ICTNET at Web Track 2009 Ad-hoc task

Feng Guan[12], Xiaoming Yu[1], Zeying Peng[12], Hongbo Xu[1], Yue Liu[1], Linhai Song[12], Xueqi Cheng[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
2. Graduate School of Chinese Academy of Sciences, Beijing 100190

Abstract

This paper is about the work done for ad-hoc task of TREC 2009 Web Track. We introduce three methods for this task, including two improved BM25 models and query expansion. The results of these models indicate that both minimum window and query expansion could improve BM25 model.

1. Introduction

An ad-hoc task in TREC investigates the performance of systems that search a static set of documents using previously-unseen topics. The goal of the task is to return a ranking of the documents in the collection in order of decreasing probability of relevance.

The document collections for web track this year are brand new. One is full collection, Category A, consists of roughly 1 billion web pages in multiple languages. The other, Category B, is a subset of the full collection which is about 50 million English-language pages. We chose the Category B for our experiment.

As the language model needs the probability of appearance of each word, it's time-unacceptable for us even on Category B. Considering BM25 is a classic and effective method for document retrieval, we investigate minimum window based on BM25 model and query expansion technique for ad-hoc task. The paper is organized as below:

Query methods including baseline and our new methods will be introduced in Section 2. Section 3 describes the experiments of traditional BM25 model and the result of our submitted runs. The conclusion will be given in Section 4.

2. Query methods

2.1 Baseline

BM25 [1] is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document based on the vector space model. It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. As it is highly effective in ad-hoc information retrieval tasks[2], one of the most prominent instantiations of the function is as follows which is also our choice as baseline.

$$sim(Q,D)=\sum_{q\in Q}\frac{f(q,D)\cdot(k_1+1)}{f(q,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{avgdl})}\cdot\log\frac{N-df(q)+0.5}{df(q)+0.5}$$

Where Q is query containing term q; f(q,D) is the term frequency of q in document D; |D| is the length

of D; avgdl is the average length of total N documents; df(q) is the document frequency of q; k1 and b are parameters.

2.2 Minimum window

One shortcoming of BM25 is that it does not take the proximity of query terms within a document into account. Considering the distance of query terms, we present a novel method named minimum window. The basic idea of this method is if all terms of a query appear in a smaller area, the document is more likely to be relevant. We rank the documents with minimum window based on BM25 as follow:

$$sim(Q, DOC_{content}) = \frac{f_w \times boost \times \sum_{q \in Q} BM25(q)}{\log(w+1)}, w \le MAXWINDOW$$

where Q is a query containing terms q; DOCconetent is the document content; BM25(q) is the BM25 score of query term q; w is the minimum window size that containing all query terms; MAXWINDOW is the upper limit of w; fw is the frequent of different minimum window in DOCcontent; boost is the weight of Q.

Our method has several parameters, including a parameter for upper limit of window size, i.e. MAXWINDOW and two parameters for adjusting precision in BM25, i.e. k1 and b.

2.3 Minimum window with URI

We notice that some terms of query may appear in the URI, and the document whose URI contains at least one query term might be more relevant than those does not, and the closer to the host name of URI the query terms are, the more likely to be relevant the document will be. The similarity of query Q and URI is calculated as follow:

$$sim(Q, DOC_{URI}) = \sum_{q \in Q} \frac{boost}{\log(Position(q)+1)+0.5}, boost \ge 3$$

Where Q is a query consisting term q; DOCURI is the URI of a document; Position(q) is the first position of q appeared in URI; boost is the weight that could distinguish between the sim (Q, DOCURI) and sim(Q, DOCcontent). The final similarity of query Q and document DOC is:

$$sim(Q, DOC) = sim(Q, DOC_{content}) + sim(Q, DOC_{URI})$$

2.4 Query expansion

Since the given topics are usually short, we expect query expansion would deal with the word mismatch problem. We make use of the method LOCOOC[3], which utilizes the local co-occurrence information in top-ranked documents and global statistical information in the whole collection to select most appropriate expansion terms. Given a query Q, our system returns a collection S of n relevant documents from the whole collection C. According to the function as below, we can obtain the similarity of one term w and the query Q. Then we select top K terms as expansion terms.

$$f(w,Q,S)=\sum_{q\in Q} idf(q)idf(w)\log(cood(w,q\mid S)+1.0)$$

Where q is the term of query Q; cood(w,q|S) and idf(q) is calculated as follow:

$$cood(w,q\mid S)=\frac{\sum_{D\in S}\log(tf(w\mid D)+1.0)\times\log(tf(q\mid D)+1.0)}{|S|}$$

$$idf(q)=\frac{\log(N)}{\log(df(q)+1.0)}$$

D is a document; tf(w|D) is the term frequent of w in document D; N is the number of documents in collection C; df(q) is the document frequency of term q.

3. Experiment and Result

As the documents of Category B are raw web pages, the dataset was processed as follow. We extracted WARC-TREC-ID, WARC-Target-URI and some important parts of HTML (title, keyword, anchor and content) to generate XML document. Meanwhile, the tags of HTML were stripped. Then we used Firtex[4] search engine to build index for experiment. Stop words were only used at query time. Queries were stopped using a standard list of common terms.

We submitted three runs. The first run (ICTNETADRun3) used minimum window. The second run (ICTNETADRun4) used query expansion, i.e. method LOCOOC. The third run (ICTNETADRun5) took URI into account while using minimum window.

Run	statMAP	statMRP	statMNDCG
baseline (BM25)	0.1340	0.2094	0.2636
ICTNETADRun3 (Minimum window)	0.1986	0.2722	0.3523
ICTNETADRun4 (Query expansion)	0.1746	0.2626	0.3409
ICTNETADRun5 (Minimum window + URI)	0.1407	0.2237	0.2976

Table 1: Results of baseline and our submitted runs.

The results of our experiment are given in Table 1. We present scores of MAP, MRP and

Run	eMAP	P@5	P@10	P@20
Best MAP	0.0476	0.3458	0.3999	0.4098
Best P@k	0.0460	0.5419	0.5282	0.5223
ICTNETADRun3 (Minimum window)	0.0433	0.4421	0.4436	0.4408
ICTNETADRun4 (Query expansion)	0.0422	0.4412	0.4427	0.4457
ICTNETADRun5 (Minimum window + URI)	0.0378	0.3568	0.3987	0.4155

Table 2: Performance of our runs

MNDCG on valid topics. We can see that both Minimum window and Query expansion can improve the traditional BM25 model.

The results of our runs are given in Table 2. We present scores of eMAP (expected MAP), P@k (expected precision at top k result).

4. Conclusion

This paper reports the methods and the experiments of our team on Web Track 2009 ad-hoc task. We focus on improving BM25 model and query expansion with LOCOOC. Our new method and query expansion could significantly improve BM25. During the minimum window method, we only focus on the appearance of all terms. Thus, our new method may be more suitable for shorter ones. That's because when query becomes longer, most of the results will contain part of the query. The minimum window will no longer work.

In the future, we will devote to improve our minimum window method and explore a better way to combine minimum window with URI or other factors.

5. Acknowledgement

Thank to the organizers of TREC 2009 Web track, the NIST assessors and the other track participants for judging the documents. In addition,our work was supported by Key Program of NSFC 60933005，863 Program of China 2007AA01Z441，973 Program of China 2007CB311100.

References

[1] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference, Gaithersburg, USA, November 1994.

[2] Stefan Büttcher, Charles L. A. Clarke, and Brad Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 621-622, New York, NY, 2006. ACM Press.

[3] Guodong Ding, Shuo Bai, Bin Wang. Local Co-occurrence based Query Expansion for Information Retrieval. Journal of Chinese Information Processing. 2006 (03).(in Chinese)

[4] http://www.firtex.org