT

In this paper we describe our efforts for TREC 2013 Web track. We focus on evaluating the effectiveness of axiomatic retrieval model on large data collection. Axiomatic approach basically searches for the retrieval functions that satisfy some reasonable retrieval constraints. We also evaluate the semantic term matching method which does the query expansion by choosing the semantically related terms. Experiment results on adhoc task and diversity task demonstrate the effectiveness of the method.

1. INTRODUCTION

TREC 2013 Web track has two main tasks - Adhoc task and Risk Sensitive (RS) task. We participate both Adhoc and RS task using the similar method, an axiomatic retrieval model with query expansion using semantically related terms to queries.

Axiomatic retrieval models have recently proposed [3, 2] and have been verified as effective models in comparing with some other well known baselines such as Okapi-BM25 and Pivoted Normalization. The main idea of axiomatic approach is to construct retrieval functions that satisfy a set of reasonable retrieval constraints. Fang and Zhai [2] proposed several basic axiomatic retrieval functions based on the existing retrieval functions and the proposed constraints. The proposed functions are less sensitive to the parameter setting than other existing retrieval functions and obtain comparable optimal performance. To further improve the the performance, the semantic term matching based query expansion method has also been proposed [3] under the axiomatic retrieval framework. In such approach, the semantic similarity between two terms are measured based on their mutual information computed over a carefully constructed working set. The weights of the semantically related terms are regulated by set of reasonable semantic term matching constraints. The performance highly depends on the choice of the working set that used to compute term mutual information since the working set affects the quality of the semantically related terms. In our experiments, we tested two different working sets: (1) the working set constructed from the test collection itself, and (2) the working set constructed from the Web search engine snippets. Experiment results show that Web working set performs better.

2. RETRIEVAL METHOD

Previous study derived several basic axiomatic retrieval functions [2]. Our preliminary experiments on the data collection of ClueWeb09 Category B show that the F2-LOG function outperform other functions. The F2-LOG retrieval function is shown as follows:

$$S(Q,D) = \sum_{t \in Q \bigcap D} C(t,Q) \times \frac{C(t,Q)}{C(t,Q) + s + s \cdot \frac{|D|}{avdl}} \times ln \frac{N+1}{df(t)} \quad (1)$$

where Q is the query, D is the document, $C(t,Q)$ is the term count of term t in Q, $|D|$ is the document length, $avdl$ is the average document length, N is the total number of documents and $df(t)$ is the document frequency of t.

The semantic term matching method [3] can connect the vocabularies between documents and queries and thus overcome the limitation of syntactic term matching. The method relies on three semantic term matching constraints to balance the importance of the semantic related terms and the original query terms. After incorporating the semantic term matching, the retrieval scores of a single term document t for query Q can be computed based on the following functions:

$$S(Q,t) = \frac{\sum_{q \in Q} s(q,t)}{|Q|},$$
$$\text{where } s(q,t) = \begin{cases} \omega(q) & \text{if } t = q \\ \omega(q) \times \beta \times \frac{s(q,t)}{s(q,q)} & \text{if } t \neq q \end{cases} \quad (2)$$

where t is a term in the document, q is a term in query Q, $\omega(q)$ is the idf of q and β is the parameter that controls how much we trust the semantically related term. $s(q,t)$ is the semantic similarity between q and t. The semantic similarity between terms, i.e., $s(q,t)$ is computed with the mutual information:

$$s(q,t) = I(X_q, X_t|W)$$
$$= \sum_{X_q, X_t \in \{0,1\}} p(X_q, X_t|W) \cdot log \frac{p(X_q, X_t|W)}{p(X_q|W)p(X_t|W)} \quad (3)$$

where X_q and X_t are two binary random variables that denote the presence/absence of query term q and term t in the document. W is the working set to compute the mutual information.

The implementation of our method basically consists of three steps:

1. The working set to compute the term similarity is contructed. An effective method [3] to build the working set is used. In particular, the working set includes R relevant documents and $N \times R$ randomly chosen documents. We set R as 20 and N as 19 as previous study

Table 1: Optimal β in training

Method	β
UDInfolabWEB1	0.1
UDInfolabWEB2	1.7

Table 2: Mean Performance of Our Runs (Adhoc)

	ERR	ERR-IA
UDInfolabWEB1	0.1149	0.4943
UDInfolabWEB2	0.1755	0.5819

indicates. The term similarity is then computed using Equation 3.

2. We choose top K similar terms for each query and combine them to form the expanding term candidates. The similarity between each candidate term and the whole query is computed using Equation 2. M most similar terms are chosen with weights $S(Q,t)$. We set K as 1000 and M as 20 in our experiments.

3. We rank documents using Equation 1 with the expanded queries.

We apply two different working sets in step 1:

- **Collection-based working set**: We can use the collection itself, i.e., ClueWeb12 Category A.
- **Web-based working set**: The other working set we use the snippets from leading Web search engines (three of them) by submitting the queries and collect the top 100 returned snippets.

3. SUBMITTED RUNS AND EXPERIMENT RESULTS

We submitted two runs **UDInfolabWEB1** and **UDInfolabWEB2**. Both runs use semantic term matching method. UDInfolabWEB1 selects semantically related terms using Collection-based working set while UDInfolabWEB2 uses Web-based working set. For test collection, we use ClueWeb12 Category A. However, when building the inverted index, we first use Indri's[1] default language model to retrieve 10,000 top ranked documents for each query and then filter out the documents that have spam score less than -130 [1]. The filtered documents are used to build a much smaller index. The preliminary experiments on ClueWeb09 Category B decides the optimal value of parameter β in Equation 2. Table 1 shows the optimal β for each run. We can see that the optimal β for Collection-based working set is only 0.1 which indicates the low quality of expanded terms. For Web-based working set the optimal β is 1.7, which is much larger, indicating the effectiveness of expanded terms from web.

The performance that average over all queries of our runs are shown in Table 2. We only include the ERR and ERR-IA. Other evaluation measurements are similar. From the table, We can see that UDInfolabWEB2 outperforms UDInfolabWEB1 for both ERR and ERR-IA. Higher ERR indicates the Web-based working set is more effective in terms of selecting semantically related terms for query expansion.

[1]http://www.lemurproject.org/indri/

Table 3: Mean Performance of Our Runs (RS)

	UDInfolabWEB1	UDInfolabWEB2
ERR($\alpha = 0$)	0.0186	0.0793
ERR-IA($\alpha = 0$)	0.1419	0.2295
ERR($\alpha = 1$)	-0.0172	0.0604
ERR-IA($\alpha = 1$)	0.0465	0.1682
ERR($\alpha = 5$)	-0.1606	-0.0149
ERR-IA($\alpha = 5$)	-0.3352	-0.0771
ERR($\alpha = 10$)	-0.3399	-0.1090
ERR-IA($\alpha = 10$)	-0.8123	-0.3837

Higher ERR-IA indicates that Web-based working set is also capable of finding expanded terms for different subtopics.

One important change of this year's Web track is that it introduces the Risk-Sensitive task. We do consider such changes. When training the parameter β on ClueWeb09 Category B, we train it with respect to $\alpha = 1$ in the risk sensitive model [4]. However, the trained β are exactly the same as the β trained for Adhoc task and thus not shown here. The more detailed results are shown in Table 3. From the table we see that our runs generally perform worse than the Adhoc task. When α is getting larger, the performances are getting worse. UDInfolabWEB2 still outperforms UDInfolabWEB1 for all α. Apparently, our method degrades the performance of some queries. We will analyze the deep reasons in the future work.

4. CONCLUSION

In this paper, we report our methods and experiments in TREC 2013 Web track. An axiomatic retrieval model F2-LOG and the semantic term matching based query expansion approach are explored and studied. When building term expansion working set, we try collection-based one and web-based one. The experiment results show that our method is effective in terms of both adhoc and diversity measurements. The collection-based working set performs better than web-based counterpart.

5. REFERENCES

[1] G. V. Cormack, M. D. Smucker, and C. L. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Inf. Retr.*, 14(5):441–465, Oct. 2011.

[2] H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 480–487, New York, NY, USA, 2005. ACM.

[3] H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 115–122, New York, NY, USA, 2006. ACM.

[4] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 761–770, New York, NY, USA, 2012. ACM.