ub-botswana participation to CLEF eHealth IR challenge 2017: Task 3 (IRTask1 : ad-hoc search)

Edwin Thuma, Nkwebi Motlogelwa, and Tebo Leburu-Dingalo

Department of Computer Science, University of Botswana
{thumae,motlogel,leburut}@mopipi.ub.bw

Abstract. In this paper, we describe the methods deployed in the different runs submitted for our participation to the CLEF eHealth 2017 Task 3: Patient-Centered Information Retrieval, IRTask 1: ad-hoc search. Specifically, we deploy DPH term weighting model with explicit relevance feedback, where the expansion terms are selected from documents which were previously identified as relevant by assessors for each query. As improvement we deployed proximity search using both Full Dependence (FD) and Sequential Dependence (SD) variants of the Markov Random Fields and the Divergence From Randomness (DFR) based dependence models to re-rank documents, which have query terms in close proximity. In another approach, we deploy pseudo relevance feedback, where the expansion terms are selected from the top 3 ranked documents after a first pass retrieval. In addition, we deploy proximity search using the SD variant of the DFR based dependence model.

Keywords: Explicit relevance Feedback, Proximity Search, Pseudo Relevance Feedback

1 Introduction

In this paper, we describe the methods used for our participation to the CLEF eHealth 2017 Task 3: Patient-Centered Information Retrieval, IRTask 1: adhoc search. Detailed task description is available in the overview paper of Task 3 [7]. This task is a continuation of the previous CLEF eHealth Information Retrieval (IR) task that ran in 2013 [3], 2014 [4], 2015 [6] and 2016 [5]. The CLEF eHealth task aims to evaluate the effectiveness of information retrieval systems when searching for health related content on the web, with the objective to foster research and development of search engines tailored to health information seeking [6, 5]. The CLEF eHealth Information Retrieval task was motivated by the problem of users of information retrieval systems formulating circumlocutory queries, using colloquial language instead of medical terms as studied by Zuccon et al. [9] and Stanton et al. [8]. In their studies, they found that modern search engines are ill-equipped to handle such queries; only 3 out of the to 10 results were highly useful for self diagnosis. In this paper, we attempt to tackle this problem by using explicit relevance feedback in order to improve the retrieval effectiveness. In addition, we deploy proximity search to further improve the retrieval effectiveness of our system. Moreover, we investigate whether pseudo relevance feedback, where the expansion terms are selected from the top 3 ranked documents after a first pass retrieval can improve the retrieval effectiveness. This paper is structured as follows. Section 2 contains a background on algorithms used. Section 3 describes the experimental environment. In Section 4, we describe the experimental the 5 runs submitted by team ub-botswana. Section 5 presents and discusses results on training data.

2 Background

In this section, we begin by presenting a brief but essential background on the different algorithms used in our experimental investigation and evaluation. We start describing the DPH term weighting model in Section 2.1. We then describe the Bose-Einstein 1 (Bo1) model for query expansion in Section 2.2.

2.1 DPH Term Weighting Model

For all our experimental investigation and evaluation we used the parameter-free DPH term weighting model from the Divergence from Randomness (DFR) framework [2]. The DPH term weighting model calculates the score of a document d for a given query Q as follows:

$$score_{DPH}(d,Q) = \sum_{t \in Q} qtf \cdot norm \cdot \left(tf \cdot \log((tf \cdot \frac{avg\_l}{l}) \cdot (\frac{N}{tfc})) + 0.5 \cdot \log(2 \cdot \pi \cdot tf \cdot (1 - t_{MLE})) \right) \quad (1)$$

where qtf, tf and tfc are the frequencies of the term t in the query Q , in the document d and in the collection C respectively. N is number of documents in the collection C, avg_l is the average length of documents in the collection C and l is the length of the document d. $t_{MLE} = \frac{tf}{l}$ and $norm = \frac{(1-t_{MLE})^2}{tf+1}$.

2.2 Bose-Einstein 1 (Bo1) Model for Query Expansion

In our experimental investiagtion and evaluation, we used the Terrier-4.0 Divergence from Randomness (DFR) Bose-Einstein 1 (Bo1) model to select the most informative terms from the topmost documents after a first pass document ranking. The DFR Bo1 model calculates the information content of a term t in the top-ranked documents as follows [1]:

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

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

where $P_n(t)$ is the probability of t in the whole collection, tfx is the frequency of the query term in the top x ranked documents, tfc is the frequency of the term t in the collection, and N is the number of documents in the collection.

3 Experimental Setting

FAQ Retrieval Platform: For all our experimental evaluation, we used Terrier-4.2, an open source Information Retrieval (IR) platform. All the documents (ClueWeb 12 B13) used in this study were first pre-processed before indexing and this involved tokenising the text and stemming each token using the full Porter stemming algorithm. Stopword removal was enabled and we used Terrier stopword list. The index was created using blocks to save positional information with each term. For pseudo relevance feedback, we used Terrier-4.2 DFR Bose-Einstein 1 (Bo1) model for query expansion to select the 10 most informative terms from the top 3 ranked documents.

4 Description of the Different Runs

Term Weighting Model: For all our runs, we used the parameter-free DPH Divergence From Randomness term weighting model in Terrier-4.2 IR platform to score and rank the documents in the ClueWeb 12 B13 document collection.

ub-botswana_IRTask1_run1: We ranked the documents using DPH DFR term weighting. As improvement, we deployed explicit relevance feedback, where we selected expansion terms from the top 3 documents that were explicitly marked relevant by assessors for each query. We used the Terrier-4.2 DFR Bose-Einstein 1 (Bo1) model for query expansion to select the 10 most informative terms from these documents. In addition, we deployed the Full Dependence (FD) variant of the Markov Random Fields for terms dependence. Full Dependence assumes all query terms are in some way dependent on each other. In this work, we experimentally selected a window size of 15, which yielded the highest retrieval performance on the training data.

ub-botswana_IRTask1_run2: We performed a first pass retrieval using DPH DFR term weighting model. As improvement, we deployed explicit relevance feedback, where we deployed DFR Bo1 model for query expansion to select the expansion terms.

ub-botswana_IRTask1_run3: We produced an initial ranking using DPH DFR term weighting. As improvement, we deployed explicit relevance feedback and used the DFR Bo1 model for query expansion to select the expansion terms. In addition, we deployed the Sequential Dependence (SD) variant of the Divergence from Randomness based dependence model. Sequential Dependence only assumes a dependence between neighbouring query terms. In this work, we experimentally selected a window size of 15, which yielded the highest retrieval performance on the training data.

ub-botswana_IRTask1_run4: We used the parameter-free DPH DFR term weighting model to produce and initial ranking. As improvement, we deployed a simple pseudo-relevance feedback on the local collection. We used the Bo1 model for query expansion to select the expansion terms. We then performed a second pass retrieval on the local collection with the new expanded query.

ub-botswana_IRTask1_run5: We used ub-botswana_IRTask1_run4 as the baseline system. As improvement, we deployed the Sequential Dependence (SD) variant of the Divergence from Randomness based term dependence model. Sequential Dependence only assumes a dependence between neighbouring query terms. In this work, we experimentally selected a window size of 15, which yielded the highest retrieval performance on the training data.

5 Results and Discussion

These working notes were compiled and submitted before the relevance judgments were released. Below we present the results of our runs using the 2016 query relevance judgments. Please note that the official results to be released will be different because new query relevance judgments will be released

Table 1. Retrieval Results for all 5 Runs using 2016 qrel

Run ID	P@5	P@10	rel_ret
DPH Baseline	0.2973	0.2710	10104
ub-botswana_IRTask1_run1	0.5093	0.4423	13661
ub-botswana_IRTask1_run2	0.4513	0.4097	13661
ub-botswana_IRTask1_run3	0.4433	0.4073	13661
ub-botswana_IRTask1_run4	0.3160	0.2903	11129
ub-botswana_IRTask1_run5	0.2873	0.2617	10104

Table 1 presents our results on the training data. From this table, we see a degradation in performance when we incorporate term dependence only in our ranking (*ub-botswana_IRTask1_run5*). However, when we deploy pseudo relevance feedback (*ub-botswana_IRTask1_run4*), we see an improvement in the retrieval performance in terms of precision at 5 (P@5), precision at 10 (P@10) and recall (rel_ret). Moreover significant improvement in the recall is obtained when explicit relevance feedback is deployed ((*ub-botswana_IRTask1_run1*), (*ub-botswana_IRTask1_run2*) and (*ub-botswana_IRTask1_run3*)). In addition, we obtain mixed results when we incorporate proximity search after deploying explicit relevance feedback. For example, there was an improvement in the retrieval performance in terms of P@5 and P@10 when we deploy the FD variant of the Markov Random Fields for term dependence using a window size of 15 (*ub-botswana_IRTask1_run1*). In contrast, we obtain a degradation in the retrieval performance in terms of P@5 and P@10 when we deploy the SD variant of the Divergence from Randomness based term dependence model using a window size of 15 (*ub-botswana_IRTask1_run5*).

References

1. G. Amati. Probabilistic Models for Information Retrieval based on Divergence from Randomness. *University of Glasgow,UK, PhD Thesis*, pages 1 – 198, June 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 *Proceedings of the 16th Text REtrieval Conference (TREC-2007)*, pages 1–10, Gaithersburg, Md., USA., 2007. Text REtrieval Conference (TREC).
3. L. Goeuriot, G.J.F Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salantera, H. Suominen, and G. Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information Retrieval to Address Patients' Questions when Reading Clinical Reports. In *CLEF 2013 Online Working Notes*, volume 8138. CEUR-WS, 2013.
4. L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G.J.F Jones, and H. Mueller. Share/clef ehealth Evaluation Lab 2014, Task 3: User-Centred Health Information Retrieval. In *CLEF 2014 Online Working Notes*. CEUR-WS, 2014.
5. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, João Palotti, and Guido Zuccon. *Overview of the CLEF eHealth Evaluation Lab 2016*, pages 255–266. Springer International Publishing, Cham, 2016.
6. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G.J.F. Jones, M. Lupu, and P. Pecina. CLEF eHealth Evaluation Lab 2015 task 2: Retrieving Information about Medical Symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS, 2015.
7. J. Palotti, G. Zuccon, Jimmy, P. Pecina, M. Lupu, L. Goeuriot, L. Kelly, and A. Hanbury. CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In *In Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*. CEUR Workshop Proceedings, 2017.
8. I. Stanton, S. Ieong, and N. Mishra. Circumlocution in Diagnostic Medical Queries. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 133–142. ACM, 2014.
9. G. Zuccon, B. Koopman, and J. Palotti. Diagnose This If You Can: On the Effectiveness of Search Engines in Finding Medical Self-Diagnosis Information. In *Advances in Information Retrieval (ECIR 2015)*, pages 562–567. Springer, 2015.