Employing query expansion models to help patients diagnose themselves

Fangmei Lu

University of Shanghai for Science and Technology, Shanghai, China

fangmei.l@gmail.com

Abstract. In the paper we use two query expansion models, Kullback-Liebler Divergence(KLD) model and parameter-free Bose-Einstein statistics-based (Bo1) model, to improve effectiveness of information retrieval systems and help lay people search relevant medical information for diagnosing themselves. Compared with Baseline BM25, the results of Bo1 models with 3 feedback documents are better than baseline but are not statistically significant, and the performance of Bo1 is generally better than KLD.

Keywords: Information Retrieval, Query Expansion, Medical Information

1 Introduction

To fuel the effectiveness of information retrieval systems, the Task2 of 2015 CLEF eHealth[1][2] is designed to support lay people who are confronted with a sign, symptom or condition to find out more information about the condition they may have. For example, when confronted with signs of jaundice, they may use queries like "white part of eye turned green" to search for information that allow them to diagnose themselves or better understand their health conditions. These queries are often circumlocutory in nature, In 2015, Task2 includes a monolingual IR task and a multilingual IR task. We participated in the former only.

2 Our approach

2.1 Dataset

The dataset for Task 2 is provided by Khresmoi project[1], which has a set of medical related documents in HTML format and its size is about 43G (uncompressed). All of the documents are crawled from well-known health and medical

sites and databases. During indexing, each term is stemmed using Porter's English stemmer, and standard English stopwords are removed.

2.2 Topics

There are 66 topics in Task2, and the format is as follows,

```
<top>
  <num>clef2015.test.1</num>
  <query>many red marks on legs after traveling from us</query>
</top>
```

2.3 Baseline

Okapi BM25 is a well-known ranking model used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others[3]. Here we use BM25 as baseline, and the value of parameter b is 0.3.

2.4 Query expansion model

We employ two models, Kullback-Liebler Divergence (KLD) model and parameter-free Bose-Einstein statistics-based (Bo1) model, to expand the query with informative terms from the pseudo-relevance documents.

2.3.1 Kullback-Liebler Divergence

Information theoretic approaches used in query expansion are based on studying the difference between the term distribution in the whole collection and in the subsets of documents that are relevant to the query, in order to, discriminate between good expansion terms and poor expansion term. One of the most interesting approaches based on term distribution analysis is using the concept the Kullback-Liebler Divergence to compute the divergence between the probability distributions of terms in the whole collection and in the top ranked documents obtained using the original user query[4]. The most likely terms to expand the query are those with a high probability in the top ranked set and low probability in the whole collection. For the term t this divergence is:

$$KLD(t) = \left[p_R(t) - p_C(t) \right] \log \frac{\frac{f(t)}{NR}}{p_C(t)}$$

Here PR(t) is the probability of t estimated from the corpus R. PC(t) is the probability of t estimated using the whole collection. To estimate PC(t) , we used the ratio between the frequency of t in C and the number of terms in C, analogously to PR(t);

$$P_R(t) = \begin{cases} \gamma \dfrac{f(t)}{NR} if t \in V(R) \\ \delta\, p_C(t) otherwise \end{cases}$$

Where
c is the set of all documents in the collection
R is the set of top retrieved documents relative to a query.
V(R) is the vocabulary of all the terms in R.
NR is the number of terms in R.
f(t) is the frequency of t in R.

We have done our experiments with one more variation in which we have used a function other than f(t) / NR, taking also into account the likely degree of relevance of the documents retrieved in the initial run[3]:

$$KLD\_\text{variation}(t) = \left[p_R(t) - p_c(t)\right] \log \frac{\dfrac{\sum_d f(t) \times score_d}{\sum_t \sum_d f(t) \times score_d}}{p_c(t)}$$

2.3.2 Bose-Einstein statistics-based (Bo1) model.

The Bo1 model calculates the weight of terms, as followings[5],

$$w(t) = tf_x \cdot log_2 \frac{1 + P_n(t)}{P_n(t)} + log_2(1 + P_n(t))$$

$$P_n(t) = \frac{tfc}{N}$$

Where tfx is the frequency of the query term t in the top-ranked documents, tfc is the frequency of term t in the collection, and N is the number of documents in the collection.

3 Submitted runs and Results

We explored different quantities of feedback documents from 3 to 20 for the two query expansion model based on BM25's first-pass run and submitted our runs to CLEF 2015. Table 1 and Table2 shows the evaluation results. Obviously, Run1 is the best result among our runs, where Bo1 model is used, the number of feedback documents is set to 3 and the number of expansion terms is 10.

Figure 1 shows the comparison of Run1 against the median and best performance(p@10) across all systems submitted to CLEF for each query topic.

Table 1. KLD model varying number of documents(Number of expansion terms:10)

# Feeback Documents	P@5	P@10	NDCG@5	NDCG@10	MAP
Baseline	0.3636	0.3045	0.3032	0.2841	0.1754
3	**0.3848**	**0.3379**	**0.3056**	**0.3000**	**0.1787**
5	0.3485	0.3030	0.2643	0.2627	0.1580
10	0.3121	0.2727	0.2279	0.2305	0.1356
20	0.2788	0.2470	0.2072	0.2082	0.1251

Table 2. KLD model varying number of documents(Number of expansion terms:10)

# Feeback Documents	P@5	P@10	NDCG@5	NDCG@10	MAP
Baseline	0.3636	0.3045	0.3032	0.2841	0.1754
3	0.3061	0.2439	0.2466	0.2220	0.0990
5	0.2636	0.1985	0.2004	0.1757	0.0788
10	0.1818	0.1439	0.1349	0.1241	0.0497
20	0.1667	0.1348	0.1178	0.1145	0.0460

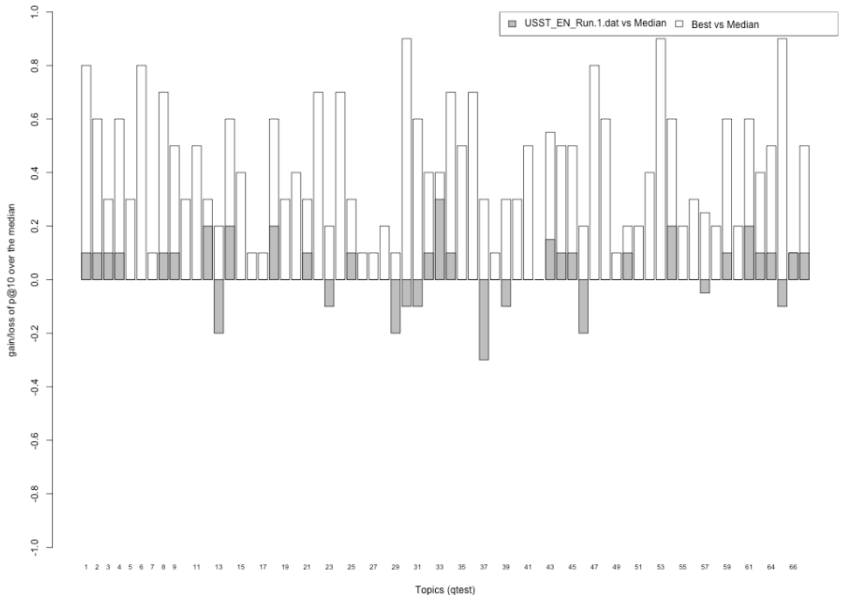

Fig. 1. The comparison of Run1 against the median and best performance(p@10) across all systems submitted to CLEF for each query topic.

4 Conclusions

Compared with Baseline BM25, the results of Ko1 models with 3 feedback documents are better than baseline while are not statistically significant, and the performance of Ko1 is generally better than KLD. In the future research, we will continue to explore other expansion models to find an effective way of supporting patients to find useful medical information .

References

1. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J. and Zuccon, G. Overview of the CLEF eHealth Evaluation Lab 2015. CLEF 2015 - 6th Conference and Labs of the Evaluation Forum.Lecture Notes in Computer Science (LNCS), Springer,2015,September.
2. Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G.JF, Lupu, M. and Pecina, P. CLEF eHealth Evaluation Lab 2015, task 2: Retrieving Information about Medical Symptoms, CLEF 2015 Online Working Notes, 2015, CEUR-WS.
3. Robertson, S. E. and Walker, S. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '94, pages 232–241, New York, NY, USA, 1994.Springer-Verlag New York, Inc.
4. Imran, H. Selecting effective expansion terms for better information retrieval, International Journal of Computer Science and Applications, Vol. 7, No. 2, pp 52 - 64, 2010
5. Amati, G. and Rijsbergen,C. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20, 4 (October 2002), 357-389.