Open Web Search at LongEval 2023: Reciprocal Rank Fusion on Automatically Generated Query Variants

Notebook for the LongEval Lab at CLEF 2023

Maik Fröbe[1], Gijs Hendriksen[2], Arjen P. de Vries[2] and Martin Potthast[3]

[1]*Friedrich-Schiller-Universität Jena*
[2]*Radboud Universiteit Nijmegen*
[3]*Leipzig University and ScaDS.AI*

Abstract

We describe the participation of the Open Web Search (OWS) team in the shared task LongEval hosted at CLEF 2023. Our submission is motivated by previous observations on static test collections that fusing search results for multiple user-generated query variants for the same information need improves the overall retrieval effectiveness. However, it is unclear whether these improvements are robust over time, or to what extent the query variant generation needs to be adjusted if queries and relevance assessments change. We evaluate the robustness of query variants generated by ChatGPT by optimizing the prompt and other retrieval hyperparameters using training data collected in June 2022, and evaluating the rank fusion effectiveness on test data from July and September 2022, respectively.

1. Introduction

The way an information need is formulated as a query to a search engine has a significant impact on its retrieval effectiveness. The same retrieval system can be very effective or ineffective for the same information need, depending on how the query is formulated [1]. This variance in retrieval effectiveness for different variants of queries has led to related research [2, 3, 4, 5, 6], and query variants have been created and studied for many commonly used static test collections in IR, e.g., the TREC Common Core [7], Robust04 [8], and the Web tracks [9, 10, 11, 12, 13, 14]. However, these studies of query variants have been conducted exclusively with static test collections, so it is not yet clear how query variants, their generation, and rank fusion techniques for their search results behave and how they may need to evolve over time. Since intentions and preferences can change considerably over time [15], retrieval pipelines using query variants may need to do so as well. For example, the effectiveness of a rank fusion technique that worked well at one point may drop a month later. We investigate the temporal robustness of query variants in the shared task LongEval at CLEF 2023 [16] using the LongEval retrieval collection [17].

In previous research on query variants, experts are usually employed to formulate meaningful queries on a given topic [4, 5]. Based on the description of a topic, the experts then formulate

---

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

maik.froebe@uni-jena.de (M. Fröbe); g.hendriksen@cs.ru.nl (G. Hendriksen)
0000-0002-1003-981X (M. Fröbe); 0000-0003-0945-3148 (G. Hendriksen); 0000-0002-2888-4202 (Arjen P. de Vries); 0000-0003-2451-0665 (M. Potthast)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

queries they would have sent to a search engine to retrieve relevant documents without seeing the actual query (i.e., the title of the topic). This way, multiple queries for the same information need are collected and their retrieval results then merged using reciprocal rank fusion. This yields a higher retrieval effectiveness compared to using only the original query [4, 5]. However, query variants are usually costly to generate and have therefore only been created for corpora with few topics, e.g., up to a few hundred information needs. Since the LongEval retrieval collection contains several thousand queries, we avoid the manual generation of query variants by using ChatGPT instead. This has the added benefits that we can systematically experiment with different ways of instructing ChatGPT on how to formulate them, and that the process can be fully automated. The manual generation of query variants constitutes the submission of a manual run in a shared task, while ours is an automatic run. We use the June 2022 training data to optimize the retrieval hyperparameters (i.e., retrieval model, prompt, number of query variants generated) and test how robust the optimized hyperparameters are on test data collected in July and September 2022, respectively.

We make all code, the query variants generated by ChatGPT, and Docker images implementing our approaches publicly available.[1] To improve reproducibility, we added the LongEval dataset to ir_datasets [18] and ran all our approaches using The Information Retrieval Experiment Platform (TIREx) [19] hosted at the TIRA Integrated Research Architecture [20].

2. Automatically Generated Query Variants with ChatGPT

The LongEval collection consists of thousands of queries, so that manually generating variants for all of them is infeasible. Therefore, we use ChatGPT to automatically generate query variants. Using the LongEval training data, we experimented with two manually developed prompts, examining the responses. Table 1 shows the prompts and ChatGPT's responses for the query "`4k video downloader`", respectively. The top prompt resulted from several rounds of manual prompt engineering and was used to collect responses for all LongEval queries. Based on issues we observed in the training data along the way, an improved prompt (Table 1, bottom) was devised to collect another set of query variants. For both prompts, the query variants are extracted from ChatGTP's response using regular expressions.

Based on the query variants, we optimized the three hyperparameters (1) retrieval model, (2) number of query variants (i.e., 3, 5, and 10 variants), and (3) prompt used (e.g., Prompt 1 or Pprompt 2) on the training data. Reciprocal rank fusion was used to combine the rankings of different query variants. Retrieval models were taken in their default configurations from PyTerrier [21]. Five runs were selected for submission using the top hyperparameter settings:

- *Runs "ows-bm25-5-variants-prompt-2" and "ows-bm25-10-variants-prompt-2".* These runs are based on PyTerrier's default BM25 retrieval model. They fuse six (eleven) rankings per topic using reciprocal rank fusion on the rankings obtained with the original query and five (ten) query variants generated with Prompt 2.
- *Run "ows-lgd-10-variants-prompt-2".* This run is based on PyTerrier's default LGD retrieval model, fusing eleven rankings per topic obtained with the original query and ten query variants generated with Prompt 2.

---

[1] https://github.com/webis-de/LONGEVAL-23

Table 1
Our two prompts and examples of query variants generated with ChatGPT given a LongEval query.

Number	1
Prompt	I wanted to search on google for "<ORIGINAL_QUERY>" but I was not able to retrieve relevant documents. Please recommend me keywords that I should additionally add to my query.
Example Input	I wanted to search on google for "4kvideo downloader" but I was not able to retrieve relevant documents. Please recommend me keywords that I should additionally add to my query.
Example Response	Sure, here are some additional keywords that you can use to refine your search for "4k video downloader": (1) Best 4k video downloader, (2) Free 4k video downloader, (3) Youtube 4k video downloader, (4) Windows/Mac 4k video downloader, (5) Chrome/Firefox 4k video downloader extension, (6) Reliable 4k video downloader, (7) Fast 4k video downloader, (8) 4k video downloader tutorial/review. I hope these suggestions help optimize your search!
Number	2
Prompt	I wanted to search the web for "<ORIGINAL_QUERY>" but I was not able to find relevant documents. Please recommend me up to ten better queries in a numbered list that retrieve relevant documents. All Queries must be English and have at least two words.
Example Input	I wanted to search the web for "4kvideo downloader" but I was not able to find relevant documents. Please recommend me up to ten better queries in a numbered list that retrieve relevant documents. All Queries must be English and have at least two words.
Example Response	1. "Best 4k video downloader", 2. "4k video download software", 3. "How to download 4k videos", 4. "Top-rated video downloader for 4k", 5. "4k video downloader for Mac/Windows", 6. "Free 4k video download tools", 7. "Recommended 4k video downloader apps", 8. "Download 4k videos from YouTube", 9. "4k video downloader review", 10. "Fastest 4k video downloader".

- *Runs "ows-pl2-5-variants-prompt-2" and "ows-pl2-10-variants-prompt-2".* These runs are based on PyTerrier's default PL2 retrieval model. They fuse six (eleven) rankings per topic obtained with the original query and five (ten) additional query variants generated with Prompt 2.

3. Evaluation

Table 2 shows the effectiveness of our five submitted runs and the BM25 baseline. We report the percentage of unjudged documents, the nDCG@10, and the condensed nDCG@10 (unjudged documents are removed from the ranking) for the June hold-out dataset and the July and September test sets. Relevance scores for all three months are derived from click logs via click models [16], so unjudged documents have a significant impact, as all runs have between 76.0% (BM25 in September) and 80.3% (ows-pl2-5-variant-prompt-2 in June) unjudged documents.

Table 2
The effectiveness of the five submitted runs and the BM25 baseline on the June, July, and September 2022 test sets, respectively. We report the proportion of unjudged documents, the nDCG@10, and the nDCG@10 when unjudged documents are removed from the ranking (Cond. nDCG@10).

Approach / Run	**Unjudged@10**			**nDCG@10**			**Cond. nDCG@10**		
	June	July	Sep.	June	July	Sep.	June	July	Sep.
ows-bm25-5-variants-prompt-2	0.791	0.798	0.792	0.155	0.166	0.165	0.441	0.474	0.473
ows-bm25-10-variants-prompt-2	0.793	0.795	0.786	0.152	0.167	0.170	0.443	0.472	0.479
ows-lgd-10-variants-prompt-2	0.796	0.799	0.791	0.161	0.163	0.166	**0.446**	0.475	0.483
ows-pl2-5-variants-prompt-2	0.803	0.800	0.790	0.158	0.163	0.164	0.440	0.479	0.486
ows-pl2-10-variants-prompt-2	0.801	0.799	0.788	**0.163**	0.164	0.167	0.440	**0.480**	**0.488**
BM25	0.782	0.771	0.760	**0.163**	**0.180**	**0.184**	**0.446**	0.476	0.478

The assumption that unjudged documents are not relevant may substantially underestimate the nDCG@10, so we also report the condensed nDCG@10 [22]. Condensed nDCG typically overestimates actual effectiveness [23, 24], so both can be used as lower and upper bound estimates to some extent. Overall, we find that the BM25 baseline improves in effectiveness over time, both in terms of nDCG@10 (from 0.163 in June to 0.180 in July and 0.184 in September) and condensed nDCG@10 (from 0.446 in June to 0.476 in July and 0.478 in September). This observation is also nearly stable for all automatically generated query variants, with one exception: the effectiveness of the ows-bm25-5-variants-prompt-2 run peaks in July, while all other runs show the same trend as the BM25 baseline. Therefore, query variant generation may indeed need adjustment over time, but most changes in effectiveness seem negligible.

3.1. Query Length

Due to the added latency of generating query variants with ChatGPT, a practical retrieval system might first assess whether query variants are necessary to effectively answer the query. If the query is already well specified, this may not be the case, and the retrieval system can speed up query processing by deciding not to generate and use the query variants. One possible heuristic for determining whether a query is detailed enough for accurate retrieval is to look at the length of the query. Intuitively, shorter queries might be more vague and insufficiently specified than longer ones, so they might benefit more from the inclusion of query variants.

To test this hypothesis, we examine the impact of query length on the effectiveness of our submitted systems relative to the BM25 baseline. Due to the relatively small number of queries longer than six tokens in the hold-out and test sets (see Figure 1), we limit our analysis to queries with six tokens or less. The results are shown in Figure 2 and indicate that there appears to be no clear correlation between query length and the gain (or loss) in effectiveness from including query variants. Our systems follow the same trends as the BM25 baseline, suggesting that there is no advantage to using query variants for queries with lengths up to six tokens. This is consistent with observations for query performance prediction, where query length has been found not to be highly correlated with the effectiveness of a retrieval system [25] (e.g., because longer queries may contain more noise that distracts or confuses the retrieval system [26]).

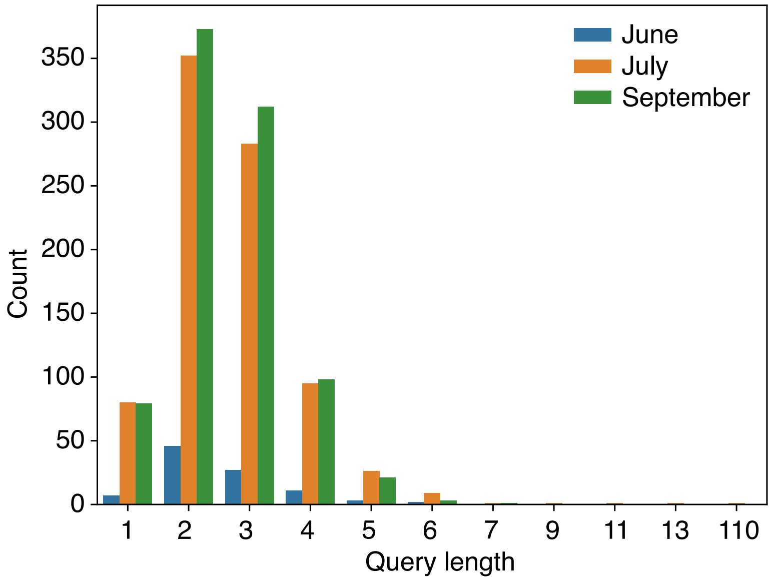

Figure 1: Query length distributions of the LongEval training (June) and test sets (July and September).

4. Conclusion

We presented the Open Web Search (OWS) team's submission to the LongEval shared task at CLEF 2023. Our motivation was to test whether automatically generated query variants are stable in their retrieval effectiveness over time. We automatically formulated query variants with ChatGPT using different prompts. The query variant generation prompt and retrieval hyperparameters were optimized using the June 2022 training data and evaluated on the July and September 2022 test data. We found that the retrieval effectiveness of automatically generated query variants largely follows the trend of the underlying retrieval model. Interesting directions for future research might focus on expanding the experiments to more retrieval collections, and making the ChatGPT prompts more dependent on the time so that drifts in queries can have a larger impact.

Acknowledgments

This work has been partially supported by the OpenWebSearch.eu project (funded by the EU; GA 101070014).

References

[1] N. Arabzadeh, A. Bigdeli, S. Seyedsalehi, M. Zihayat, E. Bagheri, Matches made in heaven: Toolkit and large-scale datasets for supervised query reformulation, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland,

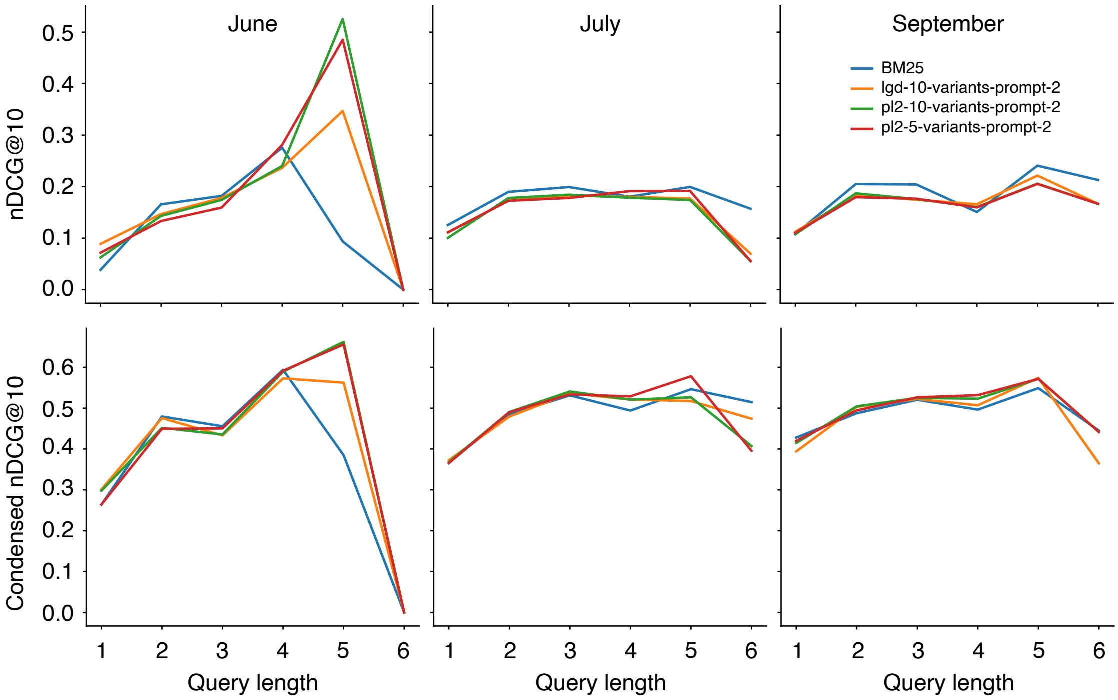

Figure 2: Effectiveness of our systems and the BM25 baseline as a function of query length. For readability, only the top three systems are shown, with the other two showing similar trends.

Australia, November 1 - 5, 2021, ACM, 2021, pp. 4417–4425. doi:10.1145/3459637.3482009.

[2] M. Alaofi, L. Gallagher, M. Sanderson, F. Scholer, P. Thomas, Can Generative LLMs Create Query Variants for Test Collections?, in: 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023.

[3] P. Bailey, A. Moffat, F. Scholer, P. Thomas, UQV100: A test collection with query variability, in: R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, J. Zobel (Eds.), Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016, ACM, 2016, pp. 725–728. doi:10.1145/2911451.2914671.

[4] R. Benham, L. Gallagher, J. M. Mackenzie, T. T. Damessie, R. Chen, F. Scholer, A. Moffat, J. S. Culpepper, RMIT at the 2017 TREC CORE track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017, volume 500-324 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2017.

[5] R. Benham, L. Gallagher, J. M. Mackenzie, B. Liu, X. Lu, F. Scholer, J. S. Culpepper, A. Moffat, RMIT at the 2018 TREC CORE track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018, volume 500-331 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2018.

[6] T. Breuer, N. Fuhr, P. Schaer, Validating simulations of user query variants, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I, volume 13185 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 80–94.

[7] J. Allan, D. Harman, E. Kanoulas, D. Li, C. V. Gysel, E. M. Voorhees, TREC 2017 common core track overview, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017, volume 500-324 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2017.

[8] E. Voorhees, Overview of the trec 2004 robust retrieval track, in: TREC, 2004.

[9] C. L. A. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2009 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009, volume 500-278 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2009.

[10] C. L. A. Clarke, N. Craswell, I. Soboroff, G. V. Cormack, Overview of the TREC 2010 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010, volume 500-294 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2010.

[11] C. L. A. Clarke, N. Craswell, I. Soboroff, E. M. Voorhees, Overview of the TREC 2011 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011, volume 500-296 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2011.

[12] C. L. A. Clarke, N. Craswell, E. M. Voorhees, Overview of the TREC 2012 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012, volume 500-298 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2012.

[13] K. Collins-Thompson, P. N. Bennett, F. Diaz, C. Clarke, E. M. Voorhees, TREC 2013 web track overview, in: E. M. Voorhees (Ed.), Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013, volume 500-302 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2013.

[14] K. Collins-Thompson, C. Macdonald, P. N. Bennett, F. Diaz, E. M. Voorhees, TREC 2014 web track overview, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014, volume 500-308 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2014.

[15] J. Huang, H. Oosterhuis, M. de Rijke, It is different when items are older: Debiasing recommendations when selection bias and user preferences are dynamic, in: K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, J. Tang (Eds.), WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February

21 - 25, 2022, ACM, 2022, pp. 381–389.

[16] R. Alkhalifa, I. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, G. Gonzalez-Saez, P. Galuscakova, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, H. T. Madabushi, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Overview of the clef-2023 longeval lab on longitudinal evaluation of model performance, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science (LNCS), Springer, Thessaliniki, Greece, 2023.

[17] P. Galuscáková, R. Deveaud, G. G. Sáez, P. Mulhem, L. Goeuriot, F. Piroi, M. Popel, Longeval-retrieval: French-english dynamic test collection for continuous web search evaluation, CoRR abs/2303.03229 (2023). `arXiv:2303.03229`.

[18] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with ir_datasets, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2429–2436.

[19] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023.

[20] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:`10.1007/978-3-031-28241-6_20`.

[21] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, Pyterrier: Declarative experimentation in python from BM25 to dense retrieval, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 4526–4533.

[22] T. Sakai, Alternatives to bpref, in: W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, ACM, 2007, pp. 71–78.

[23] T. Sakai, Comparing metrics across TREC and NTCIR: the robustness to system bias, in: J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K. Choi, A. Chowdhury (Eds.), Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008, ACM, 2008, pp. 581–590.

[24] M. Fröbe, L. Gienapp, M. Potthast, M. Hagen, Bootstrapped nDCG Estimation in the Presence of Unjudged Documents, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023.

[25] B. He, I. Ounis, Query performance prediction, Information Systems 31 (2006) 585–594.

[26] G. Kumaran, J. Allan, Effective and efficient user interaction for long queries, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 11–18.