IMS-UNIPD @ CLEF eHealth Task 2: Reciprocal Ranking Fusion in CHS

Giorgio Maria Di Nunzio[1,2], Federica Vezzani[3]

[1]Department of Information Engineering, University of Padova, Italy
[2]Department of Mathematics, University of Padova, Italy
[3]Department of Linguistic and Literary Studies, University of Padova, Italy

Abstract

In this paper, we describe the results of the participation of the Information Management Systems (IMS) group at CLEF eHealth 2021 Task 2, Consumer Health Search Task. We participated in the three subtasks: Ad-hoc IR, Weakly Supervised IR, Document credibility. The goal of our work was to evaluate the reciprocal ranking fusion approach over 1) manual query variants; 2) different retrieval functions; 3) w/out pseudo-relevance feedback; 4) reciprocal ranking fusion.

Keywords

manual query variants, pseudo relevance feedback, reciprocal ranking fusion

1. Introduction

In the CLEF eHealth 2021 edition [1], the Task 2 "Consumer Health Search" [2] provides a set of experimental collections in order to study the performance of search engines that support the needs of health consumers that are confronted with a health issue. The three subtasks available are: Ad-hoc IR, Weakly Supervised IR, and Document credibility prediction.

The contribution of our experiments to the three subtasks is summarized as follows:

- A study of a manual query variation approach similar to [3];
- An evaluation of a ranking fusion approach [4] on different document retrieval strategies, with or without pseudo-relevance feedback [5];
- A simple fusion of normalized scores for document credibility.

The remainder of the paper will introduce the methodology and a brief summary of the experimental settings that we used in order to create the official runs submitted for this task.

2. Methodology

In this section, we describe the methodology for merging the ranking lists provided by different retrieval methods for different query variants.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio); federica.vezzani@unipd.it (F. Vezzani)
http://github.com/gmdn (G. M. Di Nunzio); http://www.dei.unipd.it/~vezzanif/ (F. Vezzani)
0000-0001-9709-6392 (G. M. Di Nunzio); 0000-0003-2240-6127 (F. Vezzani)

2.1. Subtask 1: Ad-hoc IR

For this subtask, we used the original queries as well as manual reformulations to simulate simpler (lay person) queries.

2.1.1. Query Variants

We asked to an expert in the field of medical Terminology to rewrite the original English query into one variant (similarly to [3]). The aim of the query rewriting was to describe in the simplest possible way the information need expressed by the query.

2.1.2. Retrieval Models

For each query, we run three different retrieval models: the Okapi BM25 model [6], the divergence from randomness model [7], the language model using Dirichlet priors [8]. We used the RM3 Positional Relevance model to implement a pseudo-relevance feedback strategy including query expansion [9].

2.1.3. Ranking Fusion

Given different ranking lists, we used the reciprocal ranking fusion (RRF) approach to merge them [10].

2.2. Subtask 2: Weakly Supervised IR

For this subtask, we did not have the time to implement an approach to reformulate/weight the query terms given the provided query training set. Nevertheless, we submitted the same runs of subtasks 1 to provide a kind of baseline for a system that does not use any additional information.

2.3. Document Credibilty Prediction

In this subtask, we reused the runs computed in subtask 1 and grouped them in order to produce a single score for each document. Our simple hypothesis is that documents that have a higher score across different search engines are also more credible. Consider that this naïve approach does not consider any additional information about the provenance of the document.

3. Experiments

In this section, we describe the experimental settings and the results for each subtask.

3.1. Search Engine

For all the experiments, we used the PyTerrier[1] and the Terrier[2] indexes provided by the organizers of the task. We used the default parameter settings for each retrieval model:

- BM25, $k2$ = 1.2, b = 0.75
- LMDirichlet, μ = 2000
- DFR, *basic_model* = if, *after_effect* = b, *normalization* = h2

The RM3 pseudo-relevance feedback model was used with the default parameters.

For the document credibility prediction, we performed a min-max normalization before grouping the scores per document and sum them in order to obtain a single score for scenario 1, or grouping the scores per document-topic for scenario 2.

3.2. Runs

For each subtask, we submitted four runs.

3.2.1. Subtask 1

For the Ad-hoc retrieval subtask, the runs are:

- ims_original_rrf: Reciprocal Rank fusion with BM25, QLM, DFR approaches
- ims_original_rm3_rrf: Reciprocal Rank fusion with BM25, QLM, DFR approaches using RM3 pseudo relevance feedback
- ims_simplified_rrf: Reciprocal rank fusion with BM25, QLM, DFR approaches on manual variants of the query
- ims_simplified_rm3_rrf: Reciprocal Rank fusion with BM25, QLM, DFR approaches on manual variants using RM3 pseudo relevance feedback

3.2.2. Subtask 2

For the Wekly supervised IR, the runs are the same of those of subtask 1.

3.2.3. Subtask 3

For the document credibility prediction, the runs are :

- subtask1_ims_original: it is created by merging with a min-max normalization the runs provided in Task 2 subtask 1 with BM25, QLM, DFR approaches
- subtask1_ims_simplified: it is created by merging with a min-max normalization the runs provided in Task 2 subtask 1 with BM25, QLM, DFR approaches with manual reformulation
- subtask2_ims_original: same as run subtask1
- subtask2_ims_simplified: same as run subtask1

[1] https://pyterrier.readthedocs.io/en/latest/
[2] http://terrier.org

4. Final remarks and Future Work

The aim of our participation to the CLEF 2021 eHealth Task 2 was to test the effectiveness of the reciprocal ranking fusion approach together with a pseudo-relevance feedback strategy. When ground truth will be provided, we will include an analysis of the results.

References

[1] H. Suominen, L. Goeuriot, L. Kelly, L. A. Alemany, E. Bassani, N. Brew-Sam, V. Cotik, D. Filippo, G. González-Sáez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, R. Upadhyay, J. Vivaldi, M. Viviani, C. Xu, Overview of the CLEF eHealth Evaluation Lab 2021, in: CLEF 2021 - 12th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, 2021.

[2] L. Goeuriot, G. Pasi, H. Suominen, E. Bassani, N. Brew-Sam, G. Gonzalez-Saez, R. G. Upadhyay, L. Kelly, P. Mulhem, S. Seneviratne, M. Viviani, C. Xu, Consumer Health Search at CLEF eHealth 2021, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR Workshop Proceedings, 2021.

[3] G. Di Nunzio, S. Marchesin, F. Vezzani, A Study on Reciprocal Ranking Fusion in Consumer Health Search. IMS UniPD ad CLEF eHealth 2020 Task 2, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, vol. 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, URL http://ceur-ws.org/Vol-2696/paper_128.pdf, 2020.

[4] D. Frank Hsu, I. Taksa, Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval, Information Retrieval 8 (3) (2005) 449–480, doi:\let\@tempa\bibinfo@X@doi10.1007/s10791-005-6994-4, URL https://doi.org/10.1007/s10791-005-6994-4.

[5] I. Ruthven, M. Lalmas, A Survey on the Use of Relevance Feedback for Information Access Systems, Knowl. Eng. Rev. 18 (2) (2003) 95–145, ISSN 0269-8889, doi:\let\@tempa\bibinfo@X@doi10.1017/S0269888903000638, URL https://doi.org/10.1017/S0269888903000638.

[6] S. E. Robertson, H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval 3 (4) (2009) 333–389, doi:\let\@tempa\bibinfo@X@doi10.1561/1500000019, URL https://doi.org/10.1561/1500000019.

[7] G. Amati, C. J. Van Rijsbergen, Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness, ACM Trans. Inf. Syst. 20 (4) (2002) 357–389, ISSN 1046-8188, doi:\let\@tempa\bibinfo@X@doi10.1145/582415.582416, URL https://doi.org/10.1145/582415.582416.

[8] C. Zhai, J. Lafferty, A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, Association for Computing Machinery, New York, NY, USA, ISBN 1581133316, 334–342, doi:\let\@tempa\bibinfo@X@doi10.1145/383952.384019, URL https://doi.org/10.1145/383952.384019, 2001.

[9] Y. Lv, C. Zhai, Positional Relevance Model for Pseudo-Relevance Feedback, in: Proceedings

of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, Association for Computing Machinery, New York, NY, USA, ISBN 9781450301534, 579–586, doi:\let\@tempa\bibinfo@X@doi10.1145/1835449.1835546, URL https://doi.org/10.1145/1835449.1835546, 2010.

[10] G. V. Cormack, C. L. A. Clarke, S. Buettcher, Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, Association for Computing Machinery, New York, NY, USA, ISBN 9781605584836, 758–759, doi:\let\@tempa\bibinfo@X@doi10.1145/1571941.1572114, URL https://doi.org/10.1145/1571941.1572114, 2009.