A Comparative Study of Bo1 and RM3 Query Expansion using DPH Retrieval

Albert Gesk[1], Lilian Erler[1]

[1]University Kassel, Germany

Abstract

Query expansion via pseudo-relevance feedback is a well-established technique to improve retrieval effectiveness in information retrieval systems. Among classical approaches, Bo1 and RM3 are frequently used but exhibit varying effectiveness depending on the retrieval setting. In this work, we empirically compare Bo1 and RM3 query expansion within a controlled Retriever-Rewriter-Retriever pipeline using the DPH retrieval model. Experiments are conducted on the *radboud-validation-20251114-training* dataset using nDCG@10 as evaluation metric. Statistical significance testing shows that neither a significant difference nor a significant improvement in retrieval effectiveness can be observed between Bo1 and RM3 under identical parameter settings.

Keywords

Information Retrieval, Query Expansion, Bo1, RM3, PyTerrier

1. Introduction

Query expansion using pseudo-relevance feedback (PRF) is a classical approach to mitigate vocabulary mismatch in information retrieval (IR) systems. The mechanism of pseudo-relevance feedback (PRF) involves extracting expansion terms from the initial top-ranked documents to formulate a new query for a second retrieval stage [1]. Among commonly used PRF techniques, Bo1 and RM3 are widely adopted in traditional IR systems and toolkits.

Despite their popularity, the relative effectiveness of Bo1 and RM3 is known to be sensitive to retrieval models, parameter settings, and datasets. This motivates a controlled comparison of both methods under identical experimental conditions.

In this paper, we investigate the following research question: *Does the choice of Bo1 versus RM3 query expansion lead to significant differences or improvements in retrieval effectiveness when applied within a DPH-based retrieval pipeline?*

Our contributions are: (i) a controlled experimental comparison of Bo1 and RM3 query expansion using identical retrieval and feedback settings, and (ii) a statistical analysis of their impact on nDCG@10 on the *radboud-validation-20251114-training* dataset.

2. Related Work

Pseudo-relevance feedback (PRF) is a well-established technique in information retrieval to improve query effectiveness by automatically expanding the original query with terms extracted from top-ranked documents [1].

Previous studies have shown that RM3 provides robust improvements over baseline retrieval models like BM25, particularly when the number of feedback documents and expansion terms are carefully tuned. However, the quality of the expansion terms can be limited because of it's reliance on pseudo- relevant documents. This happens especially in scenarios where the top-ranked documents are not truly relevant. [2]

These findings motivate a controlled comparison between RM3 and Bo1, as they represent alternative pseudo-relevance feedback methods based on distinct term selection principles. Bo1 is based on divergence-from-randomness (DRF) and selects terms that are statistically informative in the top-retrieved documents [3], while RM3 is a probabilistic relevance model [2].

Other works have explored hybrid or transformer-based retrieval pipelines that incorporate PRF or developed a demanding pipelines, such as ColBERT-PRF, which integrates semantic pseudo-relevance feedback into dense retrieval models [4]. But controlled comparisons of classical Bo1 and RM3 within identical DPH-based pipelines remain limited.

This motivates our work to conduct a systematic and controlled evaluation of Bo1 and RM3 under identical experimental conditions, assessing not only retrieval effectiveness but also statistical significance of the observed differences.

3. Methodology

This section depicts the experimental setup designed to answer the research question.

3.1. Dataset and Preprocessing

For the experiments, we utilized the *radboud-validation-20251114-training* dataset. The indexing of the documents was based on the standard textual representation (`default_text`) provided in the dataset. No additional preprocessing steps or filtering were applied, meaning only the standard tokenization performed by PyTerrier was used.

3.2. Experimental Design

The experiment use a retriever-rewriter-retriever pipeline as foundation. The DPH retrieval model [5] implemented via PyTerrier is setup with a thousand maximum number of results to return per query and with no metadata fields to return for each search result. The pseudo-relevance feedback models,both the Bo1 [6] and RM3 [7] are implemented via PyTerrier as well. For both models, all parameters are remained at default, the number of feedback terms to use is set to ten and the number of feedback documents to use is set to three. The interpolation weight between the original query and the feedback model for RM3 is set to 0.6.

IR Lab Course, Winter Semester 2025

uk071960@uni-kassel.de (A. Gesk);
uk076601@student.uni-kassel.de (L. Erler)
https://github.com/AlbertGesk (A. Gesk);
https://github.com/lilianerler (L. Erler)

To evaluate the effectiveness of the retrieval pipelines, differences in per-topic nDCG@10 scores were assessed for statistical significance using a paired Student's t-test. The significance level was set to $\alpha = 0.05$, and Bonferroni correction was applied to account for multiple comparisons.

3.3. Hypotheses and Null-Hypotheses

Based on the research question, two hypotheses and their corresponding null hypotheses were formulated:

- H_1 Given query expansion via pseudo-relevance feedback using the top 3 retrieved documents and 10 expansion terms, there is a statistically significant difference in nDCG@10 on radboud-validation-20251114-training between (i) the DPH-based retrieval pipeline described in subsection 3.2 three with Bo1 query expansion and (ii) the same pipeline with RM3 query expansion. (α=0.05)
- H_1^0 Given query expansion via pseudo-relevance feedback using the top 3 retrieved documents and 10 expansion terms, there is no statistically significant difference in nDCG@10 on radboud-validation-20251114-training between (i) the DPH-based retrieval pipeline described in subsection 3.2 three with Bo1 query expansion and (ii) the same pipeline with RM3 query expansion. (α=0.05)
- H_2 Given query expansion via pseudo-relevance feedback using the top 3 retrieved documents and 10 expansion terms, there is a statistically significant improvement in nDCG@10 on radboud-validation-20251114-training between (i) the DPH-based retrieval pipeline described in subsection 3.2 three with Bo1 query expansion and (ii) the same pipeline with RM3 query expansion. (α=0.05)
- H_2^0 Given query expansion via pseudo-relevance feedback using the top 3 retrieved documents and 10 expansion terms, there is no statistically significant improvement in nDCG@10 on radboud-validation-20251114-training between (i) the DPH-based retrieval pipeline described in subsection 3.2 three with Bo1 query expansion and (ii) the same pipeline with RM3 query expansion. (α=0.05)

4. Results

Table 1 presents the retrieval effectiveness and statistical testing results.

The DPH-Bo1-DPH pipeline achieves a higher mean nDCG@10 score than DPH-RM3-DPH. However, Bo1 does not significantly outperform RM3. The obtained p-value (0.283) is substantially greater than the significance threshold of 0.05, indicating that the observed difference is likely due to random variation across topics. Consequently, we fail to reject the null hypotheses H_1^0 and H_2^0, and therefore find no statistical evidence to support the alternative hypotheses H_1 and H_2.

Table 1
Retrieval effectiveness and statistical significance (nDCG@10).

Method	nDCG@10	p-value
DPH-Bo1-DPH	0.4947	0.283
DPH-RM3-DPH	0.4744	

5. Conclusion

his study presented a controlled comparison of Bo1 and RM3 pseudo-relevance feedback methods within an identical DPH-based retriever–rewriter–retriever pipeline. Using three feedback documents and ten expansion terms, both approaches yielded comparable retrieval effectiveness on the *radboud-validation-20251114-training* dataset. Although the DPH-Bo1-DPH pipeline achieved a slightly higher mean nDCG@10 score than DPH-RM3-DPH, statistical significance testing showed that this difference was not significant. Consequently, neither method can be considered superior under the examined experimental conditions.

These findings suggest that, when applied with identical parameters and retrieval models, the choice between Bo1 and RM3 has limited impact on retrieval effectiveness. Future work could explore broader parameter sweeps, alternative feedback depths, or different retrieval models to better understand under which conditions one method may consistently outperform the other.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-5.2 and LanguageTool in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Junmei, A knowledge-based approach for pseudo-relevance feedback by exploiting semantic relevance, Ph.D. thesis, 2025. URL: https://doi.org/10.1007/s10115-025-02581-5.

[2] N. Sinhababu, R. Khatun, Leq: Large language models generate expanded queries for searching, in: 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024, pp. 1–4. doi:10.1109/ICCCNT61001.2024.10725314.

[3] L. de Swart, Performance comparison of different query expansion and pseudo-relevance feedback methods (2024).

[4] X. Wang, C. MacDonald, N. Tonellotto, I. Ounis, Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval, ACM Trans. Web 17 (2023). URL: https://doi.org/10.1145/3572405. doi:10.1145/3572405.

[5] G. Amati, Frequentist and bayesian approach to information retrieval, in: M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, A. Yavlinsky (Eds.), Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings, volume 3936 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 13–24. URL: https://doi.org/10.1007/11735106_3. doi:10.1007/11735106_3.

[6] G. Amati, Probability models for information retrieval based on divergence from randomness, Ph.D. thesis, University of Glasgow, UK, 2003. URL: http://theses.gla.ac.uk/1570/.

[7] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, C. Wade, Umass at TREC 2004: Novelty and HARD, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004. URL: http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf.