Performance Comparison of Different Query Expansion and Pseudo-Relevance Feedback Methods

A comparison of Bo1, KL, RM3, and Axiomatic query expansion against the BM25

INTRODUCTION

Information retrieval is the technique for querying for a document from a large collection. Query Expansion (QE) and Pseudo Relevance Feedback (PRF) are techniques to potentially get better query results by adding more keywords to the query automatically based on the data in the body of documents and the top document found by an initial query (the pseudo-relevant set). In this research paper, multiple models were put against each other to compare their performance in different situations in order to see if, why, and when these model outperform the BM25 model, which is used as a baseline. The models are tested on a variety of datasets with multiple sets of parameters to get a better overview of how they perform on average.

METHODOLOGY

- Libraries
  - PyTerrier [1]
  - ir_datasets [2]
  - ir-measures [3]
- Datasets
  - MS MARCO-passage [4]
  - ArguAna [5; 6]
  - Antique [7]
  - Deep TREC Learning Docs [4]
- Measures
  - Normalized Discount Cumulative Gain (nDCG) [8]
  - Average Precision (AP) [9]
  - Reciprocal Rank (RR)[10]

- Bo1 Query Expansion [11]
  - Expands with terms that have the right balance of being common and uncommon in the entire body of documents, and weighs that by how common the term is within the pseudo-relevant set.
- KL Query Expansion [11]
  - Expands with terms that are more common in the pseudo-relevant set than in the entirety of the collection of documents, weighted by how common the term is within the pseudo-relevant set.
- RM3 Query Expansion* [12]
  - Expands with terms using the scores that the feed model attributed to the documents from an initial query. For every document in the pseudo-relevant set, it uses these scores weighted by how common the term is within those documents.
- Axiomatic Query Expansion [13]
  - Expands with terms, mainly using the mutual information formula, which calculates how correlated 2 or more terms are. If terms are more correlated to the query terms, it has a greater chance of being used to expand the query.

*This is how it's implemented in pyterrier, which different slightly from the original version

RESEARCH QUESTIONS

The main question to answer is:

What gain in retrieval performance do different QE and PRF methods achieve compared to standard BM25 across different domains and retrieval tasks?

With the following subquestions:

How well do different QE and PRF methods perform on different datasets compared to standard BM25?

How do the different QE and PRF methods compare in terms of execution time and resources?

RESULTS

	RR@10	nDCG@10	AP@10	Time**
BM25	0.6397	0.4795	0.1083	-
Axiomatic	0.6397	0.4795	0.1083	0.0110
Bo1	0.6720	0.5120	0.1239	13.0163
KL	0.6789	0.5048	0.1218	9.8814
RM3	0.6605	0.4975	0.1150	17.6682

Results from the best performing dataset “msmarco-passage/trec-dl-2019/judged” with 50 feedback document and 50 feedback terms

**This time is corrected to not include the time of the initial query and the secondary query after the expansion

ANALYSIS

The Axiomatic model in these and most other results has identical accuracy scores and a negligible amount of execution time, implying that no expansion actually took place in all but one of the datasets, where it performed slightly worse than BM25.

Bo1 and KL are very similar in their internal logic, but have different implementations. This shows in their similarity of test results, with their different execution times coming down to the optimisation of their underlying components under the hood.

RM3 generally performs about as well as BM25, which can be attributed to it's high dependence on its feedback model.

LIMITATIONS

The models have many different sets of parameters that can be tested, as well as the added dimension of testing different feedback models in combination with the query expansion models. To properly test the models to find their most optimal performance on each dataset would take a very long time, which was not feasible for this paper and its time restrictions.

Besides that, the Axiomatic query expansion model had some problems with its performance, likely because of some bug in the source code. This bug did not get fixed in time to test the true performance of the model.

CONCLUSION

BM25 often outperforms query expansion models, but with the right parameters, query expansion can be more accurate.

The Axiomatic model in its current implementation does not seem to work most of the time and is not recommended to use unless fixed.

Bo1 and KL both have very similar results, but KL really shines with smaller datasets with many feedback documents in terms of execution time. Both models do struggle with large datasets.

RM3 doesn't really stand out from these initial experiments, but has more options when it comes to customisability and with enough trial and error could be tuned well for a specific dataset.

CREDITS

Research by: Laurens de Swart
ljpdeswart@student.tudelft.nl
Supervisors: Jurek Leonhardt, Avishek Anand
Examinor: Alan Hanjalic

REFERENCES

[1] C. Macdonald and N. Tonellotto, “Declarative experimentation ininformation retrieval using pyterrier,” in Proceedings of ICTIR 2020, 2020.
[2] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian, “Simplified data wrangling with ir datasets,” in SIGIR, 2021.
[3] S. MacAvaney, C. Macdonald, C. Clarke, B. Piwowarski, and H. Scells, “ir measures,” https://github.com/terrierteam/ir_measures, 2021, a Python library for standardized evaluation of information retrieval metrics.
[4] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, “Ms marco: A human generated machine reading comprehension dataset,” in InCoCoaNIPS, 2016.
[5] H. Wachsmuth, S. Syed, and B. Stein, “Retrieval of the best counterargument without prior topic knowledge,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2018, pp. 241-251. [Online]. Available: http://aclweb.org/anthology/P18-1023
[6] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models,” arXiv preprint arXiv:2104.08663, 4 2021. [Online]. Available: https://arxiv.org/abs/2104.08663
[7] H. Hashemi, M. Aliannejadi, H. Zamani, and B. Croft, “Antique: A non-factoid question answering benchmark,” in ECIR, 2020.
[8] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques,” ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422-446, 2002. [Online]. Available: http://doi.acm.org/10.1145/582415.582418
[9] D. Harman, “Evaluation issues in information retrieval,” Information Processing and Management, vol. 28, no. 4, pp. 439-440, 1992.
[10] P. Kantor and E. Voorhees, “The trec-5 confusion track,” Information Retrieval, vol. 2, no. 2-3, pp. 165-176, 2000.
[11] G. Amati, “Probability models for information retrieval based on divergence from randomness,” 2003.
[12] N. Jaleel, J. Allan, W. Croft, F. Diaz, L. Larkey, X. Li, M. Smucker, and C. Wade, “Umass at trec 2004: Novelty and hard,” 01 2004.
[13] P. Yang and J. Lin, Reproducing and Generalizing Semantic Term Matching in Axiomatic Information Retrieval, 04 2019, pp. 369-381.