Leveraging Prior Relevance Signals in Web Search

Notebook for the LongEval Lab at CLEF 2024

Jüri Keller[1,*], Timo Breuer[1] and Philipp Schaer[1]

[1]TH Köln (University of Applied Sciences), Claudiusstr. 1, Cologne, 50678, Germany

Abstract

This work reports our participation in the retrieval task of the second LongEval lab iteration at CLEF 2024. As part of this year's contribution, we analyze to which extent prior relevance signals on the document level and term level can be used to improve the retrieval effectiveness. In order to exploit these kinds of signals, we fetch corresponding document identifiers pointing to the same document in the different dataset slices of all timestamps. Based on several heuristics, we submit and evaluate a total of five systems that either follow our previous year's methodology or that combine baseline rankings with prior relevance signals. Our evaluations provide insights to which extent these signals can be used but let us also conclude with several recommendations for future work. Most notably, we envision a companion resource that ties together all slices of the dataset by unified document identifiers to have a better understanding of more rigorous data splits and to avoid potential data leakage that might affect the evaluation of (deep) learning-based systems.

Keywords

Web Search, Longitudinal Evaluation, Continuous Evaluation, Replicability

1. Introduction

The overall goal of Information Retrieval (IR) systems is to assist in finding relevant information for given information needs. These systems are designed to cope with the ongoing flood of information. Since the information landscape is continuously changing, the foundational data basis is ever-evolving, exposing systems to always new and updated information. While these changes directly influence the retrieval effectiveness, they are rarely considered during evaluation. Web search is an especially dynamic search scenario since websites and queries change quickly. This evolving search setting is brought to a test in the LongEval shared task that has the main goal of evaluating how systems cope with changes over time. Therefore, retrieval systems are evaluated on progressing snapshots of a test collection.

While the systems are exposed to changing data, users expect consistent, good effectiveness. To provide and maintain this, it is essential to quantify the effectiveness regularly. However, since the foundational data source, including documents, topics, and qrels is evolving, it is an open question for how long evaluations remain valid or can be generalized. While the LongEval shared task provides a one-of-a-kind test bed for the described endeavors, previous submissions mainly investigate how the changes in the test collection affect systems that remain static across time. The systems submitted last year did not use the temporal aspects of the test collection and instead treated the snapshots as independent search tasks. Exploiting the historical data as prior relevance signals in ranking systems is an important step in learning about the temporal connection between the snapshots and what differentiates longitudinal evaluations from cross test collection evaluations.

In our contribution, we aim to investigate different approaches that exploit prior relevance signals on different levels in intuitive and explainable ways. Our goal is to provide initial approaches to adapt the system to the evolving retrieval setup by using past snapshots, including ranked documents and qrels. By that, the system itself becomes an evolving component in this evaluation, like the documents,

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

jueri.keller@th-koeln.de (J. Keller); timo.breuer@th-koeln.de (T. Breuer); philipp.schaer@th-koeln.de (P. Schaer)

0000-0002-9392-8646 (J. Keller); 0000-0002-1765-2449 (T. Breuer); 0000-0002-8817-4632 (P. Schaer)

topics, and qrels. While these methods are mainly limited to topics with a known history, i.g. queries that were already issued previously, the Average Retrieval Performance (ARP) clearly improves.

In short, the contributions of this work are:

- Three experimental systems that exploit information of the past snapshots (past rankings and qrels) as prior relevance signals and two baselines are submitted to the LongEval shared task.
- An extensive evaluation of the submitted systems regarding their behavior and effectiveness across time through reproducibility and replicability measures.
- We outline and discuss in detail three directions of future work, namely the need for intellectual relevance labels, possible data leakage, and a companion resource that makes the snapshots more accessible.

To facilitate reproducibility, we make the code publicly available on GitHub.[1]

2. LongEval 2024 Dataset

The LongEval dataset is a retrieval test collection introduced in the LongEval[2] CLEF lab in 2023 [1]. It contains multiple sub-collections that resemble snapshots of different points in time of the French, privacy-focused search engine Qwant.[3] The goal of the LongEval lab are longitudinal evaluations in IR. The original test collection that was used in last year's iteration of the lab contained three sub-collections (WT, ST, and LT). In this year's iteration, three additional sub-collections are released and added to the test collection. By that, the test collection grows over time, covering longer time spans. Since no standardized naming of the different sub-collections exists, in this work, we name them by simply iterating over all sub-collections in chronological order, starting with t_0 for last year's WT sub-collection. The new snapshots that are added in this iteration are therefore:

- t_3: used as train split captured in January 2023.
- t_4: used as a test split, capturing data from June 2023. The snapshot has a gap of five months to t_3 and the runs are submitted for `lag6`.
- t_5: also used as a test split, capturing data from August 2023. The snapshot has a gap of seven months to t_3 and the runs are submitted for `lag8`.

The test collection is originally available in French, but an English automatic translation is additionally provided, on which this work mainly relies on. While we generally consider the translations to be good, some mismatches and near duplicates occur. This is also reflected in a higher effectiveness for runs using the French version [2].

The test collection statistics of the newly added snapshots are summarized in Table 1. They contain between 1.7 and 2.5 million documents and 404 to 1518 topics. Compared to last year's snapshots, it contains more documents but fewer topics on average. The qrels classify the documents as not relevant, relevant, and highly relevant. The number of qrels is balanced almost evenly. Roughly the same amount of positive and negative labels are present while regarding the positive labels, only one-third are highly relevant. The distribution across topics is highly skewed, as visualized in Figure 1.

The dataset's different snapshots overlap, strengthening the connection between them. Similar to the snapshots investigated last year, some topics appear in multiple or even all snapshots. Likewise, the document collection holds websites present in all snapshots. Additionally, many more topics and documents are added and removed over time. Besides creating and updating operations on the collection, some documents that are present in multiple snapshots also change over time. For components that were already present in a previous snapshot, the history of this component can be constructed by relating the current version to its previous ones. Tracking these different changes, even on an abstract

[1] https://github.com/irgroup/CLEF2024-LongEval-CIR
[2] https://clef-longeval.github.io/
[3] https://www.qwant.com/

Table 1
Test collection statistics of the three newly added snapshots. Besides the count of documents, topics, and qrels, the total number of qrels distributed across the three relevance labels (not_rel, rel, high_rel) and the summarizing statistics (mean, min, max) of all qrels per topic are displayed. The qrels are filtered to the ones used in the shared task.

	Documents	Topics	Qrels	not_rel	rel	high_rel	mean	min	max
t_3	2,049,729	599	9,785	5423	2891	1471	16.3	6	56
t_4	1,790,028	404	5,835	3146	1699	990	14.4	1	50
t_5	2,531,614	1518	24,861	14602	7078	3181	16.4	1	59

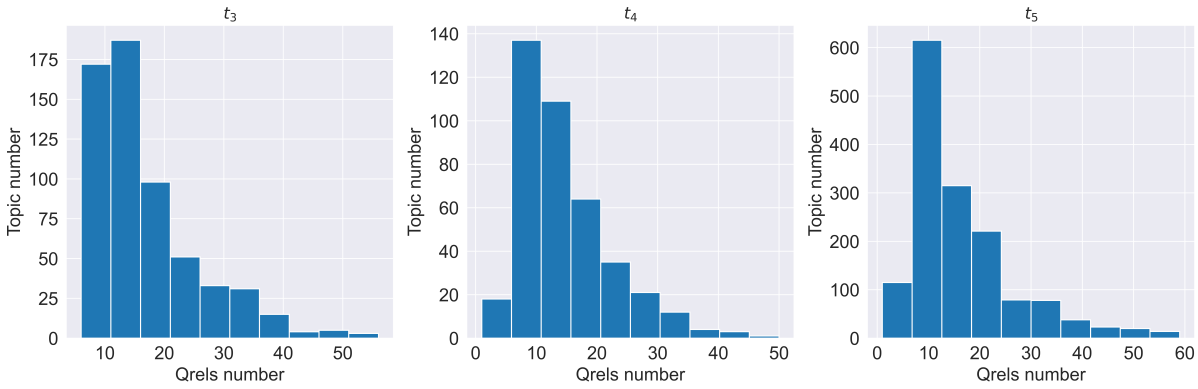

Figure 1: Histograms of the qrels distribution across topics.

level, remains challenging since each snapshot uses different identifiers. Therefore, no direct connection between topics, documents, or qrels is possible. A detour must be made via additional information to connect the components across time. We chose to link the documents by the provided URLs and the topics by exact string matches of the French query string. Linking the components by exact matches on the strings is straightforward but not reliable because this will not identify (near) duplicates for which the URL or query slightly changed. Even though the English version of the dataset is used in general, for the topic linking, the French queries are used to avoid differences occurring by the translation, especially with regards to the larger time gap between the snapshots from this year and the dataset from last years iteration in which the translation system might have changed. Linking between different temporal versions of the same topics and documents allows us to exploit the history of the test collection for the ranking.

The snapshots appear to be rather loosely connected since the overlap in topics is small. While in total 404 to 1518 topics are available, only 98 topics are present in all three new snapshots. In comparison, 124 topics could be found with the same matching method that are available in all snapshots from last year [3]. Presumably, even fewer topics are present in the snapshots of both iterations combined. We define the overlap of topics between sub-collections as core queries. Limiting the topic set to these core queries reduces the changes in the test collection to the document and relevance components.

3. Approaches and Implementations

Based on the data analysis of corresponding document identifiers described in the previous section, we can identify the same document in different snapshots of the datasets, i.e., the test collection at different timestamps. The following approaches are based on the heuristic that prior snapshots or rankings retrieved from earlier snapshots, bear some kind of relevance information that can be exploited for the ranking of the current timestamp. Generally, we derive relevance signals for the ranking at timestamp t_n from the sub-collection or a ranking at one or more earlier timestamps t_{n-1}. This history of snapshots can also include the sub-collections from last year (t_0, t_1, t_2). Compared to an earlier

snapshot, a current ranking at t_n can contain new documents that are added after t_{n-1}, documents that were also present at t_{n-1}, and documents that were present at t_{n-1} but now have different content.

An overview of all runs and their general methodology is given in Table 2.

Table 2
Overview of our submitted runs and the underlying methodology

Run	Method	Modality of Prior Relevance Signal
`CIR_BM25`	Baseline ranking based on BM25 (cf. 3.1)	-
`CIR_BM25+monoT5`	Reranking based on pre-trained monoT5 (cf. 3.1)	-
`CIR_BM25+time_boost`	Fusion based on weighting prior documents (cf. 3.2)	Documents in prior ranking
`CIR_BM25+qrel_boost`	Boosting based on weighting prior, judged documents (cf. 3.2)	Judged prior documents
`CIR_BM25+RF`	Query expansion with terms of prior, relevant documents (cf. 3.3)	Terms from prior relevant documents

3.1. "Off-The-Shelve" Baseline and Transformer-Based Reranking

Given the queries of the timestamp t_n, we run BM25 [4] on the corresponding index. BM25 is a well-established lexical retrieval method that is still competitive for many applications where training data is sparse. We run it with default parameters as implemented in the PyTerrier toolkit [5]. The corresponding run submission is entitled `CIR_BM25`.

The underlying retrieval method of the run submission `CIR_BM25+monoT5` reranks the BM25 baseline run `CIR_BM25` with monoT5 [6]. We make use of the monoT5 version that was pre-trained on MS MARCO passages.[4] To use T5 as the reranking model, the documents are truncated after 512 sub-word tokens. This system was submitted as an additional baseline to BM25 since it performed well on last year's iteration of the LongEval Lab [3]. The same implementation as last year's is reused for this run.

3.2. Direct Boost by Prior Ranking and Relevance Information

The run submission `BM25+time_boost` is based on the hypothesis that given a document appears in both the ranking at the current timestamp t_n and an earlier timestamp t_{n-1}, it is relevant — at least to the degree that it was not removed from the index and is still considered as a potential document in the ranking.

For this approach, besides the baseline BM25 ranking at timestamp t_n, we also determine rankings at the earlier timestamp t_{n-1} with the same set of queries but with the corresponding index at timestamp t_{n-1}. Both the scores and the document identifiers are normalized. This procedure results in two runs for the same set of topics that have possibly different documents ranked. Then, the documents that are also ranked at t_{n-1} are boosted by

$$\rho_{q,d}(\lambda) = \begin{cases} \lambda^2 & \text{if } d \in r_{q,t_{n-1}} \\ (1-\lambda)^2 & \text{otherwise} \end{cases} \tag{1}$$

where $r_{q,t_{n-1}}$ is the ranking r_q corresponding to the query q at timestamp t_{n-1} and λ is a free parameter. Figure 2 illustrates the weighting score $\rho_{q,d}$ based on different λ values. The basic intuition is that we can control the ratio between weights that are assigned to documents that are either contained in a ranking at an earlier timestamp or those documents that are not.

[4] https://huggingface.co/castorini/monot5-base-msmarco

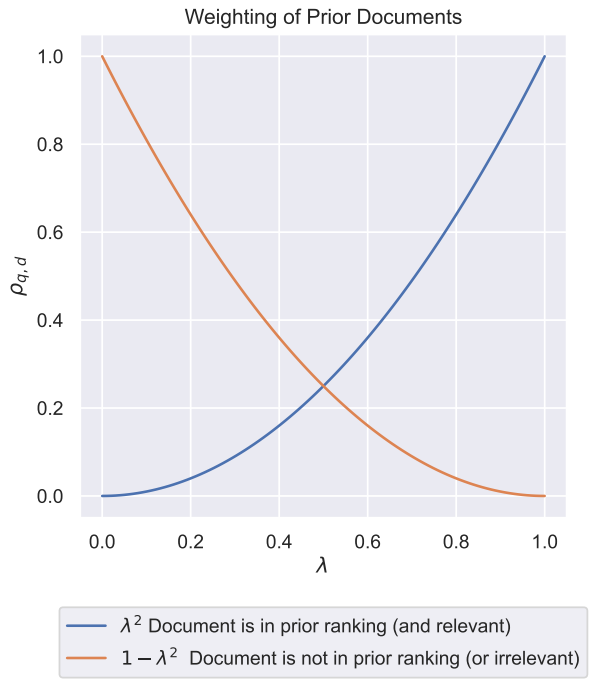

Figure 2: Weighting scheme for (relevant) documents of prior rankings

Suppose we assign a λ value larger than 0.5. In that case, we would emphasize prior, i.e., older documents in the ranking, which is a rather conservative ranking policy in a real-world setting. On the other hand, λ values lower than 0.5 yield a higher weight to new documents that were not contained in a prior ranking, corresponding to a more experimental, risky ranking policy in a real-world setting. If $\lambda = 0.5$ the ranking remains the same, and the effect of the weighting is negligible.

For the submitted run, $\lambda = 0.503$ was selected based on a grid search on the train set at t_3. Notably, this run only relies on the fact that documents were ranked earlier, but neglects any direct relevance information.

The run submission `BM25+qrel_boost` extends this rather weak heuristic by taking the relevance label directly into account and considering not only the t_{n-1} snapshot but further prior snapshots. In this sense, this run follows a similar ranking policy (*conservative vs. experimental*) but with a focus on relevance information obtained from prior labels. Independent of any prior ranking, all documents d ranked for a query q at t_n that were judged at a previous snapshot, e.g. t_{n-1} are boosted by:

$$\rho_{q,d}(\lambda) = \begin{cases} \lambda^2 & \text{if } qrel_{q,d} = 1 \\ \lambda^2\mu & \text{if } qrel_{q,d} > 1 \\ (1-\lambda)^2 & \text{if } qrel_{q,d} = 0 \end{cases}. \tag{2}$$

This extends the boost described in Eq. 1 to additionally account for highly relevant documents. The additional free parameter μ can be used to assign a different boost to documents with higher relevance labels. The boosting can be repeated for further points in time by using more distanced qrels e.g. t_{n-2}. This leads to increasing scores for documents that are labeled relevant at multiple points in time, which is in line with our initial assumption. For the submitted run, $\lambda = 0.7$ was chosen, and all available previous snapshots (t_3, t_2, t_1, t_0) as history. Due to a bug that was only found after submission, instead of highly relevant documents, all relevant documents were boosted again by $\lambda^2\mu$ with $\mu = 2$.

3.3. Query Expansion Based on Prior Relevance Feedback

In addition to the run submissions described above, we propose one additional approach that aligns with the general idea of exploiting prior relevance signals. `BM25+RF` follows a query expansion method,

making use of the relevance feedback provided by prior documents, i.e., those documents with a relevant label at earlier timestamps.

The relevant documents of all earlier timestamps are normalized and deduplicated based on the URLs. Then, the vocabulary of all remaining relevant documents corresponding to a query is unified. After stopwords are removed, the top ten expansion terms are extracted based on the TF-IDF scores and are used to expand the original query string. Hypothetically, this approach allows us to profit from the prior relevance information independent of a particular ranking or the labels for *query-document* pairs.

Since not all topics have a history with relevant documents available, the rankings for unknown topics are reranked by pseudo-relevance feedback based on RM3 [7]. Similarly, the query is expanded by ten terms extracted from the top three documents ranked by BM25.

4. Results

The ARP measured by P@10, bpref [8], nDCG [9], and also Mean Reciprocal Rank (MRR) [10] are reported in Table 3. Since the test collection has shallow pools, we report bpref instead of MAP [11]. Additionally, the MRR is reported since it relies only on the first relevant result ranked and fits the web search use case well. Following on from last year's evaluation [3], we report the replicability measures Delta Relative Improvement (ΔRI) and Effect Ratio (ER) [12] to quantify how the effectiveness changes across time. Since the snapshots are obtained from the same search engine, they are naturally related, but due to the changes over time, they are still very different. Thus, a comparison is not straightforward, and an advanced comparison strategy is needed to factor in the changed recall base [13, 14]. The ΔRI basically describes the changes in effectiveness compared to t_3 and the ER describes how well the effect measured at t_3 is recovered. Both measures rely on BM25 as a pivot system in a way that not the direct scores at two points in time are compared but the deltas to the pivot system. Since both the experimental and the pivot systems are exposed to the same snapshot, the impact of the evaluation environmental should be reduced, and the results should be more comparable. Additionally, Kendall's Tau [15] and RBO [16] are reported in Table 4 to directly compare the rankings at later points in time to the ranking at t_3 independently of any effectiveness.

The retrieval effectiveness changes over time regarding all systems and measures. For almost all systems, it is decreasing compared to t_3. It can be observed that the newer the snapshot, the lower the measured effectiveness. The ΔRI shows more minor changes in effectiveness. In contrast to the ARP trends, it indicates a slightly increasing effectiveness as indicated by the negative ΔRI values.

All runs per snapshot are tested for significantce compared to the `BM25` baseline using paired t-tests with $\alpha = 0.05$ and Bonferroni correction. Most runs show significant differences except `BM25+time_boost` at t_4 and t_5 for all measures or `BM25+monoT5` at t_4 measured by bpref or at t_3 by MRR.

The ranking of systems appears to be relatively consistent across time. The `BM25` baseline is outperformed by `BM25+monoT5`, but only by a little. The system `BM25+time_boost` initially performs the worst but is on par with `BM25` for the later snapshots. These rankings are so similar that no significant differences can be measured. The two systems, `BM25+qrel_boost` and `BM25+RF` systems, often outperform `BM25+monoT5`. In particular, the `BM25+qrel_boost` system performs substantially better.

The ranking similarity between two points in time measured by Kendall's Tau and RBO appears to be generally low. Especially the `BM25+RF` system shows a low similarity regarding Kendall's Tau, maybe because of the query rewriting that depends on the history. If the effectiveness changes are compared to the ranking similarity as displayed in Table 4, even small changes in effectiveness can lead to vastly different rankings. For example, the system `BM25+time_boost` evaluated at t_3 and t_4 show only small improvements on all measures but also among the smallest similarities in the rankings for Kendall's Tau and RBO. Interestingly, the system `BM25+monoT5` shows similar effectiveness to the system `BM25+qrel_boost` considering nDCG but different ranking similarities measured by RBO. `BM25` has the most similar rankings, followed by `BM25+monoT5` and `BM25+qrel_boost`.

The systems that rely on previous relevance signals, especially `BM25+qrel_boost` but also `BM25+RF` most often show more variance over time in the ΔRI and ER values, compared to the other systems. For

Table 3
Experimental results of the different retrieval systems across all snapshots. Statistically significant differences in the ARP from BM25 at the same point in time are marked with an asterisk (*). For the replicability measure BM25 is used as the pivot system, and the changes are measured in comparison to t_3. For these rows, the ideal values are indicated.

		P@10			bpref			nDCG			MRR		
System	t	ARP	ΔRI	ER	ARP	ΔRI	ER	ARP	ΔRI	ER	ARP	ΔRI	ER
BM25	t_3	0.1624	-	-	0.4373	-	-	0.3638	-	-	0.3660	-	-
BM25	t_4	0.1370	-	-	0.3572	-	-	0.2817	-	-	0.3162	-	-
BM25	t_5	0.1076	-	-	0.2791	-	-	0.2106	-	-	0.3116	-	-
+monoT5	t_3	0.1776*	0	1	0.4571*	0	1	0.3839*	0	1	0.3929	0	1
+monoT5	t_4	0.1591*	-0.0675	1.4513	0.3719	0.0043	0.7401	0.3081*	-0.0384	1.3122	0.3847*	-0.1436	2.5532
+monoT5	t_5	0.1246*	-0.0640	1.1155	0.2862*	0.0198	0.3591	0.2303*	-0.0385	0.9824	0.3551*	-0.0662	1.6190
+time_boost	t_3	0.1007*	0	1	0.4178*	0	1	0.2758*	0	1	0.2774*	0	1
+time_boost	t_4	0.1380	-0.3873	-0.0161	0.3583	-0.0477	-0.0562	0.2850	-0.2534	-0.0369	0.3308	-0.2882	-0.1651
+time_boost	t_5	0.1065	-0.3702	0.0171	0.2798	-0.0472	-0.0372	0.2116	-0.2469	-0.0120	0.3210*	-0.2722	-0.1065
+qrel_boost	t_3	0.1870*	0	1	0.4515*	0	1	0.3910*	0	1	0.4394*	0	1
+qrel_boost	t_4	0.1975*	-0.2906	2.4630	0.3815*	-0.0354	1.7037	0.3536*	-0.1805	2.6442*	0.4922*	-0.3562	2.3981
+qrel_boost	t_5	0.1349*	-0.1021	1.1097	0.2881*	0.0005	0.6289	0.2419*	-0.0740	1.1515*	0.4063*	-0.1033	1.2897
+RF	t_3	0.1758*	0	1	0.4819*	0	1	0.3955*	0	1	0.3997*	0	1
+RF	t_4	0.1608*	-0.0915	1.7806	0.3958*	-0.0059	0.8642	0.3197*	-0.0476	1.1966	0.3821*	-0.1164	1.9556
+RF	t_5	0.1230*	-0.0606	1.1504	0.2990*	0.0306	0.4467	0.2293*	-0.0013	0.5878	0.3373*	0.0097	0.7620

Table 4
The similarity of the document rankings from different systems according to Kendall's Tau and RBO. Since for some topics less than 1000 documents were retrieved, only the top 100 documents are considered for this comparison.

System	t	τ@100	RBO@100
BM25	t_3	1	1
BM25	t_4	0.0132	0.4203
BM25	t_5	0.0152	0.3961
+monoT5	t_3	1	1
+monoT5	t_4	0.0126	0.4211
+monoT5	t_5	0.0122	0.3960
+time_boost	t_3	1	1
+time_boost	t_4	-0.0167	0.2086
+time_boost	t_5	0.0051	0.1847
+qrel_boost	t_3	1	1
+qrel_boost	t_4	0.0234	0.3223
+qrel_boost	t_5	0.0150	0.3275
+RF	t_3	1	1
+RF	t_4	-0.0006	0.2266
+RF	t_5	0.0003	0.2336

example, the system `BM25+qrel_boost` indicates an increased effectiveness (nDCG ΔRI of -0.1805) at t_4 but only a smaller increase (nDCG ΔRI of -0.0740) at t_5. Simultaneously, the nDCG ΔRI for `BM25+monoT5` only differs by 0.0001 points.

5. Discussion and Future Work

Based on our experimental results, we basically see three directions for future work, namely:

1. the addition of deep relevance judgments pools,
2. a more in-depth analysis of potential data leakage,

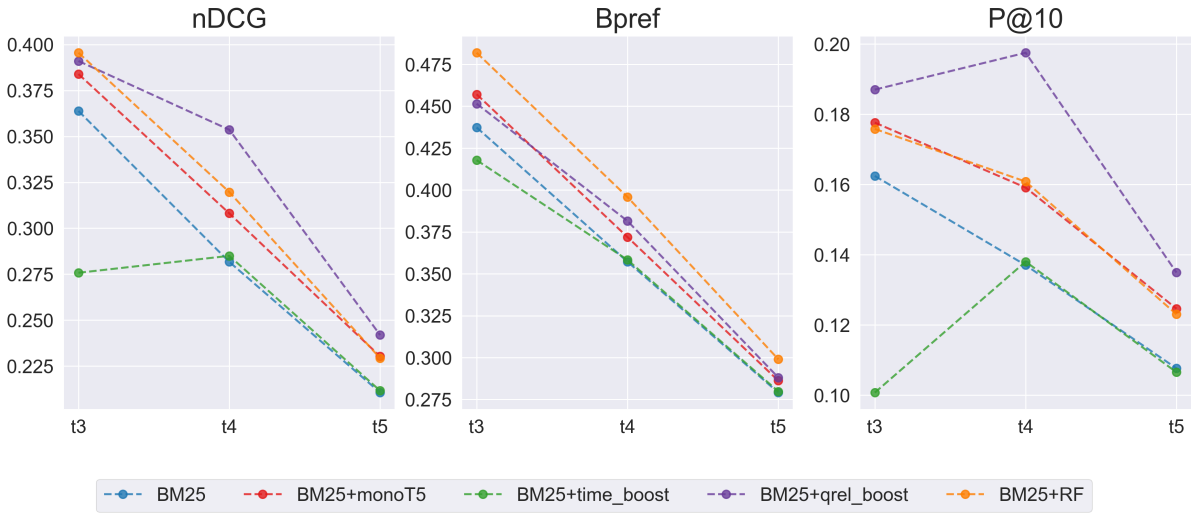

Figure 3: Effectiveness of the submitted systems across all snapshots measured by nDCG (left), Bpref (center), and P@10 (right).

3. the curation of a novel resource that ties together the different snapshots of the dataset (by identifying corresponding documents identifiers that point to the same document).

In general, we can see a performance drop for all systems and all measures over the three timestamps. None of our submitted systems specifically addressed the goal to identify new potentially relevant documents in their methodology. Instead, the submissions rather followed a retrospective approach where the relevance information of prior rankings was exploited.

As expected the baseline method BM25 is outperformed when reranking it in the second stage with monoT5, which is in line with our observations from the lab's first iteration. Even though the performance drops over time for the BM25 baseline and the monoT5 reranking alike, the latter outperforms BM25 wrt. every measure at each timestamp.

The approach underlying `CIR_BM25+time_boost` is less effective at timestamp t_3 but on par at t_4 and t_5 with BM25. This lets us conclude that it is not advisable to simply emphasize prior documents for the sake of better retrieval effectiveness without considering any other relevance information.

However, we have to point out that at timestamps t_4 and t_5 the retrieval effectiveness does not deteriorate, which is an important insight especially when considering that the rank correlation between the rankings at different timestamps is low. That means the rankings are quite different but have the same retrieval effectiveness in the end. From a user perspective, this could imply tremendous differences that are simply not reflected by the effectiveness scores.

We assume that this can be explained by the dataset currently only having shallow relevance judgment pools. Suppose that more documents would have been assigned with a relevance label. In that case, we would probably see more diverse outcomes here. In the future, we would strongly recommend to **curate a companion dataset or resource with intellectual relevance labels** that complements the dataset with deep relevance judgment pools.

A good example of an equivalent is the combination of the TripClick [17] and TripJudge [18] datasets, that cover both click-based and annotator-based relevance labels. Of course, the use of large language models for generating relevance labels could be a viable option if costs for human annotators are too high [19].

However, when comparing the monoT5 results to the other run submissions, we also see that specifically `CIR_BM25+qrel_boost` and `CIR_BM25+RF` are on par or even outperform monoT5, which opens up several interesting points for discussion. These systems are the best-performing ones among our submissions.

Regarding `CIR_BM25+RF`, we can conclude that query expansion with terms from prior relevant documents is a viable option to make use of available relevance signals. While our experimental results

show rather minor improvements over the monoT5 reranking, this approach reliably yields better scores. In the future, other language models for generating the expansion terms should be considered and evaluated.

`CIR_BM25+qrel_boost` opens up several other interesting perspectives for a more in-depth analysis. The approach of this run submission specifically boosts those documents that were judged as relevant earlier. This general approach was rather successful in the OpenSearch TREC and its CLEF predecessor LL4IR [20, 21]. Instead of (derived) relevance judgments, the original implementation used click data to approximate relevance. Nevertheless, it has to be considered that these documents and judgments could have served as training data for deep learning retrieval methods. This is critical as **data leakage might be an issue when the datasets are used as provided** and we recommend that future work should analyze this circumstance in more detail [22].

We have the concern that this circumstance was not that apparent as the documents with the same URL and the same content have different identifiers in the different versions of the datasets at different timestamps. That finally leads us to our third and final direction for future work, which would be a **companion resource to the LongEval datasets that ties together all of the six dataset's snapshots by unifying the document identifiers** and making potential document overlaps, which may cause data leakage in a learning-based scenario, explicit.

A higher variance in $\Delta\mathrm{RI}$ and ER for the systems that rely on the relevance signals of previous snapshots could be observed. This can be interpreted as validation of the comparison strategy. The runs directly utilize the relevance signals in more or less immediate. However, for both test points in time (t_4 and t_5) the same history of snapshots (t_0 to t_3) is used. This means that the relevance signals for t_4 are more up-to-date compared to t_5 and a higher improvement can be expected. In addition, the similarity of the rankings further supports this interpretation. Both, `BM25+qrel_boost` and `BM25+RF`, indicate especially consistent similar rankings over time, caused by boosting the same documents for both runs.

6. Conclusion

As part of this year's contributions to the LongEval lab, we conclude that leveraging prior relevance signals from existing logs for the sake of better retrieval effectiveness is a promising direction. Furthermore — in line with our earlier observations from the first iteration of the lab — we confirm that the retrieval effectiveness changes with different snapshots of the dataset and addressing the need for making retrieval systems more robust, reliable, and predictable is an exciting direction for future work that should receive more attention.

In particular, we applied some heuristics to make use of the available relevance information that can be obtained from prior snapshots of the dataset to improve a recent ranking. To do so, it was a requirement to identify the same document across the different snapshots. Our preliminary analysis revealed that there are indeed the same documents in different snapshots and using these (relevance) signals helps to improve retrieval effectiveness. Based on these outcomes, we envision the contribution of a companion resource that ties together the dataset's snapshots to identify documents and duplicates that would allow a more in-depth analysis about the document's relevance over time but also make a more rigorous experimental setup possible.

Acknowledgments

We would like to express our gratitude to the LongEval Shared Task organizers for their invaluable efforts in constructing the LongEval dataset. Their dedication and hard work have provided an essential foundation for our research. We also gratefully acknowledge the support of the German Research Foundation (DFG) through project grant No. 407518790.

References

[1] 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, in: SIGIR, ACM, 2023, pp. 3086–3094.

[2] R. Alkhalifa, I. M. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. E. Anke, G. G. Sáez, P. Galuscáková, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, P. Mulhem, F. Piroi, M. Popel, C. Servan, H. T. Madabushi, A. Zubiaga, Overview of the CLEF-2023 longeval lab on longitudinal evaluation of model performance, in: CLEF, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 440–458.

[3] J. Keller, T. Breuer, P. Schaer, Evaluating temporal persistence using replicability measures, in: CLEF (Working Notes), volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2441–2457.

[4] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[5] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using pyterrier, in: ICTIR, ACM, 2020, pp. 161–168.

[6] R. Pradeep, R. F. Nogueira, J. Lin, The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, CoRR abs/2101.05667 (2021).

[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: TREC, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004.

[8] C. Buckley, E. M. Voorhees, Retrieval evaluation with incomplete information, in: M. Sanderson, K. Järvelin, J. Allan, P. Bruza (Eds.), SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004, ACM, 2004, pp. 25–32. URL: https://doi.org/10.1145/1008992.1009000. doi:10.1145/1008992.1009000.

[9] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: http://doi.acm.org/10.1145/582415.582418. doi:10.1145/582415.582418.

[10] E. M. Voorhees, The TREC-8 question answering track report, in: E. M. Voorhees, D. K. Harman (Eds.), Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999, volume 500-246 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1999. URL: http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf.

[11] E. M. Voorhees, C. Buckley, Retrieval system evaluation. in harman, in: E. Voorhees, D. K. Harman, National Institute of Standards and Technology (U.S.) (Eds.), TREC: Experiment and Evaluation in Information Retrieval, Digital Libraries and Electronic Publishing, MIT Press, Cambridge, Mass, 2005.

[12] T. Breuer, N. Ferro, N. Fuhr, M. Maistro, T. Sakai, P. Schaer, I. Soboroff, How to measure the reproducibility of system-oriented IR experiments, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (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, ACM, 2020, pp. 349–358. URL: https://doi.org/10.1145/3397271.3401036. doi:10.1145/3397271.3401036.

[13] G. G. Saez, Continuous Evaluation Framework for Information Retrieval Systems, Theses, Université Grenoble Alpes [2020-....], 2023. URL: https://theses.hal.science/tel-04547265.

[14] J. Keller, T. Breuer, P. Schaer, Evaluation of temporal change in ir test collections, in: Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '24), July 13, 2024, Washington, DC, USA, ACM, 2024. doi:10.1145/3664190.3672530.

[15] M. G. Kendall, Rank correlation methods., Griffin, 1948.
[16] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Trans. Inf. Syst. 28 (2010) 20:1–20:38. URL: https://doi.org/10.1145/1852102.1852106. doi:10.1145/1852102.1852106.
[17] N. Rekabsaz, O. Lesota, M. Schedl, J. Brassey, C. Eickhoff, Tripclick: The log files of a large health web search engine, 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. 2507–2513. URL: https://doi.org/10.1145/3404835.3463242. doi:10.1145/3404835.3463242.
[18] S. Althammer, S. Hofstätter, S. Verberne, A. Hanbury, Tripjudge: A relevance judgement test collection for tripclick health retrieval, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 3801–3805. URL: https://doi.org/10.1145/3511808.3557714. doi:10.1145/3511808.3557714.
[19] O. Zendel, J. S. Culpepper, F. Scholer, P. Thomas, Enhancing human annotation: Leveraging large language models and efficient batch processing, in: P. D. Clough, M. Harvey, F. Hopfgartner (Eds.), Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2024, Sheffield, United Kingdom, March 10-14, 2024, ACM, 2024, pp. 340–345. URL: https://doi.org/10.1145/3627508.3638322. doi:10.1145/3627508.3638322.
[20] P. Schaer, N. Tavakolpoursaleh, Historical clicks for product search: GESIS at CLEF LL4IR 2015, in: L. Cappellato, N. Ferro, G. J. F. Jones, E. SanJuan (Eds.), Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015, volume 1391 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015. URL: https://ceur-ws.org/Vol-1391/26-CR.pdf.
[21] N. Tavakolpoursaleh, M. Neumann, P. Schaer, Ir-cologne at TREC 2017 opensearch track: Rerunning popularity ranking experiments in a living lab, 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. URL: https://trec.nist.gov/pubs/trec26/papers/IR-Cologne-O.pdf.
[22] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns 4 (2023) 100804. URL: https://doi.org/10.1016/j.patter.2023.100804. doi:10.1016/J.PATTER.2023.100804.