Evaluating Axiomatic Retrieval Models in the Core Track

Yue Wang, Peilin Yang, and Hui Fang

Department of Electrical and Computer Engineering
University of Delaware
140 Evans Hall, Newark, Delaware, 19716, USA
{wangyue,franklyn,hfang}@udel.edu

Abstract. Axiomatic analysis of the existing retrieval models have shown great power on understanding the retrieval models, which further shows a great opportunities on improve the retrieval performances of existing models. In this work, we tested the axiomatic retrieval models with query expansion on the newly provided data collection. The results show that the proposed method could outperform the baseline methods.

1 Introduction

The Core track organized in TREC 2017 is a new track with the purposes of providing a new test collection for the research community. The track is formulated as an ad-hoc retrieval task. With the same topic used in the Robust track, a new data collection, NY times corpus, is provided to the participants. We participated the Phrase 1 of the track using the axiomatic retrieval model with query expansion utilizing semantically related terms to the queries.

Axiomatic analysis of the existing retrieval methods have been proposed and studied recently [1, 2]. The fundamental idea of the axiomatic analysis is based on a set of reasonable constraints that the retrieval function should follow in order to achieve the satisfying results. Existing work have studied various constraints, starting from the basic ones, such as the basic term frequency constraints, term discrimination constraints, and document length normalization constraints, to the ones capture more sophisticate features, such as semantic term matching constraints and term proximity constraints. Under the same framework, Fang and Zhai [3] also studied the query expansion techniques using semantic term matching, in which the mutual information is utilized to compute the similarity between two terms.

Since the previous studies of axiomatic analysis of the retrieval methods have shown satisfying performances on different well-known data collections [4], we proposed to apply the same method on the newly constructed data set to verify if it is robust.

2 Axiomatic retrieval model and query expansion

2.1 Axiomatic retrieval model

Previous study [2] revealed that the functions in group F2, i.e., F2-EXP and F2-LOG, would usually outperforms than the ones with F1 and F3. Therefore, we selected the F2-EXP and F2-LOG as the basic retrieval functions for this track. The F2-EXP is defined as:

$$S(Q,D) = \sum_{t \in Q \cap D} C(t,Q) \times \frac{C(t,Q)}{C(t,Q) + s + s \cdot \frac{|D|}{avdl}} \times \left(\frac{N+1}{df(t)}\right)^k \tag{1}$$

while the F2-LOG is defined as:

$$S(Q,D) = \sum_{t \in Q \cap D} C(t,Q) \times \frac{C(t,Q)}{C(t,Q) + s + s \cdot \frac{|D|}{avdl}} \times ln \frac{N+1}{df(t)} \tag{2}$$

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

2.2 Query expansion of semantic term matching

Query expansion is a commonly used technique to include more relevant terms into the original query, thus the vocabulary gap of the query and the document could be overcame. The semantic term matching method proposed in [3], which is based on three semantic term matching constraints, would evaluated the relevance score of a single term document t with the query Q using the following equation:

$$S(Q,t) = \frac{\sum_{q \in Q} s(q,t)}{|Q|} \tag{3}$$

where

$$s(q,t) = \begin{cases} w(q), & t = q \\ w(q) \times \beta \times \frac{s(q,t)}{s(q,q)}, & t \neq q \end{cases} \tag{4}$$

where t is a term in the document, q is a term in query. $s(q,t)$ is the semantic similarity between q and t. It is computed using the mutual information:

$$\begin{aligned} s(q,t) &= I(X_q, X_t | W) \\ &= \sum_{X_q, X_t \in \{0,1\}} p(X_q, X_t | W) \cdot log \frac{p(X_q, X_t | W)}{p(X_q|W) p(X_t|W)} \end{aligned} \tag{5}$$

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

We selected the working set from two resources, e.g., the collection itself and a web-based search results snippets collection. We used the snippets from three commercial search engines, e.g. Google, Yahoo, and Bing!, to create a snippets collection for each query. The top 100 results from each search engine is crawled for each query. We denote the collection itself as internal, and the web-based snippet collection as external.

Each working set is constructed using R relevant documents and $N \times R$ randomly selected documents. N is set to 19 based on the previous experiments. We then computed the term similarity using equation 5 with each working set. Top K terms for each query are selected as expansion terms. The similarity score between the expansion term and the whole query is computed using equation 3. M most similar terms are chosen with weight $S(Q,t)$. Both the original query and the expanded query terms are used to evaluated the document. The documents are ranked based on equation 1 and 2.

3 Experiment

3.1 Data set pre-processing

The NYT corpus is downloaded and extracted. We used Indri to build the index. The porter stemmer is applied, while stop words is not removed from the collection when we built the index. We did not extract the content from the xml. It is directly built by using indri build index, with the file format set to xml. We did the same pre-processing to the snippets collection.

3.2 Experiment results

We only participated phase 1 of this year's track. We submitted three runs and the details of each run is summarized in Table 1. The *internal* query expansion resource refers to using the collection itself as the the working set, while the *external* query expansion resources refers to the web-based search results snippets collection. The parameters, M, R, K, and β, are tuned using the Robust04 collection.

Table 1. Details of submitted runs.

	Basic retrieval function	**Expansion resources**	M	R	K	β
UDelInfoLOGint	F2Log	Internal	17	19	21	0.4
UDelInfoLOGext	F2Log	External	21	23	22	1.5
UDelInfoEXPint	F2Exp	Internal	17	21	22	0.6

The performance of each run is reported in Table 2. To compare the performance with other runs, we also included the average of the median performance of all submitted runs, denoted as **TREC-Median**.

Table 2. Performance of submitted runs.

	AP	**NDCG**	**P10**
UDelInfoLOGint	0.2984	0.5391	0.5500
UDelInfoLOGext	0.2516	0.4923	0.5300
UDelInfoEXPint	0.2844	0.5249	0.5240
TREC-Median	0.2280	0.4787	0.5480

By comparing the performance of UDelInfoLOGint with UDelInfoEXPint, we could conclude the F2log method performs better than the F2exp method, which is consistent with the observations from previous study. In addition, all of our three runs could outperform the TREC-Median, which indicates the axiomatic retrieval model is robust on the new data collection.

The performance of using external resources is significantly worse than using the internal resources, since the UDelInfoLOGint outperforms UDelInfoLOGext. This is inconsistent with previous experiments. Therefore, we further investigate the results by re-train the methods using this year's data collection. The re-trained best performance (in terms of AP) of UDelInfoLOGint is 0.3026, and it is 0.2986 for UDelInfoLOGext. Thus, although the improvement using internal resource is still better than using external resources, the difference is not significant. By taking a close analysis, the huge improvement of using external resources comes from the the parameter β. In the training set, β is about 1.4-1.6 when it reaches the best performance, however in the new collection, it reaches the best performance around 0.4.

4 Conclusion

We applied the axiomatic retrieval models with semantic term matching on the newly proposed data collection in this year's Core track. The results show that the axiomatic methods are robust on the new collection. In addition, the submitted runs show that selecting the expansion terms from the original document set is better than from the external set, however, after training the parameters on the newly proposed data collection, we found out that the improvement of using internal resources is not significant comparing using the external resources.

References

1. Fang, H., Tao, T., Zhai, C.: A formal study of information retrieval heuristics. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '04, New York, NY, USA, ACM (2004) 49–56
2. Fang, H., Zhai, C.: 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, New York, NY, USA, ACM (2005) 480–487
3. Fang, H., Zhai, C.: 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, New York, NY, USA, ACM (2006) 115–122
4. Yang, P., Fang, H.: Evaluating the effectiveness of axiomatic approaches in web track. In: TREC. (2013)