KDEIR at CLEF eHealth 2016: Health Documents Re-ranking Based on Query Variations

Md Zia Ullah[1] and Masaki Aono[†]

Department of Computer Science and Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-Cho, Toyohashi, 441-8580, Aichi, Japan,
arif@kde.cs.tut.ac.jp[1], aono@tut.jp[†]

Abstract. In this paper, we describe our participation in the CLEF eHealth 2016 task 3: Patient-Centred Information Retrieval focusing on the clinical web documents based on user queries in the health forum. In our participation, we submitted three runs in ad-hoc search and two runs in query variation search subtasks. In ad-hoc search, the main challenge is to retrieve high quality clinical documents based on user query. For ad-hoc search, we employ multiple features based unsupervised re-ranking method to the documents retrieved by a baseline system. During the query variation search, the main challenge is to generate a ranked list of documents covering the different variations of the query. To tackle the query variation problem, first we formulate a query and a set of information needs from the query variation. Then, we re-rank the documents retrieved for the formulated query by focusing on the set of information needs.

Keywords: Health Informatics, Query variation, Re-ranking, Diversity

1 Introduction

Users often seek health related information and issue their information needs in the Web search engine. Unfortunately, the documents on the Web are mostly low-quality and contain a lot of spam information. Therefore, laypeople are usually unsuccessful in getting their health related answers. CLEF eHealth Evaluation Lab [1–5] has been organizing a task for the last couple of years to help the laypeople to obtain health related information. In 2016, task 3 includes three subtasks such as ad-hoc search, query variation search, and multilingual search. We participated in the ad-hoc search and query variation search. In particular, we desire to evaluate the following research question in both of the subtasks:

1. Is the spam identification method sufficiently reliable for the ad-hoc retrieval task?
2. Is multiple features based ranking method enough to estimate the topical relevance of the query and documents?
3. Is the diversity based ranking successfully handling query variation?

2 Our Submitted Runs

We submitted a total five runs where three runs in an ad-hoc retrieval and two runs in query variation. We make use of *Clueweb12-B13* [6] corpus by indexing with the Indri search engine [7]. To emphasize on high quality documents, we filter out the spam documents from the corpus by using the Waterloo spam score [8].

2.1 Ad-hoc Search

To prepare three runs in ad-hoc search, we apply some common procedures. Given a query, first, we tokenize the query and format it for ad-hoc retrieval with the Indri Search engine. Second, we retrieve at most 1000 documents from the *Clueweb12-B13* corpus using a query likelihood model with the Dirichlet smoothing model as baseline retrieval. Third, document with a spam score less than 70 is filtered out of the retrieved documents. The three runs are described as follows:

Run 1 (KDEIM_EN_Run1): following the common procedures described above, we re-rank the documents by fusing the page rank and documents' baseline (language model) scores and take the top 200 documents.

Run 2 (KDEIM_EN_Run2): following the common procedures stated above, we extract multiple query-independent and query-dependent features including reciprocal rank, topic cohesiveness [9], average term length, vector space similarity [10], coordinate level matching, BM25 [11], PL2 [12], DFR [12], Kullback-Laibler [13]. We re-rank the documents using the extracted features by employing a bipartite graph based ranking approach and take the top 200 documents.

Run 3 (KDEIM_EN_Run3): following the common procedures stated above, in comparison to the previous two runs, we tokenize the query and format it for expert retrieval, where we consider document title, header, body, and anchor text. Then, we re-rank the documents by combining page rank and documents' baseline (language model) score, and take the top 200 documents.

2.2 Query variation Search

To prepare two runs in query variation search, we apply some common procedures. First, we tokenize all the six query variations and formulate a vector space model representation of the query from the query variations. We also consider all the query variations as information needs (sub-queries) of the users. Second, we retrieve at most 1000 documents from the *Clueweb12-B13* corpus based on the formulated query as baseline retrieval. Third, document with a spam score less than 70 is filtered out of the retrieved documents.

Run 1 (KDEIM_EN_Run1): following the common procedures described above, we re-rank the documents using page rank and documents' baseline (language model) scores as a relevance based ranking. By considering the query variations as sub-queries (aka, information needs), we employ an explicit diversification algorithm [14] and take the top 100 documents.

Run 2 (KDEIM_EN_Run2): following the common procedures stated above, we re-rank the documents using multiple query-independent and query-dependent features including page rank, reciprocal rank, topic cohesiveness [9], average term length, vector space similarity [10], coordinate level matching, BM25 [11], PL2 [12], DFR [12], Kullback-Laibler [13]. Then, we explicitly diversify the documents based on the sub-queries and take the top-100 documents.

3 Conclusion

In this paper, we described the participation of KDEIR at CLEF eHealth 2016 Patient-Centred Information Retrieval task, where we proposed our approaches to ad-hoc search and query variations of clinical documents.

Acknowledgement

This research was partially supported by the HORI FOUNDATION of JAPAN, Grant-in-Aid C114.

References

1. Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., Deacon, A.: 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 (2016)
2. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: 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 (2016)
3. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J., et al.: Overview of the share/clef ehealth evaluation lab 2013. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer (2013) 212–231
4. Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., et al.: Overview of the share/clef ehealth evaluation lab 2014. Springer (2014)
5. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2015. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer (2015) 429–443
6. Callan, J., Hoy, M., Yoo, C., Zhao, L.: Clueweb09 data set (2012)
7. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis, Citeseer (2005) 2–6
8. Cormack, G.V., Smucker, M.D., Clarke, C.L.: Efficient and effective spam filtering and re-ranking for large web datasets. Information retrieval **14**(5) (2011) 441–465
9. Bendersky, M., Croft, W.B., Diao, Y.: Quality-biased ranking of web documents. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM (2011) 95–104

10. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5) (1988) 513–523
11. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc (2009)
12. Amati, G.: Probability models for information retrieval based on divergence from randomness. PhD thesis, University of Glasgow (2003)
13. Lafferty, J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2001) 111–119
14. Santos, R.L., Macdonald, C., Ounis, I.: Exploiting query reformulations for web search result diversification. In: Proceedings of the 19th international conference on World wide web, ACM (2010) 881–890