

A Study on Query Expansion and Rank Fusion for Precision Medicine

The IMS Unipd at TREC 2020 Precision Medicine

Giorgio Maria Di Nunzio, Stefano Marchesin

Department of Information Engineering
University of Padua, Italy
{giorgiomaria.dinunzio, stefano.marchesin}@unipd.it

Abstract. In this report, we describe the methodology and the experimental setting of our participation as the IMS Unipd team in TREC PM 2020. The objective of this work is to evaluate a query expansion and ranking fusion approach optimized on the previous years of TREC PM. In particular, we designed a procedure to (1) perform query expansion using a pseudo relevance feedback model on the first k retrieved documents, and (2) apply rank fusion techniques to the rankings produced by the different experimental settings. The results obtained provide interesting insights in terms of the different per-topic effectiveness and will be used for further failure analyses.

Keywords: Precision medicine, query reformulation, rank fusion

1 Introduction

The TREC 2020 Precision Medicine (PM) Track¹ focuses on a relevant use case in clinical decision support: providing useful precision medicine-related information to clinicians treating cancer patients. The topics considered for the task are synthetic cases created with the help of precision oncologists where each case describes the patient's disease (type of cancer), the relevant genetic variants (which genes are mutated), and the proposed treatment. The participants of the track are challenged with the retrieval of biomedical articles that provide evidence for/against the treatment in the specific population.

Our participation to the TREC 2020 PM Track focuses on the evaluation of a mixture of query expansion and rank fusion approaches optimized, in terms of number of term documents used for expansion and different evaluation measures, on previous years collections. The objective of this work is to study whether any combination can improve the precision of the search engine.

In this work, we present the experiments we carried out using a fully automated system based on our previous work that [4]: i) performs query expansion based on pseudo-relevance feedback information; ii) merges the different rankings produced by three retrieval models validated on previous TREC PM collections.

¹ <http://www.trec-cds.org/2020.html>

2 Methodology

In this section, we describe the methodology employed to merge the ranking lists provided by the different retrieval methods using query expansions based on pseudo-relevance feedback.

Query expansion : We used the RM3 model to implement a pseudo-relevance feedback strategy including query expansion [6, 5].

Retrieval models : For each query, we run three different retrieval models: the Okapi BM25 model [7], the divergence from randomness model [1], the language model using Dirichlet priors [9].

Ranking fusion : Given different ranking lists, we used the reciprocal ranking fusion (RRF) approach to merge them [2].

2.1 Parameters Optimization

Since we are using pseudo-relevance feedback (PRF) and query expansion, we wanted to find the best combination of number of terms added to the query and number of documents used for PRF. For this reason, we ran the three retrieval models on the topics and document set of TREC PM 2019 using the following parameters:

- Number of document for the PRF: 5, 10, 30, 50, 100;
- Expand the query using the title or the abstract field;
- Number of terms to add to the query: 1, 3, 5, 10, 30;
- Weight the fields title and abstract with: 0.1, 0.2, 0.3, ..., 1.0;
- Select the best combination in terms of P10, RPrec, InfNDCG.

We selected the best combination of parameters for each retrieval model and for each evaluation measure. These values are reported in Table 1.

3 Experiments

For all the experiments, we used the Elasticsearch search engine² and the indexes provided by the organizers of the task. We used the following parameter settings for each retrieval model:

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

² <https://www.elastic.co/products/elasticsearch>

Table 1: This table presents, for each model, the measure used to optimize the parameters on the TREC PM 2019 collection. The last column reports the name of the run that was produced for TREC PM 2020 with the corresponding parameters.

model	optimized for	# docs	# terms	QE field	title weight	abstract weight	run
BM25	P10	-	-	-	0.0	1.0	bm25_p10
BM25	P10	-	-	-	0.1	0.9	rrf_p10
DFR	P10	-	-	-	0.5	0.2	
QLM	P10	-	-	-	0.1	0.5	
BM25	infNDCG	5	30	title	0.1	0.8	rrf_prf_infndcg
DFR	infNDCG	5	30	title	0.3	0.5	
QLM	infNDCG	10	30	title	0.2	0.7	
BM25	P10	30	30	abstract	0.1	0.9	rrf_prf_p10
DFR	P10	10	30	title	0.5	0.2	
QLM	P10	10	10	abstract	0.1	0.5	
BM25	RPrec	10	30	title	0.2	0.5	rrf_prf_rprec
DFR	RPrec	10	30	title	0.2	0.3	
QLM	RPrec	5	30	title	0.1	0.6	

3.1 Runs

We submitted five runs:

- bm25_p10: Plain BM25 optimized for P10
- rrf_p10: Reciprocal rank fusion with BM25, QLM, DFR without PRF and without query expansion optimized for P10;
- rrf_prf_p10: Reciprocal rank fusion with BM25, QLM, DFR with PRF and query expansion optimized for P10;
- rrf_prf_rprec: Reciprocal rank fusion with BM25, QLM, DFR with PRF and query expansion optimized for RPrec;
- rrf_prf_infndcg: Reciprocal rank fusion with BM25, QLM, DFR with PRF and query expansion optimized for InfNDCG;

3.2 Results

The organizers of the TREC 2020 PM Track provided the summary of the results in terms of best, median, and worst value for each topic for three evaluation measures: inferred NDCG (infNDCG) [8], precision at 10 (P@10), and R-precision (RPrec).

In Table 2, we report the median values of the three measures averaged across topics, as well as the averaged results of the five submitted runs.

In Figures 1a, 2a, 2b, 3a, 3b, we show a barplot that displays, topic by topic, the difference between the performance of each run and the median values of

measure	median	bm25_p10	rrf_p10	rrf_prf_p10	rrf_prf_infndcg	rrf_prf_rprec
infNDCG	0.432	0.421	0.448	0.457	0.461	0.464
RPrec	0.326	0.332	0.323	0.348	0.361	0.345
P@10	0.465	0.481	0.458	0.445	0.461	0.452

Table 2: Overall comparison with average median values of the scientific literature task

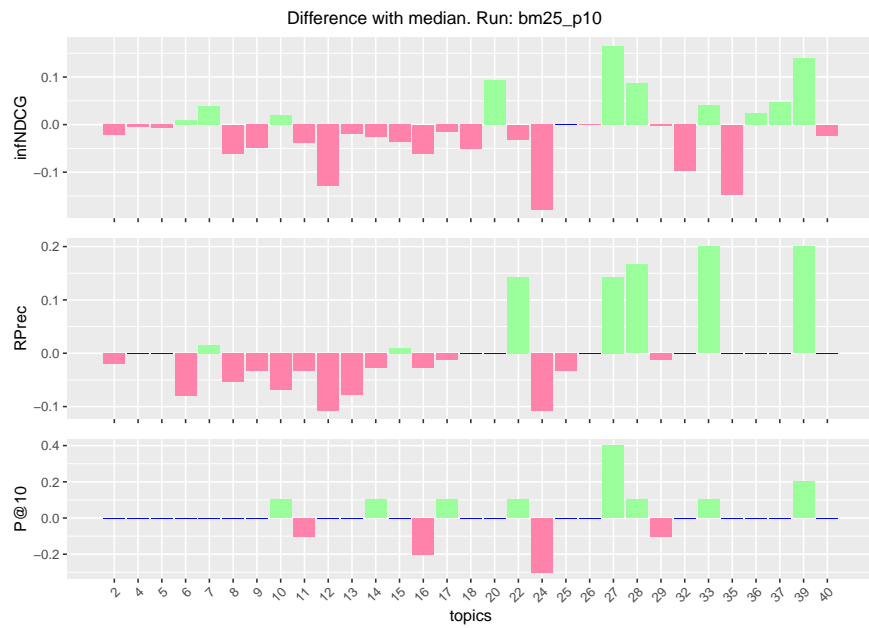
the task. For a positive difference (run better than median), a green barplot is shown, while for a negative difference (run worse than median), a red barplot is shown.

The results show that rank fusion runs achieve higher scores than median values for infNDCG and RPrec, whereas plain BM25 performs better than median for P@10. This suggests that rank fusion runs – relying on PRF to perform query expansion – are recall-oriented rather than precision-oriented. Thus, given the promising results of rank fusion runs, we plan to investigate the integration of re-ranking components in the retrieval pipeline. In particular, we will evaluate the effectiveness of applying BERT [3] to perform re-ranking.

4 Final Remarks

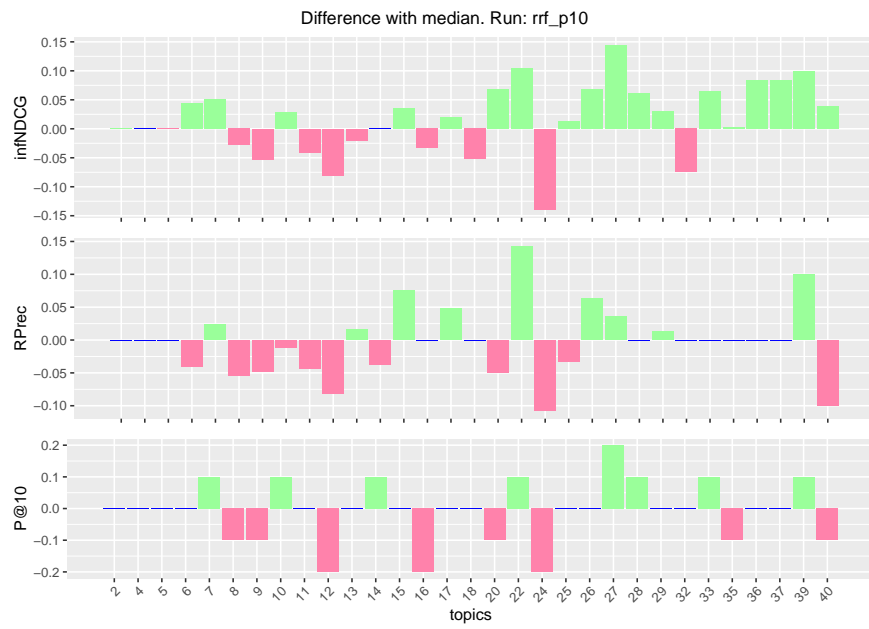
In this paper, we presented the results of our third participation in the TREC PM Track.

The analysis of the results showed the effectiveness of the rank fusion runs for infNDCG and RPrec measures, and the effectiveness of plain BM25 – although optimized – for P@10. The results suggested a recall-oriented nature for rank fusion runs, motivating the integration of a re-ranking component for future work.

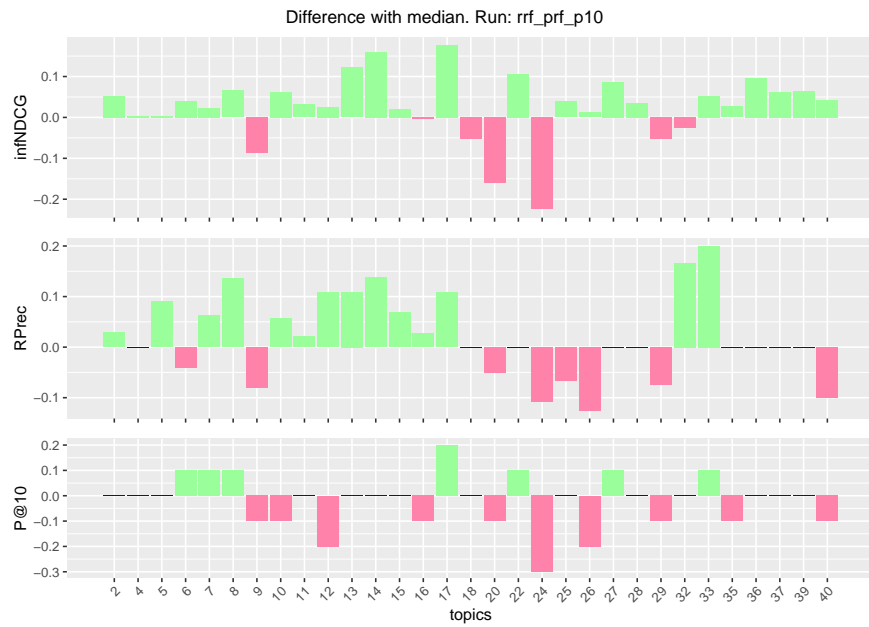


(a)

Fig. 1: Topic by topic difference between runs and median values of the clinical trials task.



(a)

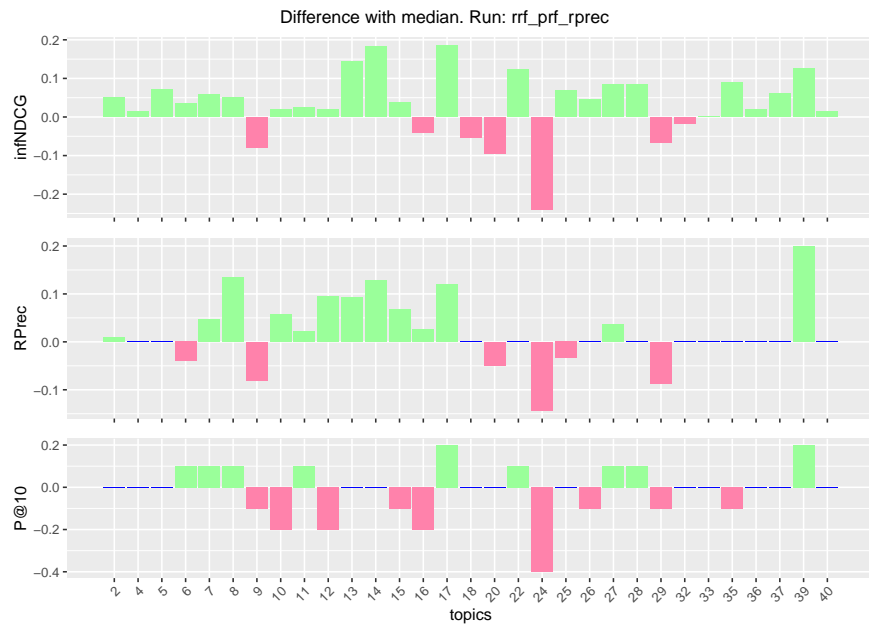


(b)

Fig. 2: Topic by topic difference between runs and median values of the clinical trials task.



(a)



(b)

Fig. 3: Topic by topic difference between runs and median values of the clinical trials task.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.* **20**(4), 357–389 (Oct 2002). <https://doi.org/10.1145/582415.582416>, <https://doi.org/10.1145/582415.582416>
2. Cormack, G.V., Clarke, C.L.A., Buettcher, S.: 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*. pp. 758–759. SIGIR '09, Association for Computing Machinery, New York, NY, USA (2009). <https://doi.org/10.1145/1571941.1572114>, <https://doi.org/10.1145/1571941.1572114>
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. pp. 4171–4186. Association for Computational Linguistics (2019). <https://doi.org/10.18653/v1/n19-1423>, <https://doi.org/10.18653/v1/n19-1423>
4. Di Nunzio, G., Marchesin, S., Vezzani, F.: A study on reciprocal ranking fusion in consumer health search. IMS unipd ad CLEF ehealth 2020 task 2. In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020* (2020), http://ceur-ws.org/Vol-2696/paper_128.pdf
5. Jaleel, N.A., Allan, J., Croft, W.B., Diaz, F., Larkey, L.S., Li, X., Smucker, M.D., Wade, C.: Umass at TREC 2004: Novelty and HARD. In: Voorhees, E.M., Buckland, L.P. (eds.) *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*. NIST Special Publication, vol. 500-261. National Institute of Standards and Technology (NIST) (2004), <http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf>
6. Lavrenko, V., Croft, W.B.: Relevance based language models. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. p. 120–127. SIGIR '01, Association for Computing Machinery, New York, NY, USA (2001). <https://doi.org/10.1145/383952.383972>, <https://doi.org/10.1145/383952.383972>
7. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* **3**(4), 333–389 (2009). <https://doi.org/10.1561/1500000019>, <https://doi.org/10.1561/1500000019>
8. Yilmaz, E., Kanoulas, E., Aslam, J.A.: A Simple and Efficient Sampling Method for Estimating AP and NDCG. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 603–610. SIGIR '08, ACM, New York, NY, USA (2008). <https://doi.org/10.1145/1390334.1390437>, <http://doi.acm.org/10.1145/1390334.1390437>
9. Zhai, C., Lafferty, J.: 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*. pp. 334–342. SIGIR '01, Association for Computing Machinery, New York, NY, USA (2001). <https://doi.org/10.1145/383952.384019>, <https://doi.org/10.1145/383952.384019>