An Exploration of New Ranking Strategies for Medical Record Tracks

Hao Wu and Hui Fang

University of Delaware, Newark DE 19716, USA
haowu@ece.udel.edu, hfang@ece.udel.edu

Abstract

We report our system and experiments at the 2011 Medical Record Track. Our goal is to return most relevant visits according to a query. In particular, we start with an axiomatic retrieval, and combine it with an aspect based term proximity strategy to improve the retrieval performance. We also propose a "disease-diversity" strategy based on the assumption that most of documents only contain information related to one main disease. Query expansion using external resources has also been studied.

1 Introduction

In this year's medical record track, the de-identified medical records from University of Pittsburgh are used as test document collection. It contains about 95,000 text electronic medical records from17,267 visits. A visit may include one or more than one records and the belonging relationship is represented in a simple ASCII table called the Report-to-Visit Mapping Key. One visit is a retrieval unit and the developed systems are required to return a rank list of visits that are relevant to user's information needs.

Symptoms and disease names are the key factors to achieve satisfying search performance. Therefore, we explored several strategies that can make use of these names. Since every symptom and disease name contains multiple terms, it is necessary to detect these names from a keyword query and treat them as an aspect rather than individual terms. To tackle this problem, we used a simple IDF-based method. Once the aspects are identified, we incorporate the proximity among terms within aspects into the retrieval function to improve retrieval performance.

Moreover, we also examined the effectiveness of query expansion using external source, i.e., Healthline.com. In particular, we start with the diseases and symptoms, find related pages and use the descriptions on the pages to find terms for expansion. Intuitively, the visits that contain more information about the disease and symptoms are more likely to receive higher scores.

Finally, we also observe that a visit is more likely to talk about only one disease. Thus we try to classify the visits by the diseases they are most likely about and develop a strategy to punish visit in which query diseases are not its main diseases.

The rest of the paper is organized as follows. We first explained the pre-processing procedure in Section 2, and then describe our methods in Section 3. Our submitted runs and experiment results are showed in Section 4 and we conclude in Section 5.

2 Preprocessing and Index Building

The task requires us to return visit as basic unit, however the data collection is store in individual records. Thus, we merge records which belong to the same visit into a document based on the official report-to-visit mapping key.

We built the index with the Lemur toolkit. Porter stemmer is applied and no stop words have been removed. Each visit is treated as one document and it contains the information from records related to the visit.

3 Method

3.1 Baseline

We use the F2-exp function [1] as our retrieval baseline:

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

where s and k are parameters. In our experiments, we set s=0.5 and k=0.35

3.2 Aspect-based term proximity

Traditional term proximity considers the distance between all the query terms. However, the importance of each query term pair distances should not be equal. More specifically, the term distances between terms within a same aspect (e.g. a disease or drug name) are more important than those between terms from different aspect.

Based on this ideal, we built up following retrieval function to only calculate term proximity within aspects:

$$S = S(Q,D) + log\left(0.1 + \sum_{As \in Q} \sum_{\substack{t1,t2 \in As \\ t1,t2 \in D}} \exp\left(-\min(dis(t1,t2,D))\right)\right)$$

where As is an aspect, dis(t1,t2,D) is the term distance between term t1 and term t2 in document D

In medical record search, the verbal query is usually consisted with a serial of key terms (e.g. disease or symptom names) and less important terms (e.g. “with”, “in”, “treated”). The

key terms compose aspects and they usually have higher IDF, while the less important terms divide aspects and they are usually with lower IDF.

Based on this observation, we simply judge the terms with an IDF threshold. That is if the term IDF is higher than the threshold, it will be judged as key term. Otherwise it will be judged as less important term. Key terms which are not divided by less important term will be considered as one aspect.

3.3 Query expansion

In medical record search, the disease and symptom names are key parts of each query. To highlight such names, we use a disease or symptom list crawled from "Healthline.com" to identify such names. Once a disease or symptom name is identified, all the terms from that description fields will be chosen as candidate. And then top 30 with highest scores are chosen to expanse the original query with normalized weight calculated by their scores.

3.4 Disease diversity

Based on our observation, one visit usually mainly talks about one single disease. Thus we can use this main disease to identify relevant documents: the visit which has query disease as its main disease is more likely to be relevant than that does not. We use the following function to implement such idea:

$$S = S(Q, D) + \beta S(D, dis)S(dis, Q)$$

In which S(D,dis) is the similarity between the visit and its main disease, while S(dis,Q) is the similarity between the main disease and the query.

4 Experiments and results

We submitted 5 runs with the method we discuss before:
UDMedBL: Use the baseline function
UDMedProx: Use the baseline function + aspect-based term proximity
UDMedExp: Expanse the original function with the information crawled from "Healthline.com"
UDMedComb: The baseline function + aspect-based term proximity + query expansion
UDMedDiv: the baseline function + disease diversity

The results are shown as following:

	MAP	bprel	R-precision	precision@10
UDMedBL	0.3457	0.4523	0.3589	0.5059
UDMedProx	0.3539	0.4574	0.3706	0.5059
UDMedExp	0.2887	0.4154	0.3010	0.4618
UDMedComb	0.2887	0.4154	0.3010	0.4618
UDMedDiv	0.3284	0.4390	0.3513	0.4853

5 Conclusion:

The results show that the aspect-based term proximity is effective to improve the retrieval performance. We will run more experiments and analyze the results to see whether other methods could be effective with different parameter settings.

Reference

[1]Hui Fang and ChengXiang Zhai. 2005. 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).