Merging Search Results Generated by Multiple Query Variants Using Data Fusion

Nkwebi Motlogelwa, Tebo Leburu-Dingalo, and Edwin Thuma

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

Abstract. In this paper, we describe the methods deployed in the different runs submitted for our participation to the CLEF eHealth2018 Task 3: Consumer Health Search Task, IRTask 3: Query Variations. In particular, we deploy data fusion techniques to merge search results generated by multiple query variants. As improvement, we attempt to alleviate the term mismatch between the queries and the relevant documents by deploying query expansion before merging the results. For our baseline system, we concatenate the multiple query variants for retrieval and then deploy query expansion.

Keywords: Query variation, Data Fusion, Query expansion

1 Introduction

The high prevalence of the internet has led to an increase in patients and health care providers seeking for health-related information online. The sources consulted include social media platforms and web pages owned and operated by diverse entities. Health Information seekers can be classified into experts and no-experts/laymen. The key distinction is that experts have rich domain knowledge whereas non-experts have limited or no domain knowledge. These two groups express their information needs by way of queries to search engines. The queries submitted often vary in content due to the diverse backgrounds of the information seekers. The challenge is thus for search engines to be able to return relevant information regardless of the type of query submitted. Cognizant of this many evaluation campaigns have been launched to enable researchers to share knowledge and develop through experiments, effective information retrieval systems to cater for this need.

We thus seek to contribute to this effort by participating in one of these campaigns, the CLEF eHealth 2018 Task 3: Consumer Health Search, IRTask 3: Query variations [10]. The campaign is aimed at building search systems that are robust to query variations. This task is a continuation of the previous CLEF eHealth Information Retrieval (IR) task that ran in 2013 [2], 2014 [3], 2015 [4], 2016 [6] and 2017 [8]. In this work we attempt to attain an effective ranking by merging search results generated by multiple query variants of the same information need using data fusion techniques. In particular, we follow earlier work by Thuma et al. [11], who deployed data fusion techniques to merge search results generated by query variants, which were formulated through the collection enrichment approach using different external resources. Furthermore, we attempt to improve the retrieval effectiveness by alleviating the term mismatch between the queries and the relevant documents by deploying query expansion.

The paper is structured as follows. Section 2 contains a background on algorithms used. Section 3 describes the test collection. Section 4 describes the experimental environment. In Section 5 we provide a description of the different runs submitted. Section 6 presents results and discussion.

2 Background

In this section, we present essential background on the different algorithms used in our experiments. We start by describing the DPH term weighting model in Section 2.1. We then describe the data fusion techniques used in this study in Section 2.2. We conclude the background by describing the Kullback-Lieber Divergence for Query Expansion in Section 2.3.

2.1 PL2 Divergence From Randomness (DFR) Term Weighting Model

In our experiments, we deploy the PL2 Divergence from Randomness (DFR) term weighting model, which applies term frequency normalisation of a term in a document [9]. The relevance score of a document d for a given query Q based on the PL2 DFR term weighting model is expressed as follows:

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

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. qtw is the query term weight and is given by $\frac{qtfn}{tfn+1}$, where 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 Data Fusion Techniques

In this work, we postulate that an effective ranking can be attained by merging search results generated by multiple query variants of the same information need.

In order to validate this hypothesis, we deploy two different data fusion techniques. In particular, we deploy CombSUM and Reciprocal Rank. CombSUM is a score aggregation technique, where the score of a document is computed the sum of the normalised scores received by the document in each individual ranking [7]. In this work, we adapted CombSUM to merge search results generated by multiple query variants of the same information need and we define the score of the final ranking as:

$$score_{CombSUM}(d) = \sum_{r(Q_{i=1}) \in R}^{Q_7} score_{r(Q_i)}(d) \tag{3}$$

where $score_{r(Q_i)}$ is the score of the document d in the document ranking $r(Q_i)$. R is the set of all the rankings generated by the query variants Q_i. In the Reciprocal Rank (RR) data fusion technique, the rank of a document in the combined ranking is determined by the sum of the reciprocal received by the document in each of the individual rankings [7]. In this work, we define the score of the final ranking after merging search results generated by multiple query variants using RR as:

$$score_{RR}(d) = \sum_{r(Q_{i=1}) \in R}^{Q_7} \frac{1}{rank_d} \tag{4}$$

where $rank_d$ is the rank of document d in the document ranking $r(Q_i)$.

2.3 Kullback-Leibler (KL) Divergence for Query Expansion

In this study, we deployed the Terrier-4.2 Kullback-Leibler divergence for query expansion to attempt to alleviate the term mismatch between the queries and the relevant documents in the collection being searched. In our deployment, we used the default terrier settings, where we select the 10 most informative terms from the top 3 documents after a first pass document ranking. The KL divergence for query expansion calculates the information content of a term t in the top-ranked documents as follows [1]:

$$w(t) = (P_x(t)) \log_2 \frac{P_x(t)}{P_n(t)} \tag{5}$$

$$P_x(t) = \frac{tfx}{x} \tag{6}$$

$$P_n(t) = \frac{tfc}{N} \tag{7}$$

where $P_x(t)$ is the probability of t estimated from the top x ranked documents, 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. The top 10 terms with the highest information content computed by $w(t)$ are then selected and used for query expansion.

3 Test Collection

In this Section, we describe the test collection used in this study. First, we describe the document collection (corpus) used for indexing and retrieval in Section 3.1. In Section 3.2 we describe the queries used for retrieval.

3.1 Document Collection

"The document collection used in CLEF 2018 consists of web pages acquired from the CommonCrawl. An initial list of websites was identified for acquisition. The list was built by submitting the CLEF 2018 queries to the Microsoft Bing Apis (through the Azure Cognitive Services) repeatedly over a period of few weeks**, and acquiring the URLs of the retrieved results. The domains of the URLs were then included in the list, except some domains that were excluded for decency reasons (e.g. pornhub.com). The list was further augmented by including a number of known reliable health websites and other known unreliable health websites, from lists previously compiled by health institutions and agencies. "[1]

3.2 Queries

In this study we used queries created from 50 topics, which were identified from queries issued by the general public to Health on the NET (HON)[2] and TRIP[3] search services. From each topic, 7 different query variations were created. The first 4 query variations were created by people with no medical knowledge, while the second 3 were created by medical experts. Details on how the queries were created can be found in Jimmy et al. [5].

4 Experimental Setting

FAQ Retrieval Platform: For all our experiments, we used Terrier-4.2 [4], an open source Information Retrieval (IR) platform. All the documents 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 query expansion, we used the Terrier-4.2 Kullback-Leibler (KL) Divergence for query expansion to select the 10 most informative terms from the top 3 ranked documents.

[1] https://sites.google.com/view/clef-ehealth-2018/task-3-consumer-health-search

[2] https://hon.ch/en/

[3] https://www.tripdatabase.com/

[4] www.terrier.org

5 Description of the Different Runs

Term Weighting Model: For all our runs, we used the Terrier-4.2 PL2 Divergence from Randomness (DFR) term weighting model to score and rank the documents in the document collection.

ub-botswana_IRTask3_run1: This is our baseline run. We concatenated all the 7 query variants for each information need. Duplicates were not removed to ensure that a query term appearing in multiple query varients has a higher query term weight (*qtw*). We then performed retrieval on the document collection using the concatenated queries. We ranked the documents using the PL2 term weighting model.

ub-botswana_IRTask3_run2: In this run, our aim was to validate the hypothesis that an effective ranking can be attained by merging search results generated by multiple query variants of the same information need. In order to achieve this, we retrieved and ranked the documents in the collection using the 7 query variants for each information need. For each information need, we merged the search results using CombSUM, which we described in Section 2.2.

ub-botswana_IRTask3_run3: This is an improvement to our second, which is *ub-botswana_IRTask3_run2:*. In particular, our aim was to improve the retrieval effectiveness by alleviating the term mismatch between the queries and the relevant documents in the document collection. We deployed query expansion using the KL divergence model before merging the results using CombSUM in an attempt to alleviate the term mismatch.

ub-botswana_IRTask3_run4: In this run, we tested the generality of our approach in order to validate whether an effective ranking can be attained by merging search results generated by multiple query variants of the same information need by deploying a second data fusion technique. In particular, we deployed the Reciprocal Rank (RR) data fusion technique. In the same vein as our third run, which is *ub-botswana_IRTask3_run3:*, we deployed query expansion using the KL divergence model before merging the results using Reciprocal Rank (RR).

6 Results and Discussion

These working notes were compiled and submitted before the relevance judgments were released. Therefore, we were unable to report on our results and evaluation.

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. Jimmy, G. Zuccon, J. Palotti, L. Goeuriot, and L. Kelly. Overview of the CLEF 2018 Consumer Health Search Task. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes*. CEUR-WS, September 2018.
6. L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, and G. Zuccon. *Overview of the CLEF eHealth Evaluation Lab 2016*, pages 255–266. Springer International Publishing, Cham, 2016.
7. C. Macdonald and I. Ounis. Voting for candidates: Adapting data fusion techniques for an expert search task. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 387–396, New York, NY, USA, 2006. ACM.
8. 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.
9. V. Plachouras and I. Ounis. Multinomial randomness models for retrieval with document fields. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 28–39, Berlin, Heidelberg, 2007. Springer-Verlag.
10. H. Suominen, L. Kelly, L. Goeuriot, E. Kanoulas, L. Azzopardi, R. Spijker, D. Li, A. Névéol, L. Ramadier, A. Robert, J. Palotti, Jimmy, and G. Zuccon. Overview of the CLEF eHealth Evaluation Lab 2018. In *CLEF 2018 - 8th Conference and Labs of the Evaluation Forum*. Lecture Notes in Computer Science (LNCS), Springer, September 2018.
11. E. Thuma, O.G. Tibi, and G. Mosweunyane. A comparison between selective collection enrichment and results merging in patient centered health information retrieval. *International Journal of Computer Applications*, 180(29):1–8, Mar 2018.