University of Glasgow at CLEF 2013: Experiments in eHealth Task 3 with Terrier

Nut Limsopatham[1], Craig Macdonald[2], and Iadh Ounis[2]

School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK
nutli@dcs.gla.ac.uk[1], firstname.lastname@glasgow.ac.uk[2]

Abstract. In our participation in the CLEF 2013 eHealth task 3, we investigate (1) the effectiveness of our Divergence from Randomness (DFR) framework on retrieving medical webpages, (2) the adoption of classical pseudo-relevance feedback for improving the representation of the queries, and (3) the exploitation of a collection enrichment technique for alleviating the mismatches between the terms in documents and queries, all within the context of our Terrier information retrieval platform.

Keywords: Pseudo-Relevance Feedback, Collection Enrichment

1 Introduction

The CLEF 2013 eHealth task 3 [13] developed a novel framework for evaluating search systems that retrieve medical web documents relevant to a query, as would be issued by patients looking to find information related to their discharge summary [13]. In our participation in the CLEF 2013 eHealth task 3, we aim to evaluate the effectiveness of classical approaches existing the Terrier platform[1] [11], which have been shown to be effective for other search tasks (e.g. web search, blog search and medical records search), on this medical web document retrieval task. In particular, building upon the effective Divergence from Randomness (DFR) framework [1], our participation has three major objectives:

1. We deploy the parameter-free DPH model [2], which has been shown to be effective for retrieval tasks (e.g. web search [12] and medical records search [7]), to the task of ranking medical web documents.
2. We investigate the effectiveness of using classical pseudo-relevance feedback (PRF) to lessen the mismatches between terms in the medical web documents and the queries.
3. We use a collection enrichment technique to improve the representation of the queries using information from different corpora, including Wikipedia and MEDLINE abstracts.

```
<query>
<id>qtest1</id>
<discharge_summary>00098-016139-DISCHARGE_SUMMARY.txt</discharge_summary>
<title>Hypothyreoidism</title>
<desc>What is hypothyreoidism</desc>
<narr>description of what type of disease hypothyreoidism is</narr>
<profile>
A forty year old woman, who seeks information about her condition
</profile>
</query>
```

Fig. 1. Query#1 from the ShARe/CLEF eHealth task 3 topics

This paper is organised as follows. In Section 2, we briefly describe the task and the document collection. Section 3 discusses the parameter-free DPH weighting model to rank medical web documents. Sections 4 and 5 explain our deployment of PRF and collection enrichment, respectively, to further improve retrieval performance. Runs and results are presented in Section 6, and the conclusions are discussed in Section 7.

2 Medical Web Search

Along with the growth of the Internet and the Web, a phenomenal expansion of Web-based medical document collections have been witnessed in the recent years. Online digital libraries, such as PubMed, provide comprehensive literature and teaching material on biomedical issues. Moreover, the number of publicly available websites that provide information about healthcare and treatment (e.g. `http://www.patientslikeme.com/`, `http://www.webmd.com/`) have been increasing. Furthermore, the number of users using search engines to search for information related to personal health has been growing. Hersh [4] reported that 80% of search engine users have searched for websites or documents related to their health condition, while about 98% of US physicians use the Internet to find documents related to healthcare. Searchers of these medical web collections are desirable to retrieve documents pertaining to a specific medical scenario [10]. For example, patients may search the Internet for an explanation of their diagnosed disease in order to understand their health condition better.

To facilitate these phenomenon, the ShARe/CLEF eHealth Evaluation Lab [13] introduced a standard framework for evaluating medical webpage search systems (Task 3) in 2013. Specifically, the aim of the task is to retrieve medical web documents that can answer patients queries about their disorders, after they have examined their discharge summary. The queries are the representative of real patient information needs after reading their discharge summary. Figure 1 shows an example of the ShARe/CLEF eHealth task 3 queries. Indeed, each query contains three different levels of details of an information need in the different tags

[1] `http://terrier.org`

(i.e. *title*, *desc* and *narr*). In addition, the reference to the original discharge summary is also provided (i.e. *discharge_summary* tag). The description of the search can be obtained from the *profile* tag. The collection consists of 1.2 M. webpages from online medical resources, including Health On the Net Foundation-certified websites and other well-known medical websites (e.g. Genetics Home Reference).

3 The Effective Parameter-free DPH Term Weight Model

As a representation of a classical ranking approach, we apply the DPH [2] hypergeometric parameter-free document weighting model to rank medical web documents, since it has been shown to be effective in both web search and medical records search (e.g. [7, 8, 12]). DPH is a weighting model from the Divergence from Randomness (DFR) [1] framework, which calculates the score for a document d as follows [2]:

$$score_{DPH}(d,Q) = \sum_{t \in Q} tfq \cdot norm \cdot \Bigg(tfd \cdot \log((tfd \cdot \frac{avg\_dl}{dl}) \cdot (\frac{N}{tfc})) + 0.5 \cdot \log(2 \cdot \pi \cdot tfd \cdot (1-f)) \Bigg) \quad (1)$$

where tfq is the frequency of term t in the query Q, tfd is the frequency of term t in document d, N is the number of documents in the collection, tfc is the frequency of term t in the collection, avg_dl is the average length of documents in the collection, dl is the length of the document d, $f = \frac{tfd}{dl}$, and $norm = \frac{(1-f)^2}{(tf+1)}$.

4 Pseudo-Relevance Feedback using the Bo1 Model

Medical terminology is known to be inconsistent [9] (e.g. practitioners can use different terms to refer to a particular medical term). In order to improve the representation of a given query, we deploy the parameter-free Bose-Einstein statistics-based (Bo1) model from the Divergence from Randomness (DFR) framework [1] to expand the query with informative terms from the pseudo-relevance documents (i.e. top-ranked documents). Specifically, the Bo1 model calculates the weight of terms, as followings [1]:

$$w(t) = tf_x \cdot log_2 \frac{1 + P_n(t)}{P_n(t)} + log_2(1 + P_n(t)) \quad (2)$$

$$P_n(t) = \frac{tfc}{N} \quad (3)$$

where tf_x 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. Following Amati [1], we extract the 10 most informative terms (i.e. terms having highest $w(t)$ scores) from the top 3 retrieved documents to

reformulate the query. The original query terms can also appear in the 10 extracted terms. Then, the query term weight qtw of each expanded query term can be calculated as [1]:

$$qtw(t) = \frac{qtf}{qtf_{max}} + \frac{w(t)}{lim_{F \to tfx} w(t)} \tag{4}$$
$$= F_{max} \cdot log_2 \frac{1 + P_{n,max}}{P_{n,max}} + log_2(1 + P_{n,max})$$

$$P_{n,max} = \frac{F_{max}}{N} \tag{5}$$

where $lim_{F \to tfx}$ is the upper bound of $w(t)$, F_{max} is the frequency F of the term with the maximum $w(t)$ in the top-ranked documents. If an original query term t does not appear in the most informative terms extracted from the top-ranked documents, its query term weight qtw remains equal to the original one.

5 Collection Enrichment

To further improve the representation of the queries, we deploy collection enrichment (CE) [3], which has shown to be effective for medical records search [6, 14]. Indeed, the collection enrichment aims to expand a query with informative terms extracted from an external corpus, by deploying a query expansion technique. Intuitively, collection enrichment should alleviate the mismatch between terms in the relevant documents and a given query, since informative terms may be extracted from an external corpus. In this work, we use the DFR Bo1 model [1] (Equation (4) in Section 4) to expand the queries with the top 10 informative terms from the top 3 ranked documents retrieved from two different external corpora, namely Wikipedia 2008 and the MEDLINE abstract collection of the TREC 2005 Genomics track [5], respectively. Specifically, the Wikipedia collection contains 3,588,998 Wikipedia pages, while the MEDLINE abstract collection consists of 4,591,008 documents.

6 Runs and Results

We perform all runs using the Terrier retrieval platform [11][2], applying Porter's English stemmer and removing stopwords. For each topic, we use only the terms in the *title* tag (see Figure 1) as the query terms, since web search engine users normally use a very few terms as a query. Indeed, we submitted 4 title-only runs, as followings:

1. **uogTr.1.3.noadd**: A baseline run, which applies the effective parameter-free DPH weighting model discussed in Section 3. All other submitted runs build upon this run.

[2] `http://terrier.org`

Table 1. Comparing the results of the submitted runs, in terms of P@10 and NDCG@10

Run	P@10	NDCG@10
uogTr.1.3.noadd (DPH)	0.4360	0.3826
uogTr.5.3.noadd (DPH + PRF)	**0.4400**	**0.3858**
uogTr.6.3.noadd (DPH + PRF + CE:MEDLINE abstracts)	0.4040	0.3536
uogTr.7.3.noadd (DPH + PRF + CE:Wikipedia)	0.3500	0.3220

2. **uogTr.5.3.noadd**: This run improves the representation of the queries using PRF. Indeed, the Bo1 model is deployed to expand the queries with informative terms extracted from the corpus, as discussed in Section 4.
3. **uogTr.6.3.noadd**: This run further enhances the query representation using both PRF and the collection enrichment, introduced in Sections 4 and 5, respectively. Specifically, the MEDLINE abstract collection is used for the collection enrichment.
4. **uogTr.7.3.noadd**: This run deploys the same approaches as the previous run (**uogTr.6.3.noadd**); however, Wikipedia collection is used for the collection enrichment, instead of the MEDLINE abstract collection.

Table 1 compares the results of our submitted runs, in terms of P@10 and NDCG@10, which are the official measures of CLEF eHealth 2013 task 3 [13]. Firstly, we find that our baseline (uogTr.1.3.noadd), which deploys only a classical ranking model (the DFR DPH model, discussed in Section 3), achieves 0.4360 and 0.3826 in term of P@10 and NDCG@10, respectively. Secondly, we observe that PRF using the Bo1 model (see Section 4) further improves the retrieval performance for both official measures. However, the combination of both PRF and collection enrichment (see Section 5) to enhance query representation appears to be difficult, as neither uogTr.6.3.noadd nor uogTr.7.3.noadd outperforms the baseline. In particular, we observe topic drift to be occurring, where the expanded terms are not relevant to the query. For example, for query #qtest6: 'dysplasia and multiple sclerosis', run uogTr.7.3.noadd leverages the Wikipedia collection to expand the query with general terms such as syndrome, type, and disease, thereby excessively changing the focus of the query.

7 Conclusions

For our participation in the CLEF eHealth 2013 task 3, we focus on examining the effectiveness of classical approaches in the searching of medical web documents. Specifically, using Terrier, we evaluate the performances of the DFR DPH weighting model as well as PRF and collection enrichment using the Bo1 model. Overall, we find that PRF helps to improve retrieval effectiveness. However, we observe the combining both PRF and the collection enrichment to further improve retrieval performance remains an open problem.

References

1. G. Amati.: Probability models for information retrieval based on Divergence From Randomness. PhD thesis, University of Glasgow, (2003)
2. G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi.: FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track. In: TREC (2007)
3. F. Diaz and D. Metzler.: Improving the estimation of relevance models using large external corpora. In: SIGIR (2006)
4. W. Hersh.: Ubiquitous but unfinished: grand challenges for information retrieval. In: Health Information and Libraries Journal, 25 (2008)
5. W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, and M. Hearst.: TREC 2005 Genomics Track Overview. In: TREC (2005)
6. N. Limsopatham, C. Macdonald, and I. Ounis.: Inferring Conceptual Relationships to Improve Medical Records Search. In: OAIR (2013)
7. N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, M. M. Bouamrane.: University of Glasgow at Medical Records Track 2011: Experiments with Terrier. In: TREC (2011)
8. N. Limsopatham, C. Macdonald, R. McCreadie, and I. Ounis.: Exploiting Term Dependence while Handling Negation in Medical Search. In: SIGIR (2012)
9. N. Limsopatham, R. L. T. Santos, C. Macdonald, and I. Ounis.: Disambiguating biomedical acronyms using EMIM. In: SIGIR (2011)
10. Z. Liu and W. Chu.: Knowledge-based query expansion to support scenario-specific retrieval of medical free text. In: Information Retrieval, 10(2) (2007)
11. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: OSIR at SIGIR (2006)
12. R. L. T. Santos, R. McCreadie, C. Macdonald, and I. Ounis.: University of Glasgow at TREC 2010: Experiments with Terrier in Blog and Web tracks In: TREC (2010)
13. H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, D. Mowery, J. Leveling, L. Goeuriot, L. Kelly, D. Martinez, and G. Zuccon.: ShARe/CLEF eHealth Evaluation Lab 2013: Three Shared Tasks on Natural Language Processing and Machine Learning to Make Clinical Reports Easier to Understand for Patients. In: CLEF (2013)
14. D. Zhu and B. Carterette.: Combining Multi-level Evidence for Medical Record Retrieval. In: SHB at CIKM (2012)