Task 3: Patient-Centered Information Retrieval, IRTask 1: ad-hoc search - TEAM ub-botswana.

Edwin Thuma, Nkwebi Peace Motlogelwa, and Tebo Leburu-Dingalo

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

Abstract. This paper describes the methods used for our participation to the CLEF (Conference and Labs of the Evaluation Forum) eHealth 2016 Task 3: Patient-Centered Information Retrieval, IRTask 1: ad-hoc search. For this participation, we evaluate the effectiveness of three different retrieval strategies. In particular, we deploy PL2 with Boolean Fallback as our baseline system. In another approach, we deploy the collection enrichment approach, where the original query is expanded with additional terms from an external collection (collection not being searched). To deliver an effective ranking, we combine the first two rankers using data fusion techniques.

Keywords: Collection Enrichment, Data Fusion, Query Expansion

1 Introduction

In this paper, we describe the methods used for our participation to the CLEF eHealth 2016 Task 3: Patient-Centered Information Retrieval, IRTask 1: ad-hoc search. Detailed task description is available in the overview paper of Task 3 [5, 12]. This task is a continuation of the previous CLEF eHealth Information Retrieval (IR) task that ran in 2013 [2], 2014 [3] and 2015 [4]. 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 [2–4]. 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. [11] and Stanton et al. [10]. In their study, 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 query expansion to try to add medical terms to the original query in order to improve the retrieval effectiveness of such systems. In addition, we deploy data fusion techniques to combine multiple rankers in order to further improve the retrieval effectiveness.

This paper is structured as follows. Section 2 contains a background on algorithms used. Section 3 describes the 3 runs submitted by team ub-botswana. In Section 4, we describe the experimental environment. Section 5 reports our results.

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 by describing the PL2 term weighting model in Section 2.1. In Section 2.2, we describe the Bose-Einstein 1 (Bo1) model for query expansion, followed by a description of the CombSUM for data fusion in Section 2.3. A description of the runs is provided in Section 3.

2.1 PL2 Term Weighting Model

For our baseline system and all our experimental investigation and evaluation, we used the PL2 term weighting model to score and rank medical documents. For a given query Q, the relevance score of a document d based on the PL2 Divergence from Randomness (DFR) term weighting model is expressed as follows [8]:

$$score_{PL2}(d,Q) = \sum_{t \in Q} \frac{qtfn}{tfn+1} \left(tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn) \right) \quad (1)$$

where $score(d,Q)$ is the relevance score of a document d for a given query Q. $\lambda = \frac{tfc}{N}$ is the mean and variance of a Poisson distribution, tfc is the frequency of the term t in the collection C while N is the number of documents in the collection. The normalised query term frequency is given by $qtfn = \frac{qtf}{qtf_{max}}$, where qtf_{max} is the maximum query term frequency among the query terms and qtf is the query term frequency. tfn is the Normalisation 2 of the term frequency tf of the term t in a document d and is expressed as:

$$tfn = tf \cdot \log_2 \left(1 + b \frac{avg\_l}{l} \right), (b > 0) \quad (2)$$

In the above expression, l is the length of the document d, avg_l is the average document length in the collection and b is a hyper-parameter.

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 on an external collection. The DFR Bo1 model calculates the information content of a term t in the top-ranked documents as follows [1]:

$$w(t) = tfx \cdot \log_2 \frac{1 + P_n(t)}{P_n(t)} + \log_2(1 + P_n(t)) \quad (3)$$

$$P_n(t) = \frac{tfc}{N} \quad (4)$$

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.

2.3 CombSUM

In another approach, we used data fusion to combine document rankings of two different rankers. In particular, we used CombSUM, which is a data fusion technique that sums the scores of each document in the constituent ranking based on the following equation:

$$score(d, Q) = \sum_{r \in R} score_r(d, Q) \tag{5}$$

where r is a ranking in R, R being the set of ranking being considered. $score_r(d, Q)$ is the score of document d for query Q in ranking r. If a document d is not in ranking r, the $score_r(d, Q) = 0$. Hence, a document scored highly in many rankings is likely to be scored highly in the final ranking. In contrast, a document with low scores, or that is present in less rankings is less likely to end up highly in the final ranking.

3 Description of the Different Runs

ub-botswana_EN_Run1: This is the baseline system. We used PL2 Divergence from Randomness term weighting model in Terrier-4.0 IR platform to score and rank the documents in the ClueWeb 12 B13 document collection. In order to improve the retrieval effectiveness of our system, we deployed a boolean fallback score modifier. With this score modifier, if any of the retrieved documents contain all undecorated query terms (ie query terms without any operators), then we remove from the result set documents that do not contain all undecorated query terms. Otherwise, we do nothing. The intuition is that when we combine this ranker with another ranker using any data fusion technique, documents retrieved and ranked by this score modifier are likely to be ranked higher when they appear in both rankings.

ub-botswana_EN_Run2: We used the baseline system without boolean fallback. As improvement, we used the collection enrichment approach [6], where we selected the expansion terms from an external collection, which was made up of the CLEF 2015 eHealth dataset. We used the Terrier-4.0 Divergence from Randomness (DRF) Bose - Einstein 1 (Bo1) model for query expansion to select the 10 most informative terms from the top 3 ranked documents after the first pass retrieval (on the external collection). We then performed a second pass retrieval on the local collection (ClueWeb 12 B13) with the new expanded query.

ub-botswana_EN_Run3: As improvement to ub-botswana_EN_{Run1 and Run2}, we diployed a simple CombSUM data fusion technique to combine the rankings for aforementioned rankers.

4 Experimental Setting

FAQ Retrieval Platform: For all our experimental evaluation, we used Terrier-4.0[1][7], 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 [9]. Stopword removal was enabled and we used Terrier stopword list. The hyper-parameter for PL2 was set to its default value of b = 1.0.

5 Results

These working notes were compiled and submitted before the relevance judgments were released. However, results for our different runs and how our approaches performed as compared to other participating teams are presented in the Task 3: Patient-Centered Informational Retrieval overview paper entitled *"The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centered Health Information Retrieval"* [12].

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. 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.
3. 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.
4. L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth Evaluation Lab 2015. In *CLEF 2015 - 6th Conference and Labs of the Evaluation Forum*. Lecture Notes in Computer Science (LNCS), Springer, September 2015.
5. L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth Evaluation Lab 2016. In *CLEF 2016 - 7th Conference and Labs of the Evaluation Forum*. Lecture Notes in Computer Science (LNCS), Springer, September 2016.
6. K.L. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 250–256, New York, NY, USA, 1998. ACM.

[1] http://terrier.org/

7. I. Ounis, G. Amati, Plachouras V., B. He, C. Macdonald, and Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519, Berlin, Heidelberg, 2005. Springer-Verlag.
8. Vassilis Plachouras and Iadh Ounis. Multinomial Randomness Models for Retrieval with Document Fields. In *Proceedings of the 29th European Conference on IR Research*, pages 28–39, Berlin, Heidelberg, 2007. Springer-Verlag.
9. M.F. Porter. An Algorithm for Suffix Stripping. *Readings in Information Retrieval*, 14(3):313–316, 1997.
10. 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.
11. 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.
12. G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu, P. Pecina, H. Mueller, J. Budaher, and A. Deacon. The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval. In *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes*. CEUR-WS, September 2016.