York University at TREC 2012: Medical Records Track

Jun Miao, Zheng Ye, Jimmy Huang
Information Retrieval and Knowledge Managment Lab
York University, Toronto, Canada
{jun97, yezheng, jhuang}@yorku.ca

Abstract

In this paper, we present our participation in the Medical Records Track of TREC 2012. This is the second time we take part in this track. 50 new topics have been published in this year. The goal of this track is still to find relevant patients that have particular diseases and/or treatments. To achieve this goal, we try four methods which include popular techniques like query expansion, concept recognition and so on. Four runs have been submitted and they are based on our previous work. Detailed discussion has been made to show the effectiveness of different techniques on the Medical Records dataset.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software;

Keywords

BM25, Query Expansion, Medical Concept, Proximity Information

1 Introduction

With the rapid growth of electronic clinic reports, advanced information retrieval and knowledge discovery systems are extensively needed. In order to discover effective methods for Electronic Health Record (EHR) search, TREC started a new track, Medical Record Track, in 2011 and published a dataset containing 101,712 clinic reports and 35 topics. In 2012, 50 more topics are published. The goal of this track is retrieving relevant patients for a particular topic like "Patients who developed disseminated intravascular coagulation in the hospital". To protect the privacy of patients, the data provider use "visit" which represents an individual patient's single stay at a hospital. Although the documents in the dataset are clinic reports, the unit of retrieval for this task is "visit" instead of "report". Thus, how to map retrieved "reports" to "visits" is a challenge.

This is the second year that York University participates in this task. We submitted four runs based on different information retrieval models and techniques. All our four runs are obtained automatically. Details about our runs will be discussed in the following sections. There are some changes in evaluation metrics in this year. Bpref is not a criterion anymore while infAP and infNDCG are introduced as new metrics.

This paper is organized as follows: Section 2 presents our methods applied in the four submitted runs. Section 3 details our experimental results on different evaluation metrics. Finally, in Section 4, we conclude the paper and present our future work.

2 Four runs submitted by York University

Unlike in last year, we do not use age/gender information to filter retrieved report this time because this information was not so effective as we expect. A possible reason is that the retrieved reports have already contained the correct age/gender information. For instance, most patients who have breast cancer are females. So, "breast cancer" in the topic "Female patients with breast cancer with mastectomies during admission" has already covered the "Female" information. Our four submitted runs are denoted as: YorkUMB1, YorkUMC2, YorkUMQ3 and YorkUMP4. Table 1 presents the four official submitted runs. Detailed descriptions about these runs are as follows.

Table 1: Submitted runs

Official Runs	**Description**
YorkUMB1	Use BM25 algorithm to obtain a baseline run.
YorkUMC2	Use concept relationships in queries to improve the result of baseline run.
YorkUMQ3	Use an extended Rocchio's feedback framework with the BM25 weighting model.
YorkUMP4	Use an extended Rocchio's feedback framework and proximiy information with the BM25 weighting model.

2.1 BM25

We obtain our first run, YorkUMB1, by applying the classic probabilistic model BM25 [2] [3]. The first run is considered as a baseline run so that we can investigate how other techniques performs over BM25. In BM25, search term is assigned weight based on its within-document term frequency and query term frequency. The corresponding weighting function is as follows.

$$w = \frac{(k_1+1)*tf}{K+tf} * \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} * \frac{(k_3+1)*qtf}{k_3+qtf} \\ \oplus \quad k_2 * nq * \frac{(avdl-dl)}{(avdl+dl)} \tag{1}$$

where w is the weight of a query term, N is the number of indexed documents in the collection, n is the number of documents containing a specific term, R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, qtf is within-query term frequency, dl is the length of the document, $avdl$ is the average document length, nq is the number of query terms, the k_is are tuning constants (which depend on the database and possibly on the nature of the queries and are empirically determined), K equals to $k_1 * ((1-b) + b * dl/avdl)$, and $\oplus$ indicates that its following component is added only once per document, rather than for each term.

2.2 Concept-based Method

In order to take medical concept information into account, we use MeSH [1], a very popular medical vocabulary, to identify concepts. Since it takes a long time to recognize all the concepts in the whole collection, we just try to find concepts in the topics. Then, if a document doesn't contain the complete medical concepts in the corresponding topic, it will be removed from the retrieved document list. For instance, if a topic contains a medical concept "heart disease", a document is not acceptable if "heart disease" doesn't appear in it. This document may be in the

[1] http://www.nlm.nih.gov/mesh/

retrieved list obtained by BM25 and have "heart" or "disease" separately, but it will be filtered and not in the final result.

The second run YorkUMC2 is obtained based on YorkUMB1. In the first run, each document is ranked by its BM25 score. But in YorkUMC2, if a document does not contain the complete medical concepts in the topic, it will not appear in the final result no matter how much score it gets. This is a very strict condition. The purpose of this post-process is to see whether doctors are always eager to use a particular medical concept in all the clinic reports. If so, it is easier for researcher to find relevant document in the clinic report search because the matching process is simpler and clearer.

2.3 A Hybrid Retrieval Model

For the 3rd and 4th runs, we evaluate a recently proposed hybrid retrieval model [8], which has shown to be very effective on a large number of TREC datasets for ad hoc information retrieval.

In particular, this hybrid model extends the Rocchio's feedback method by incorporating three kinds of IR techniques, which are proximity, feedback document quality estimation and query performance prediction techniques, under the pseudo relevance feedback (PRF) framework to boost the overall performance. In our experiments, we test different setting of this hybrid model on the medical dataset. In the rest of this section, we briefly describe this hybrid model.

Rocchio's algorithm [4] is a classic framework for implementing (pseudo) relevance feedback via improving the query representation. Although the Rocchio's model has been introduced for many years, it is still effective in obtaining relevant documents. According to [9], "BM25 term weighting coupled with Rocchio feedback remains a strong baseline which is at least as competitive as any language modeling approach for many tasks". The formula of the Rocchio's model is as follows:

$$Q_1 = \alpha * Q_0 + \beta * \sum_{r \in R} \frac{r}{|R|} \tag{2}$$

where Q_0 and Q_1 represent the original and first iteration query vectors, r is the expansion term weight vector, and α and β are tuning constants controlling how much we rely on the original query and the feedback information.

However, the traditional Rocchio's model can still be reformed to be better. First, the query term proximity information which has proven to be useful is not considered. Second, Rocchio's algorithm views terms from different feedback documents equally. Intuitively, a candidate expansion term in a document with better quality is more likely to be relevant to the query topic. Third, the interpolation parameter α is always fixed across a group of queries.

In order to address these problems, Ye et al. [8] extend Rocchio's algorithm by refining the query representation as follows.

$$Q_1 = \alpha * (\beta * Q_0 + (1 - \beta) * Q_p) + (1 - \alpha) * \sum_{r \in R} \frac{r * q(d_r)}{|R|} \tag{3}$$

where β controls how much we rely on the query term proximity information [6], α controls how much we rely on the original query, Q_p is an n-gram of original query terms and $q(d_r)$ is the quality score of document d.

As we can see from Equation 3, this hybrid model is very flexible and can evaluate different techniques. In our experiments, we adopt the co-occurrence interpretation of term proximity to compute Q_p, where the proximity among query terms is represented by the n-gram frequencies and BM25 is used as the weighting model [1]. Full dependencies of query terms are taken into account. For the document quality factor $q(d_r)$, we simple use the normalized scores from the first-pass retrieval for approximation as describe in [7]. For the term weighting formula in the query expansion component, we simply use the Lemur TFIDF formula, which was shown to be surprisingly effective on a number of standard TREC collections in our preliminary experiments. In our submissions, we did not use the proximity model in run 3, while all the components were used in run 4 with the same parameter settings.

3 Experiments

We use Porter stemming and stopword removal to preprocess the original dataset when indexing all the documents. Only the contents in the report_text field are indexed. In all our four runs, we retrieve 2500 reports for each topic, and then map these reports to their according visits. When mapping the retrieved clinic reports to visits, if more than one report of a visit are retrieved, we choose the top ranked one to represent the rank of this visit. Only top 1000 visits are retained if more are returned.

For the first run, YorkUMB1, we set the parameter b to 0.75, k1 to 1.2, k2 to 0 and k3 to 8. To obtain YorkUMC2, we first get the same document list as in YorkUMB1 with the same settings for BM25, and then filtered documents which do not contain the complete medical concepts in the according topics. For YorkUMQ3, we empirically set α to 0.6 and b in BM25 to 0.3. To combine the proximity information with the extended Rocchio's framework, we set β to be 0.2 in run YorkUMP4. We obtain YorkUMQ3 and YorkUMP4 by using pesudo relevance feedback with 10 feedback documents and 30 feedback terms. The performance of each run is shown in Table 3.

Table 2: Results

Official runs	**P@10**	**R-prec**	**infAP**	**infNDCG**
YorkUMB1	0.4340	0.2895	0.1716	0.3982
YorkUMC2	0.4149	0.2647	0.1574	0.3810
YorkUMQ3	0.4787	0.3284	0.2044	0.4634
YorkUMP4	0.4979	0.3407	0.2127	0.5552

We use different b values for run1, 2 (0.75) and run 3, 4 (0.3) for the BM25 model. This can conduct some differences in the performance. However, the differences will not be extensive according to our previous research. So it is still reasonable to compare all the four runs while we do not have the golden standard yet.

In the last year, BM25 performed well on the topics from 101 to 135 which prove that it can provide a strong baseline. This is also the reason that we use it to obtain the first run. In the medical search domain, medical concepts are always considered very important because they can help to retrieve documents more accurately. Some teams identified these concepts, indexed them and conducted concept-based search in the last year. But it is not easy to apply concepts in the medical search. The first thing is its efficiency. In order to implement a concept matching method, researchers have to convert both the queries and the medical corpora into concepts. To this end, some powerful tools such as MetaMap [2], OpenNLP [3] or Biolabeler [4] are widely used [5]. Generally, these tools segment a medical document into sentences, then phrases and finally identify medical concepts from these phrases or terms. Natural language processing (NLP) technologies are extensively used in these tools. For example, MetaMap use chunking technology to slice a sentence into phrases. Meanwhile, part-of-speech tagging methods are also used to pick out potent concepts from phrases and identify them in plenty of medical vocabularies by applying word sense disambiguation.

In order to simplify the scenario, we just identify medical concepts in the topics instead of the whole collection, and use them to filter documents which do not contain them. Our intent is to verify whether topic terms are always appear as medical concepts in relevant documents. As we can see from Table 3, the performance of the second run is not so good as the baseline even on the P@10 metric. A possible reason is that topic terms can express the same meaning even if they are not in a concept format. Also, this restriction indeed decreases the retrieved document list extensively. Some relevant documents are removed which is possible to lower the final performance. To make use of concept better, we need a more careful and elaborate way to investigate what is the correct scenario to use medical concepts.

[2]http://metamap.nlm.nih.gov/
[3]http://opennlp.apache.org/
[4]http://www.biolabeler.com/bioLabeler/

The extended Rocchio's framework has prove its effectiveness on the web datasets in our previous work. We are glad to see that it also performs well on domain-based datasets like the medical collection. Both YorkUMQ3 and YorkUMP4 outperform YorkUMB1 significantly. The improvement is more than 10% on all the four evaluation metrics. YorkUMP4 is obviously better than YorkUMQ3 which indicates that the proximity information can help to improve the overall performance not only on general collections but also the medical dataset.

4 Conclusions

The TREC Medical Records Track presents a challenging ad-hoc retrieval task where the focus is on the reasoning part of the system. In this paper, we present our participation in the Medical Records Track of TREC 2012. In this year, we submit four runs which were obtained by applying different information retrieval techniques. The performance of the baseline run obtained by BM25 is decent. However, when we utilize medical concepts in the topics, the performance drops. This indicates that we need to find a more elaborated way to make full use of these concepts. Based on our previous work, we apply the extended Rocchio's framework on the medical dataset and generate another two runs, YorkUMQ3 and YorkUMP4. Both of them outperform the BM25 run significantly. YorkUMP4 is better than YorkUMQ3 while the proximity information of topic terms are taken into account. The techniques which are effective on the general datasets can also perform well on the dataset of a particular domain, e.g., the medical domain.

Future work will focus on exploring different strategies for identifying the most relevant disease synonym to append the query where we plan to consider how to utilize the relationships in the medical ontology tree like MeSH. Also, we plan to use some external resources to reformulate the original topics so that the search intents can be expressed more clearly.

5 Acknowledgements

This research is jointly supported by NSERC of Canada and the Early Researcher/Premier's Research Excellence Award.

References

[1] Ben He, Jimmy Xiangji Huang, and Xiaofeng Zhou. Modeling term proximity for probabilistic information retrieval models. *Information Sciences*, 181(14):3017–3031, 2011.

[2] X. Huang and S. E. Robertson. Application of probabilistic methods to chinese. *Journal of Documentation*, 53(1):74–79, 1997.

[3] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *The Forth Text REtrieval Conference (TREC-4)*, 1996.

[4] J. Rocchio. *Relevance feedback in information retrieval*, pages 313–323. Prentice-Hall Englewood Cliffs, 1971.

[5] Kirk Roberts Sanda M. Harabagiu Travis Goodwin, Bryan Rink. Cohort shepherd: Discovering cohort traits from hospital visits. In *TREC 2011 Proceedings*, 2011.

[6] C. J. van Rijsbergen. A theoretical basis for the use of co-occurence data in information retrieval. *Journal of Documentation*, 1977.

[7] Zheng Ye, Ben He, Xiangji Huang, and Hongfei Lin. Revisiting rocchio's relevance feedback algorithm for probabilistic models. In *AAIRS*, pages 151–161, 2010.

[8] Zheng Ye, Jimmy Xiangji Huang, and Jun Miao. A hybrid model for ad-hoc information retrieval. In William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *SIGIR*, pages 1025–1026. ACM, 2012.

[9] ChengXiang Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2:137–213, 2008.