Counterfactual Query Rewriting to Use Historical Relevance Feedback

Jüri Keller[1], Maik Fröbe[2], Gijs Hendriksen[3], Daria Alexander[3], Martin Potthast[4], Matthias Hagen[2], and Philipp Schaer[1]

[1] TH Köln - University of Applied Sciences, Cologne, Germany
[2] Friedrich-Schiller-Universität Jena
[3] Radboud Universiteit Nijmegen
[4] University of Kassel, hessian.AI, ScaDS.AI

Abstract. When a retrieval system receives a query it has encountered before, previous relevance feedback, such as clicks or explicit judgments can help to improve retrieval results. However, the content of a previously relevant document may have changed, or the document might not be available anymore. Despite this evolved corpus, we counterfactually use these previously relevant documents as relevance signals. In this paper we proposed approaches to rewrite user queries and compare them against a system that directly uses the previous qrels for the ranking. We expand queries with terms extracted from the previously relevant documents or derive so-called keyqueries that rank the previously relevant documents to the top of the current corpus. Our evaluation in the CLEF LongEval scenario shows that rewriting queries with historical relevance feedback improves the retrieval effectiveness and even outperforms computationally expensive transformer-based approaches.

Keywords: Query Rewriting · Keyqueries · Longitudinal Evaluation

1 Introduction

Many queries received by a search engine have been seen before [29]. Analyzing how users interact with the results of these queries provides valuable relevance indicators that can improve the search engine's effectiveness, e.g., by constructing click models that synthesize a relevance indicator [7]. These synthesized labels can be used to boost documents [21], to train learning-to-rank models [24], or to fine-tune transformer models [23]. While the training and fine-tuning of models often require huge amounts of labeled data, boosting also works with few labels [21]. However, it may only be effective shortly after relevance information has been observed, as queries, documents, and relevance may evolve [20]. For instance, boosting would fail when documents are deleted or change their content such that they are not relevant to the query anymore (exemplified in Table 1).

To address this challenge, we explore how historical relevance feedback can guide query rewriting approaches. Incorporating the historical relevance feedback

Table 1. Historical relevance feedback observed at timestamp t_0 for a query $q :=$ `bird song` that is to be applied at timestamp t_1. Transferring t_0 to t_1 would introduce errors, whereas our counterfactual query rewriting always uses the correct t_0 observations.

Document		**Comment**	**Type**
Timestamp t_0	Timestamp t_1		
Alphabetical list of *bird songs* you may like ...	—	Document deleted from the web	`DELETE`
Best phone ring tone? Enjoy *bird songs* ...	Get phone tones from the charts for free ...	Document became non-relevant	`UPDATE`
311 *songs by birds* from France by species ...	312 *songs by birds* from France by species ...	Document remains relevant	`UPDATE`

into query rewriting works, analogously to boosting, already with few feedback documents (e.g., RM3 in PyTerrier uses 3 feedback documents as default [25]). The document corpus naturally evolves over time for various reasons, such as newly created content, updated websites, or the removal of outdated information. While the conventional retrieval setting relies on the factual, current state of the collection. In contrast, we additionally use previous versions of the collection and thereby counterfactually assume that superseded documents remain relevant or are at least suitable for constructing relevance indicators. This counterfactual assumption is motivated by the observation that many document changes are incremental and may not substantially alter their relevance to the original query.

We apply our counterfactual relevance feedback in three scenarios: (1) via boosting, (2) via explicit relevance feedback, and (3) via keyqueries. Boosting re-weights known (non-)relevant documents, which can not generalize to new or deleted documents. Explicit relevance feedback extends the query with terms from known relevant documents but does not test the resulting ranking. Therefore, we use so-called keyqueries [17] which reformulate queries until the target documents appear in the top positions.

We evaluate our counterfactual query rewriting approaches on the LongEval test collection. Our results show that counterfactual query rewriting is as effective as boosting but generalizes to new documents, making it more robust than boosting and substantially outperforming neural transformers while being much more efficient as queries can be pre-computed. Our code is publicly available.[5]

2 Related Work

Web Dynamics. Temporal dynamics in web search are an established research topic. Websites change constantly, often more than hourly [1], making them relevant for only a limited time [32]. This relates directly to the observation that

[5] `https://github.com/webis-de/ECIR-25`

many queries are not unique but frequently reissued [8,29]. Even the same users tend to repeat the same queries at different points in time [33].

Temporal Information Retrieval. The observed dynamics motivate Temporal Information Retrieval (TIR) aiming to use temporal information to improve the ranking quality [19,6], e.g., by using temporal patterns for the term weighting [9]. While TIR focuses on leveraging temporal properties, our work takes a complementary approach by exploiting past document versions rather than directly addressing their temporal aspects.

Query Rewriting with Keyqueries. Given a set of target documents, a *keyquery* is the minimal query that retrieves the target documents in the top positions [17,18]. We adapt this approach to the web search setting by using previously relevant documents as target documents. Keyqueries use terms generated via RM3 or other query expansion approaches as vocabulary for an efficient enumeration of query candidates [12,14]. Beyond other query expansion approaches, the keyquery approach also generates a ranking for each candidate to test if all criteria are fulfilled and thereby fully leveraging historical data.

Evaluations in Dynamic Settings. Although temporal dynamics can strongly influence the effectiveness of IR systems, they are rarely considered during evaluation. Soboroff [30] initially investigated how temporal dynamics influence test collection evaluations and hypothesized how they could be maintained. Fröbe et al. [10] studied the case when relevance judgments are re-used between different snapshots of crawled documents. Recently, the LongEval shared task [2,3] provides a test bed of an evolving web search scenario covering over a year. Changes in evolving test collections are described through create, update, and delete operations on documents, topics, and relevance judgments [20]. While the LongEval dataset contains changes in all types of components, this study is mainly concerned with changing documents and ignore, counterfactually, changes in the relevance label.

3 Query Rewriting on Historical Relevance Feedback

We present three approaches that incorporate historical relevance feedback, from (1) boosting (can not generalize), over (2) relevance feedback (might underfit), towards (3) keyqueries (trade-off under- vs. overfitting). These approaches were previously applied in diverse retrieval scenarios (see Section 2). However, we are the first to adapt them to a retrieval scenario that evolves over time.

For a set H of historical relevance feedback with observations $(q, d, t) \in H$, $rel(q, d, t)$ is the (graded) relevance judgment for document d for query q at timestamp t. Document d may have been deleted or substantially changed in the current corpus (Table 1), which is why we counterfactually use the version of the document at the timestamp t where the relevance observation was made.

Boosting Previously Relevant Documents. Given a ranking r and a set of historical relevance feedback H, boosting incorporates a document's historical relevance label, regardless of how much the document has changed or how old the relevance feedback is. We apply boosting to adjust the scores of documents based on their historical relevance. For a document d previously observed for the query q at the timestamps $t_1, \ldots, t_k$, we increase the score if it is relevant, respectively decrease the score if it is not-relevant at the corresponding timestamps using a weighting factor λ^2 and additionally boost highly relevant documents by a factor of μ. The boosted score of document d for query q is then computed by:

$$\mathrm{score}(q,d) = \mathrm{score}_0 \times \prod_{t=t_1}^{t_k} \begin{cases} (1-\lambda)^2, & \text{if } rel(q,d,t) = 0, \\ \lambda^2, & \text{if } rel(q,d,t) = 1, \\ \lambda^2\mu, & \text{if } rel(q,d,t) = 2. \end{cases} \tag{1}$$

While this qrel boosting is highly effective when documents do not change [4,21], it cannot generalize to newly created or deleted documents.

Previously Relevant Documents as Relevance Feedback. Given a retrieval model, the current document corpus D, and historical relevance feedback H, we expand each query by adding k terms with the highest tf-idf scores of previously known relevant documents. For a query q, the set $D^+ = \{d|(q,d,t) \in H \wedge rel(q,d,t) > 0\}$ specifies the previously positive documents on which we calculate the tf-idf scores. The top k terms with the highest tf-idf scores are obtained for query expansion and appended to the original query. The expanded query is submitted to the retrieval system on the current document corpus D to produce the final ranking. Since this expansion relies solely on tf-idf scores from previous corpora, these scores can be calculated offline. Improving upon boosting, this allows to generalize from the historical relevance feedback to potentially deleted or newly created documents.

Keyqueries for Previously Relevant Documents. Given a retrieval model, the current document corpus D, and a set of historical relevance feedback H, we construct keyqueries [14,17,18] against the previously relevant documents. A query q_k is a keyquery for the set of target documents $D^+ = \{d|(q,d,t) \in H \wedge rel(q,d,t) > 0\}$ previously known relevant for a query q against the corpus $D^+ \cup D$ for the given retrieval model, iff (1) every $d \in D^+$ is in the top-k results, (2) q_k has more than l results, and (3) no subquery $q_k^{'} \subset q_k$ satisfies the above. The first criterion ensures the specificity, the second the generality, and the third the minimality of the keyquery, together trading off generalizability versus specificity. Traditional relevance feedback does not verify the position of feedback documents in the resulting rankings, whereas keyqueries uses them to remove overfitted or underfitted candidates. To generate candidates, we re-implemented a previous algorithm [12] in PyTerrier [25] which generates candidates from the top-10 RM3 terms. If multiple candidates are keyqueries, we use the one with the highest nDCG@10 on D^+. Finally, the keyquery is submitted against the corpus D, which may, or may not, contain documents that were previously relevant.

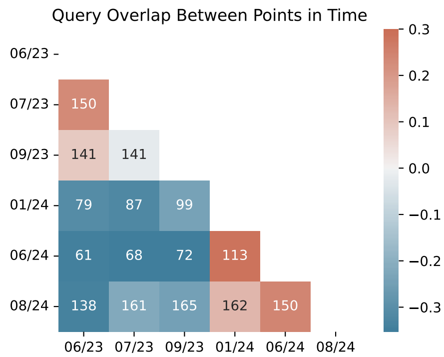

Fig. 1. Frequency of queries over time.

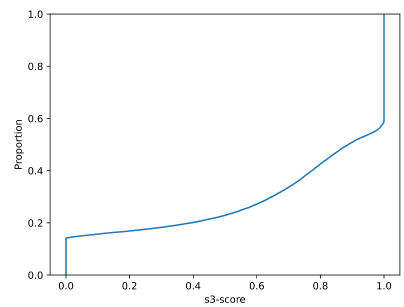

Fig. 2. S_3 Similarities of documents with overlapping URLs as eCDF plot.

4 Evaluation

We evaluate our counterfactual query rewriting in the LongEval scenario that comes with overlapping queries across six points in time between June 2022 and August 2023 [2,4,16]. We modify the LongEval datasets to focus on queries that re-occur across multiple timestamps to study the effects of evolving documents. We evaluate the retrieval effectiveness of all approaches and use an ablation study to investigate if they generalize beyond previously known relevant documents.

4.1 Experimental Setup

The LongEval test collection [16] samples documents, queries, and clicks from the French web search engine Qwant. For each timestamp, we remove queries that did not occur at least in one earlier timestamp, leaving 5 timestamps for evaluation between July 2022 and August 2023. Figure 1 overviews the overlapping queries for the timestamps. For instance, 138 queries from June 2023 re-occur in August 2024, forming the biggest time gap in our evaluation scenario.

We contrast five baselines with our three approaches. We use BM25 [27], BM25 with RM3 expansion (implemented in PyTerrier [25]), ColBERT [22], List-in-T5 [31], and monoT5 [26]. We use the default hyperparameters for all baselines (exporting ColBERT, List-in-T5, and monoT5 from TIRA/TIREx [13,15]). We also implement our three approaches in PyTerrier using BM25 as the underlying retrieval model. For boosting (BM25$_{Boost}$), we set $\lambda = 0.7$ and $\mu = 2$ based on previous experiments [21]. For relevance feedback (BM25$_{RF}$), $k = 10$ feedback terms are used as this is also the default for RM3 in PyTerrier. For keyqueries (BM25$_{keyquery}$), we use 10 feedback terms aiming at queries that retrieve the target tocuments to the top-10 while having more than 25 results.

4.2 Evolution of Documents in the LongEval Corpora over Time

The documents in the corpora may evolve via deletion, creation, or updates. The corpus comprises between one and 2.5 million documents, with a total 2.6 mil-

Table 2. Retrieval effectiveness of the five baselines and our three approaches as nDCG@10 with and without unjudged documents (nDCG@10$'$) on the LongEval timestamps. We report Bonferroni corrected significance against BM25 (†) and monoT5 (‡).

System	**nDCG@10**					**nDCG@10$'$**				
	07/22	09/22	01/23	06/23	08/23	07/22	09/22	01/23	06/23	08/23
BM25	.155	.184	.172	.175	.134	.471	.492‡	.516‡	.486‡	.379‡
$BM25_{RM3}$	.147‡	.181	.163	.174	.134	.478‡	.490‡	.524‡	.492‡	.388‡
ColBERT	.198	.207	.201	.184	.151	.402†	.409†	.420†	.408†	.315†
List-in-T5	.203	.204	.202	.198	.161	.401†	.413†	.425†	.413†	.317†
monoT5	.202	.219	.197	.202	.154	.405	.410†	.415†	.411†	.314†
$BM25_{Boost}$	**.355**†‡	.372†‡	**.287**†‡	**.364**†‡	**.271**†‡	.529‡	.546‡	.541‡	.540‡	.412‡
$BM25_{RF}$	.303†‡	.332†‡	.241†	.262†‡	.191†‡	.606†‡	.611†‡	**.590**†‡	.552†‡	**.426**†‡
$BM25_{keyquery}$	.350†‡	**.391**†‡	.233	.262	.185	**.642**	**.655**‡	.574†	**.554**	.422‡

lion created and 1.7 million deleted over time. We measure how the re-occuring documents changed by inspecting their pairwise similarities. We measured the similarity with the S_3 score [5] implemented in CopyCat [11] at default configuration (1 indicates identical, 0 no overlap) as this score aims to identify redundant documents in retrieval scenarios [5]. Figure 2 shows the S_3 similarities for all document pairs, indicating that 40 % do not change their content ($S_3 = 1.0$), whereas around 50 % have an S_3 similarity below 0.8 that indicate non-negligible changes (prior research used 0.82 as near-duplicate threshold on the Web [11]). Given that the LongEval corpora evolve only slightly, we include an ablation study that removes all overlap to analyze how approaches generalize.

4.3 Retrieval Effectiveness

We evaluate the effectiveness of our five baselines and our three approaches using nDCG@10. As the relevance labels of the LongEval corpus are derived from click logs, unjudged documents strongly impact the evaluation. In this scenario, it is recommendet to remove unjudged documents [28] which we report as nDCG@10$'$. Table 2 shows the results. ColBERT, List-in-T5, and monoT5 outperform the BM25 baseline in most cases, whreas BM25 with RM3 expansion does not substantially differ from BM25. Our three approaches substantially outperform all five baselines (nDCG$'$ is always higher). After removing the undesired impact of unjudged documents, both $BM25_{RF}$ and $BM25_{keyquery}$ outperform $BM25_{Boost}$, indicating that these approches generalize to newly created or modified documents. Keyqueries are the most effective approach in all cases, outperforming the best transformer by a large margin.

We conduct an ablation study to analyze if the improvements of $BM25_{RF}$ and $BM25_{keyquery}$ come from a generalization beyound previously known relevant documents. We remove all documents that occur in previous timestamps from the runs and relevance judgments and evaluate nDCG$'$. This way, all remaining

Table 3. Ablation study showing the nDCG@10$'$ improvement upon BM25 (± std-dev) on newly added documents that were never seen before, analyzing how approaches generalize. * marks Bonferroni corrected significance for students t-test.

System	07/23	09/23	01/24	06/24	08/24
BM25$_{Boost}$	+0.000$_{\pm.000}$	+0.000$_{\pm.000}$	+0.000$_{\pm.000}$	+0.000$_{\pm.000}$	+0.000$_{\pm.000}$
BM25$_{RF}$	-0.034$_{\pm.111}$	+0.000$_{\pm.135}$	+0.022$_{\pm.146}$	+0.012$_{\pm.081}$	+0.006$_{\pm.146}$
BM25$_{keyquery}$	-0.010$_{\pm.084}$	+0.012$_{\pm.153}$	+0.032$^*_{\pm.105}$	-0.001$_{\pm.065}$	+0.002$_{\pm.085}$

documents have never been seen before. Table 3 shows the results as improvement upon BM25 for our three approaches. As BM25$_{Boost}$ can not generalize to new documents, they never improve (improvement is always +0.0). However, both BM25$_{RF}$ and BM25$_{keyquery}$ generalize to unseen documents.

5 Conclusion and Future Work

We explored the capabilities of query rewriting approaches for recurring queries. We counterfactually assume that previously relevant documents are still available to use them as explicit relevance feedback.

The current analysis is subject to several limitations: currently it is restricted to a web search context, investigates a limited time frame, and relies on a single test collection. Moreover, comparing the approaches against systems that continuously learn from the evolving corpus could yield deeper insights. The dynamics of the collection and the ways users interact with the search system strongly influences the effectiveness of the proposed approaches. As collections evolve for various reasons, previously relevant documents may no longer be available for further use. Expiring licenses may prevent the system from using the licensed documents, even if they are not ranked, whereas outdated or shutdown websites may still be available. Similarly, the query types users issue and how the system processes them may be differently well suited. For instance, while short keyword queries, as used in the experiments, may benefit from direct expansion, natural language queries should not be expanded in the same way. While the results indicated improved effectiveness over the observed time frame, the approaches are essentially vulnerable to the Matthews effect, where older documents tend to accumulate exposure. Promoting older documents may create an increased entry barrier for new documents. Addressing such effects and biases over extended time frames is crucial for ensuring a secure and sustainably effective system.

Our approaches only need a few documents as feedback, and our results show that this already suffices to significantly outperform expensive transformer-based models. The ablation study suggests that the advanced approaches generalize beyond known query-document pairs, making them effective for new documents as well. Interesting directions for future work would be also to handle scenarios where the intent of a query might change, take into account how the documents evolve, or how relevance feedback observed for similar queries can be re-used.

References

1. Adar, E., Teevan, J., Dumais, S.T., Elsas, J.L.: The web changes everything: understanding the dynamics of web content. In: WSDM. pp. 282–291. ACM (2009)
2. Alkhalifa, R., Bilal, I.M., Borkakoty, H., Camacho-Collados, J., Deveaud, R., El-Ebshihy, A., Anke, L.E., Sáez, G.N.G., Galuscáková, P., Goeuriot, L., Kochkina, E., Liakata, M., Loureiro, D., Mulhem, P., Piroi, F., Popel, M., Servan, C., Madabushi, H.T., Zubiaga, A.: Extended overview of the CLEF-2023 longeval lab on longitudinal evaluation of model performance. In: CLEF (Working Notes). CEUR Workshop Proceedings, vol. 3497, pp. 2181–2203. CEUR-WS.org (2023)
3. Alkhalifa, R., Borkakoty, H., Deveaud, R., El-Ebshihy, A., Anke, L.E., Fink, T., Galuscáková, P., Sáez, G.G., Goeuriot, L., Iommi, D., Liakata, M., Madabushi, H.T., Medina-Alias, P., Mulhem, P., Piroi, F., Popel, M., Zubiaga, A.: Extended overview of the CLEF 2024 longeval lab on longitudinal evaluation of model performance. In: CLEF (Working Notes). CEUR Workshop Proceedings, vol. 3740, pp. 2267–2289. CEUR-WS.org (2024)
4. Alkhalifa, R., Borkakoty, H., Deveaud, R., El-Ebshihy, A., Anke, L.E., Fink, T., Galuscáková, P., Sáez, G.G., Goeuriot, L., Iommi, D., Liakata, M., Madabushi, H.T., Medina-Alias, P., Mulhem, P., Piroi, F., Popel, M., Zubiaga, A.: Extended overview of the CLEF 2024 longeval lab on longitudinal evaluation of model performance. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024. CEUR Workshop Proceedings, vol. 3740, pp. 2267–2289. CEUR-WS.org (2024)
5. Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: Herzog, O., Schek, H., Fuhr, N., Chowdhury, A., Teiken, W. (eds.) Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005. pp. 736–743. ACM (2005)
6. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Comput. Surv. **47**(2), 15:1–15:41 (2014)
7. Chuklin, A., Markov, I., de Rijke, M.: Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers (2015). https://doi.org/10.2200/S00654ED1V01Y201507ICR043
8. Dumais, S.T.: Putting searchers into search. In: SIGIR. pp. 1–2. ACM (2014)
9. Elsas, J.L., Dumais, S.T.: Leveraging temporal dynamics of document content in relevance ranking. In: WSDM. pp. 1–10. ACM (2010)
10. Fröbe, M., Akiki, C., Potthast, M., Hagen, M.: Noise-reduction for automatically transferred relevance judgments. In: Barrón-Cedeño, A., Da San Martino, G., Esposti, M.D., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022). Lecture Notes in Computer Science, vol. 13390, pp. 48–61. Springer, Berlin Heidelberg New York (Sep 2022). https://doi.org/10.1007/978-3-031-13643-6_4
11. Fröbe, M., Bevendorff, J., Gienapp, L., Völske, M., Stein, B., Potthast, M., Hagen, M.: CopyCat: Near-Duplicates within and between the ClueWeb and the Common Crawl. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) 44th International ACM Conference on Research and Development in

Information Retrieval (SIGIR 2021). pp. 2398–2404. ACM (Jul 2021). https://doi.org/10.1145/3404835.3463246
12. Fröbe, M., Günther, S., Bondarenko, A., Huck, J., Hagen, M.: Using Keyqueries to Reduce Misinformation in Health-Related Search Results. In: 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2022). CEUR Workshop Proceedings, CEUR-WS.org (Apr 2022)
13. Fröbe, M., Reimer, J.H., MacAvaney, S., Deckers, N., Reich, S., Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: The Information Retrieval Experiment Platform. In: Chen, H.H., Duh, W., Huang, H.H., Kato, M., Mothe, J., Poblete, B. (eds.) 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023). pp. 2826–2836. ACM (Jul 2023). https://doi.org/10.1145/3539618.3591888
14. Fröbe, M., Schmidt, E.O., Hagen, M.: Efficient query obfuscation with keyqueries. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. pp. 154–161 (2021)
15. Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., Hagen, M., Stein, B., Potthast, M.: Continuous Integration for Reproducible Shared Tasks with TIRA.io. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023). pp. 236–241. Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Apr 2023). https://doi.org/10.1007/978-3-031-28241-6_20
16. Galuscáková, P., Deveaud, R., Sáez, G.G., Mulhem, P., Goeuriot, L., Piroi, F., Popel, M.: Longeval-retrieval: French-english dynamic test collection for continuous web search evaluation. In: Chen, H., Duh, W.E., Huang, H., Kato, M.P., Mothe, J., Poblete, B. (eds.) Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023. pp. 3086–3094. ACM (2023). https://doi.org/10.1145/3539618.3591921
17. Gollub, T., Hagen, M., Michel, M., Stein, B.: From Keywords to Keyqueries: Content Descriptors for the Web. In: Gurrin, C., Jones, G., Kelly, D., Kruschwitz, U., de Rijke, M., Sakai, T., Sheridan, P. (eds.) 36th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2013). pp. 981–984. ACM (Jul 2013). https://doi.org/10.1145/2484028.2484181
18. Hagen, M., Beyer, A., Gollub, T., Komlossy, K., Stein, B.: Supporting Scholarly Search with Keyqueries. In: Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016). Lecture Notes in Computer Science, vol. 9626, pp. 507–520. Springer, Berlin Heidelberg New York (Mar 2016). https://doi.org/10.1007/978-3-319-30671-1_37
19. Kanhabua, N., Blanco, R., Nørvåg, K.: Temporal information retrieval. Found. Trends Inf. Retr. **9**(2), 91–208 (2015). https://doi.org/10.1561/1500000043
20. Keller, J., Breuer, T., Schaer, P.: Evaluation of temporal change in IR test collections. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024. pp. 3–13. ACM (2024). https://doi.org/10.1145/3664190.3672530
21. Keller, J., Breuer, T., Schaer, P.: Leveraging prior relevance signals in web search. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF

2024), Grenoble, France, 9-12 September, 2024. CEUR Workshop Proceedings, vol. 3740, pp. 2396–2406. CEUR-WS.org (2024)
22. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. pp. 39–48. ACM (2020). https://doi.org/10.1145/3397271.3401075
23. Lin, J., Nogueira, R.F., Yates, A.: Pretrained Transformers for Text Ranking: BERT and Beyond. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2021). https://doi.org/10.2200/S01123ED1V01Y202108HLT053
24. Liu, T.: Learning to Rank for Information Retrieval. Springer (2011). https://doi.org/10.1007/978-3-642-14267-3
25. Macdonald, C., Tonellotto, N.: Declarative experimentation in information retrieval using pyterrier. In: Balog, K., Setty, V., Lioma, C., Liu, Y., Zhang, M., Berberich, K. (eds.) ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020. pp. 161–168. ACM (2020). https://doi.org/10.1145/3409256.3409829
26. Nogueira, R.F., Jiang, Z., Pradeep, R., Lin, J.: Document ranking with a pretrained sequence-to-sequence model. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. Findings of ACL, vol. EMNLP 2020, pp. 708–718. Association for Computational Linguistics (2020). https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.63
27. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. (Jan 1994)
28. Sakai, T.: Alternatives to bpref. In: Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (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. pp. 71–78. ACM (2007). https://doi.org/10.1145/1277741.1277756
29. Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum **33**(1), 6–12 (1999)
30. Soboroff, I.: Dynamic test collections: measuring search effectiveness on the live web. In: SIGIR. pp. 276–283. ACM (2006)
31. Tamber, M.S., Pradeep, R., Lin, J.: Scaling down, litting up: Efficient zero-shot listwise reranking with seq2seq encoder-decoder models. CoRR **abs/2312.16098** (2023). https://doi.org/10.48550/ARXIV.2312.16098
32. Tikhonov, A., Bogatyy, I., Burangulov, P., Ostroumova, L., Koshelev, V., Gusev, G.: Studying page life patterns in dynamical web. In: SIGIR. pp. 905–908. ACM (2013)
33. Tyler, S.K., Teevan, J.: Large scale query log analysis of re-finding. In: WSDM. pp. 191–200. ACM (2010)