International Journal of Computer Science and Software Engineering (IJCSSE), Volume 5, Issue 10, October 2016
ISSN (Online): 2409-4285 www.IJCSSE.org Page: 227-232

Data Fusion in Patient Centered Health Information Retrieval

Nkwebi Peace Motlogelwa[1], Edwin Thuma[2] and Tebo Leburu-Dingalo[3]

[1, 2, 3] Department of Computer Science, University of Botswana, Gaborone, P/BAG UB 00704, Botswna

[1]*motlogel,thumae,leburut@mopipi.ub.bw*

ABSTRACT

When faced with medical ailments, a majority of lay people often submit circumlocutory queries to modern web search engines for self-diagnosis. However, such queries often fail to retrieve relevant documents to answer their information need. In this articles, we attempt to improve the retrieval effectiveness of such systems by enriching these circumlocutory queries with the most informative terms from two different external collections to produce two different expanded queries. In addition, we submit these expanded queries to the local collection (collection being searched) to produce two different rankings. Furthermore, we deploy data fusion techniques to combine these multiple rankings in order to further improve the retrieval effectiveness of such systems. Our empirical evaluation shows marked improvement in the retrieval performance of such systems when we fuse these multiple rankings. In particular, we see an improvement in both precision at 5 (P@5), precision at 10 (P@10) and recall when the number of rankers deployed for each data fusion technique is increased.

Keywords: *Data Fusion, Collection Enrichment, Query Expansion.*

1. INTRODUCTION

Development of effective Information Retrieval (IR) systems has been an active area of research for a number of years. This is evidenced by a number of evaluation campaigns such as Text Retrieval Conference (TREC)[1], the Conference and Labs of the Evaluation Forum (CLEF) [2], the Forum for Information Retrieval Evaluation (FIRE) [3], and the NII Testbeds and Community for Information access Research (NTCIR)[4]. These evaluation campaigns encourage researchers across the globe to systematically evaluate information access systems, primarily through experimentation on shared tasks. This has led to the development of several novel search strategies whose benefits are being realized in modern search engines such as Google, Yandex and Bing. These search engines provide users with relevant information from the vast amount of heterogeneous information available on the Internet, which is often used as a central source of information for a variety of domains including health related issues. However, access to useful health information by modern search engine users who try to self-diagnose is often difficult because they typically construct circumlocutory queries, using colloquial language devoid of medical terms [1, 2]. This often makes it difficult to retrieve relevant documents, which are more readily retrieved using medical terms. According to Stanton et.al [1], current search engines are unable to handle this type of queries and thus often return information that is not useful as medical advice to users. For example, if a user submits the following query to a modern search engine: “Baldness in multiple spots” instead of “alopecia”, it is likely the search engine would retrieve few relevant documents [2].

In line with these findings, the CLEF (Conference and Labs of the Evaluation Forum) eHealth Information Retrieval (IR) task that ran in 2013 [3,4], 2014 [5,6], 2015 [7] and 2016 [8] aim to evaluate the effectiveness of information retrieval systems when searching for health related content on the web. Its main objective is to foster research and development of search engines tailored to health information seeking. This paper is part of our contribution to the research, with a focus on improving IR performance through query enrichment with multiple external resources and application of data fusion techniques to combine multiple rankers in order to further improve the retrieval effectiveness of such systems.

The remainder of this paper is organized as follows: In Section 2, we provide a review on related work and brief but essential background on the different IR approaches used in this study. This is followed by a description of the dataset used in this study in Section 3. In Section 4, we describe our experimental investigation and evaluation, followed by a description of our experimental setting in Section 5. In Section 6, we present an analysis of our experimental results. This is followed by a discussion and conclusion in Section 7.

[1] http://trec.nist.gov
[2] http://www.clef-initiative.eu
[3] http://fire.irsi.res.in/fire/2016/home
[4] http://research.nii.ac.jp/ntcir/index-en.html

2. BACKGROUND AND RELATED WORK

In this section, we begin by presenting related work and background on the different algorithms used in our experimental investigation and evaluation. Section 2.1 reviews related work, this is followed by a description of the PL2 and the BM25 term weighting model in Section 2.2 and 2.3 respectively. In Section 2.4, we describe the Bose-Einstein 1 (Bo1) model for query expansion, followed by a description of Data Fusion techniques in Section 2.5.

2.1 Related Work

The CLEF eHealth challenge was motivated by the inability of modern search engines (such as Google, Yandex and Bing) to handle circumlocutory queries [3,4]. According to Zuccon et al. [2], these queries are often submitted to modern search engines by laypeople without sufficient knowledge of medical terminology when they attempt to self-diagnose ailments. A series of studies through the CLEF tasks aimed at improving non-expert user access to health information have attempted to solve this problem. In 2013 CLEF initiated a study aimed at developing techniques specific to user's medical queries based on information from patient discharge summaries [3,4]. Several teams participated in the task and only Team-Mayo submitted runs, which performed better than the baseline BM25 term weighting model with pseudo-relevance feedback deployed [4]. In their approach, they used the query likelihood term weighting model [4,9]. As an improvement they deployed proximity search using markov random fields and they also expanded the original query with terms from the Mesh ontology [4,9]. Several other teams also used other term weighting models and query expansion with different external resources such as MEDLINE, Wikepedia, UMLS and MetaMap. However, there was no improvement in the retrieval performance compared to the baseline BM25 [4].

In 2014 further studies were performed with Shen et.al [10], being able to achieve better performance using Language Modelling and an external resource to enhance users' queries. The success of Shen et.al [10] and Verberne [11] who both used query expansion with UMLS illustrated that addition of medical terms to user's queries is essential to the effectiveness of a health related search. This was further observed by Thuma et.al [12] who got similar results when they expanded the original query with additional medical terms from MEDLINE as opposed to Wikipedia2008. Similarly, in the CLEF2015 and CLEF2016 eHealth tasks, several teams that participated in these tasks investigated the effectiveness of using different external resources for query expansion in order to improve the retrieval effectiveness [7,8]. However as noted by Palotti et.al [7] in the overview of the study, further work is needed to improve the precision of the actual search results as most studies have shown that even with query expansion only half of the top 10 results retrieved were found to be relevant. In this work, we differ with prior work by using multiple external resources for query expansion. In addition we produce different rankings using the expanded queries. Furthermore we combine these rankings using data fusion techniques in order to improve the overall retrieval effectiveness. In Section 2.2 and 2.3, we describe the baseline term weighting models used in this study.

2.2 PL2 Term Weighting Model

For our experimental investigation and evaluation, we used two different state of the art term weighting models namely PL2 and BM25 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 [13]:

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

where $score_{PL2}$ (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 normalized query term frequency is given by $qtfn = \frac{qtf}{qtf_{max}}$, where qtf_{max} is the maximum query term frequency among the query term and qtf is the query term frequency. tfn is the Normalization 2 of the term frequency tf of the term t in a document d and is expressed as:

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

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.

2.3 BM25 Term Weighting Model

For a given query q, the relevance score of a document d, based on the BM25 term weighting model is expressed as [14]:

$$score_{BM25}(d,Q) = w^{(1)} \sum_{t \in Q} \frac{(k_1+1)tfn}{k_1+tfn} \cdot \frac{(k_3+1)qtf}{k_3+qtf}$$
(3)

where qtf is the number of occurrences of a given term t in the query Q. $k_1$and $k_3$are parameters of the model. tfn is the normalized within document term frequency. $w^{(1)}$denotes the Robertson-Spark Jones (RSJ) weights, which is an inverse document frequency (IDF) factor and is given by:

$$w^{(1)} = \log \frac{N-dft+0.5}{dft+0.5}$$
(4)

Where N is the number of documents in the collection and dft is the number of documents in the collection that have a term t.

2.4 Query Expansion

In this paper, we enrich our queries with terms from multiple external resources to try to improve the retrieval performance. In particular, we used the Terrier-4.0 Divergence from Randomness (DFR) Bose-Einstein 1 (Bo1) model for query expansion to select the most informative terms from the topmost document 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 [15]:

$$w(t) = tfx \cdot \log_2 \frac{1+P_n(t)}{P_n(t)} + \log_2(1+P_n(t))$$
(5)

$$P_n(t) = \frac{tfc}{N}$$
(6

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.

2.5 Data Fusion

In this paper, we use data fusion techniques to combine rankings of different rankers in order to improve the retrieval effectiveness of our system. Our aim is to ensure that documents scored highly in many rankings are likely to be scored highly in the final ranking. In contrast, documents with low scores, or that are present in less rankings are less likely to end up highly in the final ranking. In Table 1 [16], we provide a description of the different data fusion techniques used in this study.

Table 1
Table 1: Data Fusion Techniques used in this paper.

Name	Description
RR	Sum of inverse of ranks of documents retrieved by each system.
CombSUM	Uses the summation of relevance scores by each system as the fused relevance score.

3. DATASET

In this section we present the dataset used in this study. In Section 3.1, we present the document collection used for searching and retrieval, this is followed by a description of the queries and the query relevance judgments in Section 3.2.

3.1 Document Collection

In this study, we used the ClueWeb12-B13[1], a collection of more than 52 million web pages. This ClueWeb12-B13 web crawl is a higher-fidelity representation of a common Internet craw, making the dataset more in line with the content current web search engines index and retrieve. This dataset was made available to the CLEF eHealth task participants by the organizers of the task through cloud computing instances provided by Microsoft Azure.

3.2 Queries and Query Relevance Judgements

We used a total of 300 circumlocutory queries that users may pose when faced with signs and symptoms of a medical condition. These queries were extracted from post published in the AskDocs section of Reddit[2], which is a public health web forum. These queries and their query relevance judgments were collected as described in Zuccon et al [8]. In Table 2, we present a selection of queries used in this study.

Table 2: A selection of test queries used in this study

Test Queries
102001 : anal Tag removal options
103001 : headaches relieved by blood donation
104002 : coughing laying down

[1] http://lemurproject.org/clueweb12/
[2] https://www.reddit.com/r/AskDocs

4. Experimental Investigation and Evaluation

Recall from Section 1 that we aim to improve the retrieval performance through query enrichment with multiple external resources and deploying data fusion techniques to combine multiple rankers in order to further improve the retrieval effectiveness. In this article, we differ with our prior work in Thuma et al. [17], where we only used one external resource and one data fusion technique. In order to archive this, the following research questions were identified:

RQ1: Can we improve the retrieval performance of our system by enriching the original query terms with terms from an external collection of documents.

RQ2: Can we improve the retrieval performance of our system by using data fusion techniques to combine the retrieval results of several rankers.

RQ3: Does increasing the number of different rankers used in our data fusion approach improve the retrieval performance.

In Section 4.1, we provide a description of the runs designed to answer the aforementioned research question RQ1, RQ2 and RQ3.

4.1 Description of the Different Run

PL2_Baseline and BM25_Baseline: We used the PL2 and BM25 term weighting models in the Terrier-4.0 IR platform as our baseline systems. In our preliminary investigations, the PL2 term weighting model outperformed the BM25 term weighting model in terms of P@5, P@10 and recall. Therefore, we deploy the PL2 term weighting model in subsequent runs.

CLEF_Run: In order to answer research question one (RQ1), we used the PL2 term weighting model in the Terrier-4.0 IR platform. As improvement, we deployed the collection enrichment approach [18], where we selected the expansion terms from an external collection, which was made up of a collection of documents from CLEF_2015_eHealth. We used the Terrier-4.0 Divergence from Randomness (DFR) 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 (ClueWeb12-B13) with the new expanded query. In Table 3, we present a selection of expanded, preprocessed test queries (stemmed and tokenized). These queries have been assigned term weights by the Bose – Einstein 1 (Bo1) model for query expansion.

Table 3: A selection of expanded, stemmed test queries with term weights assigned by the Bo1 model for query expansion

Test Queries
102001 : anal^1.422081564 tag^1.000000000 remov^1.000000000 option^1.000000000 anu^0.189241539 fistula^0.170902342 sphincter^0.090926864 rectum^0.065447220 faecal^0.056256694 incontin^0.032928088 perian^0.031780140 procedur^0.029974398 cancer^0.028866764
103001 : headach^1.000000000 reliev^1.000000000 blood^1.157894694 donat^1.220645852 apheresi^0.093626771 mayo^0.091765421 bloodborn^0.080712685 encodeuri^0.075139893 donor^0.074911750 needl^0.072440809 rh^0.063180225 givelife2^0.061542378
104002 : cough^1.179193820 lai^1.000000000 down^1.000000000 flu^0.218899307 went^0.161063267 throat^0.148441101 feel^0.136354148 sore^0.122559512 night^0.104422729 headach^0.080504868 hurt^0.073353096 sick^0.072107820

Wiki_Run: In order to answer research question one (RQ1), we used the PL2 term weighting model in the Terrier-4.0 IR platform. As improvement, we deployed the collection enrichment approach [18], where we selected the expansion terms from an external collection, which was made up of a collection of documents from Wikipedia2008. We used the Terrier-4.0 Divergence from Randomness (DFR) 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 (ClueWeb12-B13) with the new expanded query.

Data_Fusion_Runs.RQ2. In order to answer research question two (RQ2), we used CombSUM to combine rankings generated by the PL2_Baseliene and CLEF_Run rankers. We call this new ranking PL2_CLEF_Run.csum. We test the generality of our approach by deploying the reciprocal rank (RR) data fusion technique on the aforementioned rankers, which produce the PL2_CLEF_Run.RR run. These data fusion techniques were also used to fuse rankings generated after expanding the query terms with Wikipedia2008 and the CLEF_2015_eHealth document collections to generate CLEF_Wiki_Run.csum and CLEF_Wiki_Run.RR.

Data_Fusion_Runs.RQ3: In order to answer research question three (RQ3), we used CombSUM to combine rankings generated by the PL2_Baseline, CLEF_Run and Wiki_Run rankers. We test the generality of our approach by deploying the reciprocal rank (RR) data fusion technique on the aforementioned rankers, which produce PL2_CLEF_Wiki_Run.csum and PL2_CLEF_Wiki_Run.RR for CombSUM and RR respectively.

5. EXPERIMENTAL SETTING

For all our experimental evaluation, we used Terrier-4.0, an open source IR platform. All the document collections used in this study namely; the ClueWeb12-B13, CLEF_2015_eHealth and Wikipedia2008 were first pre-processed before indexing and this involved tokenizing the text and stemming each token using the full porter stemming algorithm. We enabled stopword removal by using the default Terrier stopword list. The hyper-parameter of the PL2 term weighting model was set to its default value of 1.0.

6. AN ANALYSIS OF EXPERIMENTAL RESULTS

*Table 4: The mean retrieval performance for each run.P@5, P@10 and Recall were computed using **Trec_eval** [1].*

RUN_ID	P@5	P@10	RECALL
PL2_Baseline	0.3207	0.2963	0.5041
BM25_Baseline	0.3080	0.2890	0.4944
CLEF_Run	0.3313	0.2963	0.4778
Wiki_Run	0.3213	0.2940	0.4744
PL2_CLEF_Run.csum	0.3307	0.3017	0.5148
PL2_CLEF_Run.RR	0.3227	0.2957	0.5117
CLEF_Wiki_Run.csum	0.3347	0.3077	0.5185
CLEF_Wiki_Run.RR	0.3207	0.2950	0.5034
PL2_CLEF_Wiki_Run.csum	0.3347	0.3083	0.5190
PL2_CLEF_Wiki_Run.RR	0.3333	0.2980	0.5239

In Table 4, we present the results for our empirical evaluation. As depicted in Table 4, the PL2 term weighing model outperformed the BM25 term weighting model in terms of P@5, P@10 and recall for our baseline systems. Based on these results, we subsequently use PL2 term weighting model as our term weighting model in other runs. In order to answer RQ1 (Can we improve the retrieval performance of our system by enriching the original query terms with terms from an external collection of documents), we enriched the original query terms with terms from two external collections (CLEF_Run and Wiki_Run). In Table 4, we see an improvement in terms of P@5 when the original query terms are enriched. However, there is a degradation in recall for these expanded queries. To take advantage of the two rankings generated by the expanded and unexpanded queries, we investigate whether we can improve the retrieval performance of our system by using data fusion techniques to combine the retrieval results of these rankers (RQ2). As can be seen in Table 4, there is an improvement in the retrieval performance in term of P@5, P@10 and recall when several rankers are combined with data fusion techniques (PL2_CLEF_Run.csum, PL2_CLEF_Run.RR, CLEF_Wiki_Run.csum and CLEF_Wiki_Run.RR). Furthermore, we see an improvement in the retrieval performance in terms of P@10 and recall when we increase the number of rankers with our data fusion technique (PL2_CLEF_Wiki_Run.csum and PL2_CLEF_Wiki_Run.RR).

[1]http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz

7. CONCLUSIONS

In this paper, we attempt to improve the retrieval effectiveness of an IR system by enriching circumlocutory queries with the most informative terms from an external collection (RQ1). The results of this study show that we can improve the retrieval performance of such systems when the original queries are enriched with terms from an external collection. Furthermore, we investigate whether deploying data fusion techniques to combine these rankings generated by the expanded and unexpanded queries can further improve the retrieval effectiveness of such systems (RQ2). Our empirical evaluation shows marked improvement in the retrieval performance of such systems when we fuse these multiple rankings. In particular, we see a significant improvement in both precision at 5 (P@5), precision at 10 (P@10) and recall when the number of rankers deployed for each data fusion technique is increased (RQ3).

REFERENCES

[1] I. Stanton, S. Ieong, and N. Mishra, “Circumlocution in Diagnostic Medical Queries,” In Proc. of the 37th ACM SIGIR. ACM, New York, USA, 2014, pp. 133 – 142.

[2] 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 Proc. of the 37th ECIR, Springer-Verlag, Berlin, Heidelberg, 2015, pp. 562 – 567.

[3] H. Suominen, S. Salantera, S. Velupillai, W.W Chapman, G. Savova, N. Elhadad, et al. “Overview ShARe/CLEF eHealth Evaluation Lab 2013,” In Proc. of 4th Int. of the CLEF, Springer-Verlag, Berlin, Heidelberg, 2013, pp.212 – 213.

[4] L. Goeuriot, G.J.F Jones, L. Kelly, J. Leveling, A. Hanbury, H. Muller, 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, vol. 1179, CEUR-WS, 2013.

[5] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. Mowery, et al. “Overview ShARe/CLEF eHealth Evaluation Lab 2014,” In Proc. of 4th Int. of

the CLEF, Springer-Verlag, Berlin, Heidelberg, 2014, pp.172 – 191.

[6] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G.J.F Jones and H. Muller, “ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-Centered Health Information Retrieval” In CLEF 2014 Online Working Notes, vol. 1180, CEUR-WS, 2014, pp. 43 – 61.

[7] 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, vol. 1391, CEUR-WS, 2015.

[8] G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu and P. Pecina, H. Muller, J. Budaher and A. Deacon “The IR Task at the CLEF eHealth Evaluation Lab 2016: User-Centered Health Information Retrieval” In CLEF 2016 Online Working Notes, vol. 1609, CEUR-WS, 2016, pp. 15 – 27.

[9] D. Zhu, S. Wu, M. James, B. Carterette and H. Liu, “Using Discharge summaries to improve information retrieval in clinical domain” In CLEF 2013 Online Working Notes, vol. 1179, CEUR-WS, 2013.

[10] W. Shen, J.Y. Nie, X. Liu and X. Liui, “An Investigation of the Effectiveness of Concept-Based Approach in Medical Information Retrieva GRIUM @ CLEF2014eHealth Task 3” In CLEF 2014 Online Working Notes, vol. 1180, CEUR-WS, 2014, pp. 236 – 247.

[11] S. Verberne “A Language-Modelling Approach to User-Centered Health Information Retrieval” In CLEF 2014 Online Working Notes, vol. 1180, CEUR-WS, 2014, pp. 269 – 275.

[12] E. Thuma, G. Anderson and G. Mosweunyane, “UBML Participation to CLEF eHealth IR Challenge 2015: Task 3” In CLEF 2015 Online Working Notes, vol. 1391, CEUR-WS, 2015.

[13] V. Plachouras and I. Ounis, “Multinomial Randomness Models for Retrieval with Document Fields,” In Proc. of the 29th ECIR, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 28 – 39.

[14] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Found. and Trends in Info. Retr., vol 3, no. 4, pp. 333 – 389, Apr. 2009.

[15] G. Amati, “Probabilistic Models for Information Retrieval Based on Divergence from Randomness,” University of Glasgow, PhD Thesis, June. 2003, pp. 1–198.

[16] C. Macdonald and I. Ounis, “Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task,” In Proc. of the 15th ACM Int. CIKM. ACM, New York, USA, 2006, pp. 387 – 396.

[17] E. Thuma, N.P. Motlogelwa, and T. Leburu-Dingalo “Task 3: Patient-Centered Information Retrieval, IRTask 1: Ad-hoc Search” In CLEF 2016 Online Working Notes, vol. 1609, CEUR-WS, 2016, pp. 162 – 166.

[18] K.L. Kwok, and M. Chan, “Improving Two-Stage Ad-hoc Retrieval for Short Queries,” In Proc. of the 21st ACM SIGIR. ACM, New York, USA, 1998, pp. 250 – 256.