ECNU at 2016 eHealth Task 3: Patient-centred Information Retrieval

Yang Song[1,2], Yun He[1,2], Hongyu Liu[1,2], Yueyao Wang[1,2], Qinmin Hu[1,2], Liang He[1,2], and Guihua Luo[3,4]

[1] Shanghai Key Laboratory of Multidimensional Information Processing
[2] Department of Computer Science & Technology, East China Normal University, Shanghai, 200241, China
{ysong,yhe,liuhy,yywang}@ica.stc.sh.cn, {qmhu,lhe}@cs.ecnu.edu.cn
[3] Science and Technology Commission of Shanghai Municipality
[4] Shanghai Digital Trade Co. Ltd, Shanghai, 200241,China
ghluo@metinform.cn

Abstract. The 2016 CLEF eHealth Task 3 aims to evaluation the effectiveness of information retrieval systems when searching for health content on the web. The ClueWeb12 B13 data set is utilized in this task. This paper presents our work on the 2016 CLEF e-Health Task 3. We propose a Web-based query expansion method and a combination method to better understand and satisfy the task. In particular, we test our Web-based query expansion method in many medical data sets and achieve an outstanding performance.

Keywords: Web-based, Query Expansion, Combination

1 Introduction

The 2016 CLEF eHealth Information Retrieval Task aims to evaluate the effectiveness of information retrieval systems when searching for health content on the web, with the objective to foster research and development of search engines tailored to health information seeking [1][2].

This task continues the previous CLEF eHealth information retrieval (IR) tasks that ran in 2013, 2014 and 2015. With a shared collection of documents and queries, this year's task embraces the TREC-style evaluation process, the contribution of runs from participants and the subsequent formation of relevance assessments and evaluation of the participants submissions.

ClueWeb12 B13 which is more representative of the current state of health information online is select as this year's IR task new corpus. The organizers mine health web forums to extract topic stories and generate the associated (English) queries to identify example information needs to be used in the task.

1.1 IRTask 1: Ad-hoc Search

In order to identify example information needs, queries for this task are generated by mining health web forums. This task extends the evaluation framework used in 2015 to consider further dimensions of relevance such as the reliability of the retrieved information.

1.2 IRTask 2: Query Variation

This task explores query variations for an information need. Different query variants are generated for the same forum entry, thus capturing the variability intrinsic in how people search when they have the same information need. Our participants should take these variations into account when building our systems. In this task we are told which queries relate to the same information need and we have to produce one set of results to be used as answer for all query variations of an information need. This task aims to foster research into building systems that are robust to query variations.

1.3 IRTask 3: Multilingual Search

This task, similar to last year, offers parallel queries in several languages (Czech, French, Hungarian, German, Polish and Swedish).

Our experiments on Task 3 aim to investigate in effectiveness of our Web-based query expansion method and the combination method for medical IR. Figure 1 presents our system architecture. Particularly, we take advantage of the Web search engine to obtain better expansion terms. At the same time, we adopt multiple classic IR models, such as BM25 [6], PL2 and BB2 [7], to perform retrieval, and we make combination in order to get rid of the influence of single model. Finally, we submit three runs for each subtask.

2 Methods

2.1 Web-based Query Expansion

The task of this year is patient-centered health IR. We intent to improve the retrieval precision by expanding the queries. To achieve a better expansion term we propose a Web-based query expansion method as follows. Note that we apply the similar model in the 2014 TREC Microblog track [3], 2015 TREC Clinical Decision Support track [4], and 2015 CLEF eHealth task 2 [5] which achieve better results than most of the runs.

- Query is searched by Google and the top 10 concurrent Web titles and snippets (if existed) are crawled from the Web page.
- By applying the MeSH database, the medical terms are extracted from both the titles and the snippets.

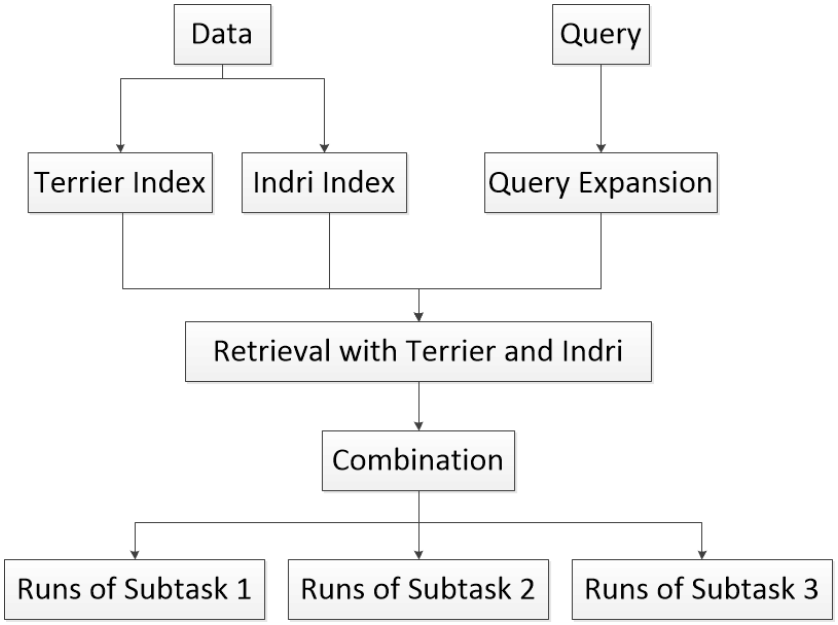

Fig. 1. System Architecture

- The frequency of each stemmed medical term is calculated. Only the terms appearing more than n times are kept for expanding, which can be denoted as Q_{web}.
- The final query is formulated as $Q = Q_0 \cup Q_{web}$, where Q_0 represents the initial query.

In addition, since some of the queries are to find out "what is the patient's diagnosis?", "what tests should the patient receive?" and "how should the patient be treated?", we manually add the keywords 'diagnose', 'test' and 'treatment' as the regular expansion terms to all queries.

2.2 Combination

We apply Equation 1 to normalize the scores of each candidate. Then, we add up all the normalized scores of each document among these candidates, followed by the documents which are ranked by the total normalized scores. Finally, the top 1000 documents for each query are extracted as final results for evaluation.

$$score\_normalized_i = \frac{score_{max} - score_i}{score_{max} - score_{min}} \tag{1}$$

3 Experiments

The corpus of the 2016 eHealth Task 3 is ClueWeb12-B13 which is the subset of ClueWeb12. The organizers provide us with the Microsoft Azure server and indexes in Terrier and Indir formats respectively. We adopt Terrier and Indri to conduct our experiments on the given resource.

In IRTask 1, we submit three runs where the description for each run is as follows.

- ecnu_EN_Run1: The baseline with the BM25 model and Terrier search engine.
- ecnu_EN_Run2: The query we utilizing is those removed punctuation. We adopt Terrier to perform BM25, PL2, BB2, DFR_BM25 [7]and Indri to perform TFIDF and Language Model [8]. The final submission is the combination of those results with the method proposed in section 2.2.
- ecnu_EN_Run3: We utilize the Google search engine to conduct query expansion, where there terms of 'diagnose', 'test' and 'treatment' are added by mandatory into queries. We use the same method as ecnu_EN_Run2 to perform retrieval and combination.

In IRTask 2, we also submit three runs.

- ecnu_EN_Run1_Task2: On the basis of the ecnu_EN_Run1 in task 1, we combine results of queries in the same group with the method proposed in section 2.2. For instance, we combine the results achieved by using queries "101001", "101002", "101003", "101004", "101005" and "101006" as the final result of query "101".
- ecnu_EN_Run2_Task2: We conduct combination on the basis of ecnu_EN_Run2 achieved in task 1. The combination method is the same as ecnu_EN_Run1_Task2.
- ecnu_EN_Run3_Task2: We conduct combination on the basis of ecnu_EN_Run3 achieved in task 1. The combination method is the same as ecnu_EN_Run1_Task2.

In IRTask 3, we utilize the Google Translate to translate the non-English queries into English. At the same time, we adopt the same method used in IRTask 1 to obtain our submissions.

4 Conclusions and Future Work

In 2016 CLEF eHealth task 3, We propose a Web-based query expansion model and a combination method to achieve the better performance for medical IR. In the future, we will continue on the Web-based query expansion method for better understand the queries and discover more effective methods to perform combination.

5 Acknowledgement

This research is funded by National Key Technology Support Program (No. 2015BAH12F01-04) and Science and Technology Commission of Shanghai Municipality (No.14DZ1101700). We also thank anonymous reviewers for their review comments on this paper.

References

1. Kelly, L., Goeuriot, L., Suominen, H., Nvol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. CLEF 2016 - 7th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September, 2016.
2. 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. CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September, 2016.
3. Chen, Q., Hu, Q.M., Pei, Y.J., Yang, Y., He, L.:ECNU at TREC 2014: Microblog Track. (2014)
4. Yang, S., He, Y., Hu, Q.M., He, L., Haacke, E.M.: ECNU at 2015 eHealth Task 2: User-centred Health Information Retrieval. Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2015).
5. Yang, S., He, Y., Hu, Q.M., He, L.: ECNU at 2015 CDS Track: Two Re-ranking Methods in Medical Information Retrieval. Proceedings of the 2015 Text Retrieval Conference. (2015).
6. Stephen E., Robertson, S.W., Susan J., Micheline H.B., Mike G.: Okapi at TREC-3. Proceedings of the Third Text REtrieval Conference (TREC 1994). Gaithersburg, USA.
7. Amati, G., Cornelis J., Van R.: Probabilistic models for information retrieval based on divergence from randomness. (2003).
8. Katz, S.M.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. Acoustics, Speech and Signal Processing, IEEE Transactions on 35.3 (1987): 400-401.