# Team OpenWebSearch at LongEval: Using Historical Data for Scientific Search

Daria Alexander[1], Maik Fröbe[2], Gijs Hendriksen[1], Matthias Hagen[2], Djoerd Hiemstra[1], Martin Potthast[3] and Arjen P. de Vries[1]

[1]*Radboud Universiteit Nijmegen*
[2]*Friedrich-Schiller-Universität Jena*
[3]*University of Kassel, hessian.AI, ScaDS.AI*

**Abstract**

We describe the submissions of the OpenWebSearch team for the CLEF 2025 LongEval Sci-Retrieval track. Our approaches aim to explore how historical data from the past can be re-used to build effective rankings. The Sci-Retrieval track uses click-data and documents from the CORE search engine. We start all our submissions from rankings of the CORE search engine that we crawled for all queries of the track. This has two motivations: first, we hypothesize that a good practical search engine should only make minor improvements in the ranking at a time (i.e., we would like to only make small adjustments to the production ranking), and, second, we hypothesize that only documents that are in the top ranks of the CORE ranking can be relevant in the setup of LongEval where relevance is derived from clicks (i.e., we try to incorporate the position bias of the clicks into our rankings). Based on this crawled CORE ranking, we try to make improvements via qrel-boosting, RM3 keyqueries, clustering, monoT5 re-ranking and user intent prediction. Our evaluation shows that qrel-boosting, RM3 keyqueries, clustering and intent prediction improve the CORE ranking that we re-rank.

**Keywords**

Keyquery, User Intent Prediction, Counterfactual Query Rewriting

## 1. Introduction

Historical query logs can help to improve future retrieval models as the relevance labels might be transferrable across time. The CLEF LongEval retrieval task [1, 2, 3, 4, 5, 6, 7, 8] studies this scenario and provides relevance labels mined from click longs of past user interactions, allowing retrieval systems to optimize rankings for future queries. Especially for recurring queries, which a retrieval system observed in the past and future test periods, there is substantial opportunity to exploit past relevance judgments.

This year, LongEval introduced a scientific search retrieval task [8, 9]. Scientific search focuses on retrieving and profiling information objects related to scholarly research [10]. These systems are designed to locate scientific papers, theses, technical reports, and other academic materials from curated and often domain-specific repositories. Unlike general web search, scientific search emphasizes high precision and relevance, often including bibliometric data such as citations, authorship, publication venues, and institutional affiliations. Scientific search engines play an important role in supporting the discovery, navigation, and evaluation of scientific literature across different fields.

In this paper, we describe the submissions of the OpenWebSearch team for the CLEF 2025 LongEval Sci-Retrieval track. We build on top of our submissions to LongEval 2023 and 2024 [11, 12, 13], focussing again on leveraging historical interaction data to improve document rankings in a realistic and constrained retrieval setting. The Sci-Retrieval track provides query logs and click data derived from the CORE search engine, along with its associated document corpus. We build all our submissions starting from the top-ranked results of the CORE search engine, which we crawled for every query in the track across different document fields (title, abstract, and full text). This design choice is grounded in two main hypotheses. First, we assume that a practical search engine should make only incremental improvements to existing rankings – reflecting realistic production constraints and user expectations. Second, we believe that most relevant documents are already among the top results shown by the CORE

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*

engine, especially since LongEval defines relevance based on user clicks (i.e., we try to incorporate the position bias into our rankings).

Building on these initial rankings, we explore multiple techniques to improve retrieval effectiveness: qrel-boosting based on past relevance (as proposed by Keller et al [12, 14]), RM3 keyquery expansion, cluster-based boosting, monoT5 re-ranking of top results, and user intent prediction (such as suggested by [15, 16]). Each of these methods is designed to re-rank or selectively boost documents within the CORE results to increase the likelihood of retrieving relevant content. Our evaluation shows that qrel-boosting, RM3 keyqueries, clustering, and intent prediction consistently improve the re-ranked CORE results, supporting our hypothesis that lightweight, targeted adjustments can lead to meaningful gains in effectiveness. Our code is available online.[1]

## 2. Related Work

We review related work on redundancy in information retrieval setups, keyqueries, and user intent classification that form the basis of our submissions.

**Redundancy in Information Retrieval Setups.** Normally, it is good practice to avoid redundancy between training, validation, and test splits in experiments, as otherwise, the effectiveness could be overestimated due to train-test leakage [17, 18]. Especially for IR experiments, redundant documents might cause effectiveness scores to be overestimated because retrieval models get a reward for showing the same document multiple times [19, 20]. Similar problems can occur for learned models that might overfit to redundancy in the training data [21]. However, in the LongEval scenario, redundancy emerges naturally, as queries and documents might overlap over time, which is no form of train-test leakage as the datasets are partitioned over time [1, 3]. In this setting, redundant data might be especially helpful, e.g., as previously showcased when relevance judgments were transferred from the ClueWeb09 corpus to ClueWeb12 via near-duplicate detection [22]. We followed this approach and transferred the relevance judgments to the newer dataset splits in the LongEval scenario.

**Keyqueries** The concept of keyqueries [23] aims to formulate a query that retrieves a set of target documents at the top positions and has been applied to scholarly search [24], medical search [25], privacy scenarios [26], etc. For a set $D$ of documents, a query $q$ is a *keyquery* against some retrieval system $S$, if $q$ fulfills the following three conditions [23]: (1) every $d \in D$ is in the top-$k$ results returned by $S$ for $q$, (2) $q$ has at least $l$ results, and (3) no $q' \subset q$ fulfills the first two conditions. The first two conditions (i.e., the parameters $k$ and $l$) determine the desired specificity and the generality of a keyquery, while the third condition is a minimality constraint to avoid adding further terms to a query that already retrieves the target records at high ranks. Previous work applied this concept only to static corpora, but we extended it to evolving corpora in the LongEval scenario.

**User Intent Classification** When users type queries in search engines, they often have a specific intent in mind. Broder [27] divided queries into three categories according to their intent: navigational, transactional and informational. An informational intent refers to acquiring some information from a website, a navigational intent consists of searching for a particular website, a transactional intent refers to obtaining some services from a website (e.g. downloading a game). The follow-up studies that utilised Broder's taxonomy to classify user intent usually chose two out of three Broder's categories: either informational and navigational [28, 29], or informational and transactional [30]. They adopted different techniques such as computing the scores of distribution of query terms [29], classification of queries into topics [30] as well as tracking past user-click behavior and anchor-link distribution [28].

Broder taxonomy's categories were also used in scientific search. Khabsa et al. [31] classified the queries in CiteSeerX logs into two categories: navigational and informational. According to them,

[1] https://github.com/OpenWebSearch/LONGEVAL-25

navigational queries in scientific search contain full titles of the papers, or some keywords from the title and authors' names, while informational queries usually contain concepts. There are also domain specific user intent taxonomies in scientific search. Rohatgi et al. [32] suggested title, concept and author intent categories; Xiong et al. [33] proposed concept, author, exploration, title, and venue categories.

## 3. Methodology

During our last year participation at LongEval, we incorporated relevance information from past click logs into the query reformulation process using keyqueries [11]. We also utilised this information for the indexing process via a reverted index that contains the top ranked documents per query [34]. Finally, we incorporated both approaches into learning-to-rank pipelines, ensuring that retrieval is also possible for novel queries that were not seen before. This year, we decided to participate in the LongEval-Sci task, as improving retrieval effectiveness in scientific search is less explored compared to web search (for instance, there are only a few corpora for this, and we have some prior experience from building and evaluating the IR Anthology [35, 36]). For this task we aimed to continue with keyqueries and are also adding qrel boosting, cluster boosting, and an intent-aware layer. All our implementations used the `ir_datasets` [37] extensions[2] for LongEval [38]. We submitted all our approaches as run submissions and tracked the execution in most cases with the TIREx Tracker [39] in the `ir_metadata` format [40], but we intend to submit our best approaches as software submission to improve reproducibility to TIRA/TIREx [41, 42] after the deadline to re-run them without modification next year.

### 3.1. Fusion with CORE

We began by crawling the top-25 results returned by the CORE search engine[3] for each query in the 2025 Sci track of the LongEval benchmark. To capture diverse relevance signals, we performed this retrieval separately for different document fields: title, abstract, and full text. Each field-specific retrieval represented a distinct ranking produced by the CORE search engine. We verified that the IDs are the same by looking at the titles of the documents from the CORE search engine and the titles of the Longeval corpus. We further removed all documents from the CORE ranking that were not present in the LongEval corpus (we make the original rankings available online for better reprodicibility[4]). To construct a more robust and comprehensive final ranking, we applied a fusion strategy that combines the rankings from the individual fields (i.e., title, abstract, and full text). After combining the results, we further enhanced the final ranking by filling in any missing positions with additional documents retrieved using a BM25 model over the full text (i.e., we appended the BM25 results to the CORE ranking). This process ensures broad coverage and incorporates complementary evidence from multiple content fields to improve the overall ranking quality.

### 3.2. Reranking with MonoT5

To further refine the quality of our fused rankings, we applied a neural reranking step using the `castorini/monot5-base-msmarco` model [43]. Specifically, we reranked the top-10 documents (abstracts) from the fused results generated by the CORE search engine. The monoT5 model is a sequence-to-sequence transformer trained on the MS MARCO dataset for passage reranking tasks. It takes a query and a candidate document (the default text as implemented in the `ir_datasets` LongEval package) as input and outputs a relevance score by predicting whether the document is relevant to the query. By applying MonoT5 to the highest-ranked candidates, we aimed to better capture semantic relevance beyond lexical overlap, leveraging the model's deep language understanding to improve precision at the top of the ranking.

[2] https://github.com/clef-longeval/ir-datasets-longeval
[3] https://core.ac.uk/
[4] https://huggingface.co/datasets/gijshendriksen/LongEval

### 3.3. Qrel Boosting

We applied the qrel boosting approach [14] provided as the official baseline[5] in the default configuration without modification to the fused CORE rankings. Conceptually, this means that for the queries that have relevant documents in the past that still exist in the corpus, we move those known relevant documents to the top. Independent of any prior ranking, all documents $d$ ranked for a query $q$ at $t_n$ (timeslot) that were judged at a previous timeslot, e.g. $t_{n-1}$, were boosted by:

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

The additional free parameter $\mu$ can be used to assign a different boost to documents with higher relevance labels. For queries without known previous rankings, the CORE-fusion ranking remained unchanged.

### 3.4. Keyqueries

As in the previous editions, some queries overlapped over different time slots, and in case their intent stayed the same, we aimed to transfer their relevance to the new time slots. Consecutively, for those queries we knew what documents had been clicked clicked a few months ago. This run applied RM3 in default configuration inspired by the counterfactual query rewriting with keyqueries by [12] on top of the qrel-boost-core run (i.e., we use the documents that were previously clicked as feedback documents so that RM3 reformulates the query to move the previously clicked documents higher in the ranking). In the previous years, documents could be deleted in later timestamps, therefore, in our previous years submissions we had to insert the clicked documents into the current corpus. However, in this year's Science retrieval task documents are not removed and also do not change their content; therefore, we did not modify the corpus and applied RM3 without changes.

### 3.5. Cluster Boosting

We implemented a cluster boosting strategy using the PyTerrier framework. This approach involved partitioning the document collection into clusters (or shards) based on content similarity or other clustering methods. During training, for each query, we tracked which clusters contained relevant documents across previous timestamps or query interactions. At retrieval time, documents that belonged to these historically relevant clusters were assigned a higher score or were explicitly boosted. The reasoning is that if a cluster has previously provided relevant results for a given query, other documents within the same cluster are more likely to be relevant as well (following the Cluster Hypothesis [44]). This method aims to improve retrieval effectiveness by using implicit relevance signals at the cluster level, rather than treating documents independently.

### 3.6. Adding User Intent

**User Intent Classification** For user intent classification we used Broder's [27] taxonomy that has informational, navigational and transactional categories. We manually annotated 50 queries from the training set and classified their intent. In this manually annotated sample we found informational and navigational queries but no transactional, so we decided to classify the queries into informational and navigational categories. Overall, transactional queries are rarely present in scientific search [31], and typically involve downloading data, such as a dataset or metadata.

As the queries were very short (mean length 1.57 words in the training set), we used additional information from relevant documents, such as the titles, the name of the authors, the text of the abstract. We noticed that some queries had only a few clicked documents, contained the names of the authors

[5] https://github.com/clef-longeval/longeval-code/tree/main/clef25/qrel-boost

and were searching for one specific paper by those authors. Since in those cases the users were likely searching for specific papers, we classified such queries as navigational. Queries that received a high number of clicks and also those that targeted general concepts were classified as informational.

Since informational queries are often categorized by their degree of specificity, such as finding specific facts versus exploring the subject to gather information and learn [45, 46], and the informational queries in LongEval-Sci collection are not aimed at finding specific facts, we refined the informational category and defined those queries as exploratory. As a result, we have two categories for our classification: exploratory and navigational.

To perform query classification for the train and the test sets, we did automatic classification using weak supervision. Weak supervision is an approach in machine learning where noisy, limited, or imprecise sources are used instead of (or along with) gold labelled data. We used Snorkel [47], which is an end-to-end system for creating labelling functions and training and evaluating the labelling model. In Snorkel the final intent of the queries is determined by the labeling functions using different labeling systems, such as majority voting.

For the classification, we adapted the same approach as Rohatgi et al. [32] and filtered out the queries of one intent category, assuming that other queries belong to the other intent category. We filtered out navigational queries, by establishing the following characteristics for them:

- Only a few document clicks (up to 3 clicks).
- A query of more than two words that is fully contained within the title of a clicked paper as an exact sequence.
- A query is an author name (or author names), and only one paper by this author (or these authors) is clicked.

After the classification, we found that 12 % of the queries in the training set were navigational. That corresponded to the manually annotated subset, which contained 10 % of navigational queries. Since there were overlapping queries in the 2025-01 test set timeslot, we were able to use information from the previously relevant documents to classify their intent. This test set timeslot contained 5 % of navigational queries, which is understandable, as for 337 out of 492 queries we did not have information about previously relevant documents. Since there was no overlap between the training set and the 2024-11 test set timeslot, we could not rely on previously relevant documents and could not filter out any navigational queries.

**An Intent-Aware System.** We chose 2 different strategies for the intent-aware system. For the navigational queries, as the aim was to find a specific paper, we decided to move known relevant documents to the top. For the exploratory queries, we used RM3 on top of qrel boosting, as expanding such queries with terms from relevant documents can broaden their scope and improve retrieval effectiveness. In cases where there was no information about previously relevant documents (such as in the 2024-11 test set), it was not possible to separate navigational from exploratory queries. As a result, those queries were classified as exploratory by default.

## 4. Results

We evaluate our retrieval systems in terms of nDCG@10 and a condensed version of nDCG@10 where we remove all unjudged documents from the ranking. The condensed list evaluation was proposed by Sakai [48], and we apply it here because documents that are not retrieved by the original search engine cannot be considered relevant in the evaluation setting of LongEval (still, it is known that this condensed evaluation overestimates effectiveness [49], so a realistic evaluation score is likely between the nDCG@10 and its condensed counterpart).

Table 1 provides an overview of our results. The results indicate that *qrel-boost-core* is the most effective approach when considering documents that have already been judged for relevance. This is

**Table 1**
The effectiveness of six submitted runs on the 2024-11 and 2025-01 test sets. We report the nDCG@10 as well as nDCG@10 when unjudged documents are removed (Cond. nDCG@10).

| Approach / Run | nDCG@10 | | Cond. nDCG@10 | |
|---|---|---|---|---|
| | 2024-11 | 2025-01 | 2024-11 | 2025-01 |
| qrel-boost-core | 0.0610 | 0.2311 | 0.8188 | 0.8146 |
| fusion-with-core | 0.0425 | 0.0592 | 0.8173 | 0.8062 |
| monot5-in-core | 0.0273 | 0.0306 | 0.7993 | 0.7699 |
| query-intent-fusion | 0.1503 | 0.2636 | 0.6942 | 0.6735 |
| rm3-on-qrel-boost | 0.1503 | 0.2526 | 0.6942 | 0.6669 |
| ows-cluster-boosting | 0.1682 | 0.2223 | 0.6549 | 0.6397 |
| BM25 | 0.1615 | 0.2158 | 0.6402 | 0.5991 |

expected, as the method directly promotes documents known to be relevant in the past. For judged documents, *fusion-with-core* also outperforms the baseline, highlighting the benefit of combining documents from the collection with results crawled from the CORE search engine.

Interestingly, reranking the *fusion-with-core* results using MonoT5 did not lead to further improvements. We suspect this is due to the short length of the queries, which may limit the effectiveness of deep semantic models like monoT5. Similarly, *rm3-on-qrel-boost* performed worse than *qrel-boost-core*, likely for the same reason.

When previously relevant documents allowed to distinguish between navigational and exploratory queries, *query-intent-fusion* outperformed *rm3-on-qrel-boost*, but still did not reach the effectiveness of *qrel-boost-core*. This aligns with expectations, as most queries in the dataset are exploratory, and *rm3-on-qrel-boost* is applied to these. Still, applying *qrel-boost-core* to navigational queries leads to improvements for judged documents, emphasizing the value of tailoring retrieval strategies to different types of user intent.

For nDCG@10 across all documents, intent-aware retrieval achieves the best performance when previously relevant documents are known, highlighting the importance of aligning retrieval strategies with user intent. Please note that the intent-aware retrieval only has overlapping queries available for the 2025-01 timeslot, but in this scenario, it performs well. When such relevance information is not available (i.e., in the 2024-11 setting), *ows-cluster-boosting* proves to be the most effective method, demonstrating its usefulness in initial retrieval scenarios.

## 5. Conclusion

Our experiments for the CLEF 2025 LongEval Sci-Retrieval track demonstrate that incorporating historical relevance information can considerably increase retrieval effectiveness. To improve retrieval effectiveness in scientific search, we performed RM3 keyquery expansion, cluster-based boosting, monoT5 re-ranking of top results, and user intent prediction. Overall, our findings suggest that modest, targeted interventions, especially those guided by relevance history and user intent, can lead to substantial improvements over production rankings in real-world search settings. In our experiments, we explicitly boost the existing position bias of systems. Therefore, interesting directions for future work might be to verify how alternative relevance judgments (i.e., not derived from the production search engine) can be applied to the evaluation, for instance, via simulations or large language model relevance assessors.

## 6. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to perform a grammar and spelling check and find synonyms to make the vocabulary more diverse. After using this tool the authors

reviewed and edited the content as needed and take full responsibility for the publication's content.

## Acknowledgments

This work has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

## References

[1] 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, H. T. Madabushi, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Longeval: Longitudinal evaluation of model performance at CLEF 2023, 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 Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 499–505. URL: https://doi.org/10.1007/978-3-031-28241-6_58. doi:10.1007/978-3-031-28241-6_58.

[2] R. Alkhalifa, I. M. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. E. Anke, G. N. 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, Extended overview of the CLEF-2023 longeval lab on longitudinal evaluation of model performance, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2181–2203. URL: https://ceur-ws.org/Vol-3497/paper-184.pdf.

[3] 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: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 3086–3094. URL: https://doi.org/10.1145/3539618.3591921. doi:10.1145/3539618.3591921.

[4] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. E. Anke, T. Fink, G. G. Sáez, P. Galuscáková, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Longeval: Longitudinal evaluation of model performance at CLEF 2024, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part VI, volume 14613 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 60–66. URL: https://doi.org/10.1007/978-3-031-56072-9_8. doi:10.1007/978-3-031-56072-9_8.

[5] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. E. Anke, T. Fink, G. G. Sáez, P. Galuscáková, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Overview of the CLEF-2023 LongEval Lab on Longitudinal Evaluation of Model Performance, in: G. F. N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

[6] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, T. Fink, P. Galuščáková, G. Gonzalez-Saez, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, A. Zubiaga, Overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2024.

[7] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. E. Anke, T. Fink, P. Galuscáková, G. G. Sáez, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, A. Zubiaga, Extended overview of the CLEF 2024 longeval lab on longitudinal evaluation of model performance, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2267–2289. URL: https://ceur-ws.org/Vol-3740/paper-213.pdf.

[8] M. Cancellieri, A. El-Ebshihy, T. Fink, P. Galuscáková, G. G. Sáez, L. Goeuriot, D. Iommi, J. Keller, P. Knoth, P. Mulhem, F. Piroi, D. Pride, P. Schaer, Longeval at CLEF 2025: Longitudinal evaluation of IR model performance, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V, volume 15576 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 382–388. URL: https://doi.org/10.1007/978-3-031-88720-8_58. doi:10.1007/978-3-031-88720-8_58.

[9] M. Cancellieri, A. El-Ebshihy, T. Fink, P. Galuščáková, G. Gonzalez-Saez, L. Goeuriot, D. Iommi, J. Keller, P. Knoth, P. Mulhem, F. Piroi, D. Pride, P. Schaer, Overview of the CLEF 2025 LongEval Lab on Longitudinal Evaluation of Model Performance, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[10] X. Li, B. J. Schijvenaars, M. de Rijke, Investigating queries and search failures in academic search, Information processing & management 53 (2017) 666–683.

[11] D. Alexander, M. Fröbe, F. Schlatt, M. Hagen, D. Hiemstra, M. Potthast, A. P. de Vries, Team OpenWebSearch at CLEF 2024: LongEval, in: Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR Workshop Proceedings, 2024.

[12] J. Keller, M. Fröbe, G. Hendriksen, D. Alexander, M. Potthast, M. Hagen, P. Schaer, Counterfactual query rewriting to use historical relevance feedback, in: European Conference on Information Retrieval, Springer, 2025, pp. 138–147.

[13] M. Fröbe, G. Hendriksen, A. P. de Vries, M. Potthast, Open web search at longeval 2023: Reciprocal rank fusion on automatically generated query variants, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2432–2440. URL: https://ceur-ws.org/Vol-3497/paper-195.pdf.

[14] J. Keller, T. Breuer, P. Schaer, Leveraging prior relevance signals in web search, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2396–2406. URL: https://ceur-ws.org/Vol-3740/paper-220.pdf.

[15] D. Alexander, W. Kusa, A. P. de Vries, Orcas-i: queries annotated with intent using weak supervision, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3057–3066.

[16] D. Alexander, W. Kusa, A. P. de Vries, Orcas-i query intent predictor as component of tira, in: S. M. Farzana, M. Fröbe, M. Granitzer, G. Hendriksen, D. Hiemstra, M. Potthast, S. Zerhoudi (Eds.), 1st International Workshop on Open Web Search, number 3689 in CEUR Workshop Proceedings, 2024, pp. 23–29. URL: https://ceur-ws.org/Vol-3689/.

[17] K. Krishna, A. Roy, M. Iyyer, Hurdles to progress in long-form question answering, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 4940–4957. URL: https://doi.org/10.18653/v1/2021.naacl-main.393. doi:10.18653/V1/2021.NAACL-MAIN.393.

[18] M. Fröbe, C. Akiki, M. Potthast, M. Hagen, How Train-Test Leakage Affects Zero-shot Re-

trieval, in: D. Arroyuelo, B. Poblete (Eds.), 29th International Symposium on String Processing and Information Retrieval (SPIRE 2022), volume 13617, Concepción, Chile, 2022. doi:10.1007/978-3-031-20643-6_11.

[19] Y. Bernstein, J. Zobel, Redundant documents and search effectiveness, in: O. Herzog, H. Schek, N. Fuhr, A. Chowdhury, W. Teiken (Eds.), Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005, ACM, 2005, pp. 736–743. URL: https://doi.org/10.1145/1099554.1099733. doi:10.1145/1099554.1099733.

[20] M. Fröbe, J. Bittner, M. Potthast, M. Hagen, The Effect of Content-Equivalent Near-Duplicates on the Evaluation of Search Engines, in: J. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. Silva, F. Martins (Eds.), Advances in Information Retrieval. 42nd European Conference on IR Research (ECIR 2020), volume 12036 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2020, pp. 12–19. doi:10.1007/978-3-030-45442-5_2.

[21] M. Fröbe, J. Bevendorff, J. Reimer, M. Potthast, M. Hagen, Sampling Bias Due to Near-Duplicates in Learning to Rank, in: 43rd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2020), ACM, 2020, pp. 1997–2000. doi:10.1145/3397271.3401212.

[22] M. Fröbe, J. Bevendorff, L. Gienapp, M. Völske, B. Stein, M. Potthast, M. Hagen, CopyCat: Near-Duplicates within and between the ClueWeb and the Common Crawl, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), 44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021), ACM, 2021, pp. 2398–2404. doi:10.1145/3404835.3463246.

[23] M. Hagen, A. Beyer, T. Gollub, K. Komlossy, B. Stein, Supporting Scholarly Search with Keyqueries, in: N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. Di Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016), volume 9626 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2016, pp. 507–520. doi:10.1007/978-3-319-30671-1_37.

[24] M. Völske, T. Gollub, M. Hagen, B. Stein, A keyquery-based classification system for CORE, D Lib Mag. 20 (2014). URL: https://doi.org/10.1045/november14-voelske. doi:10.1045/NOVEMBER14-VOELSKE.

[25] M. Fröbe, S. Günther, A. Bondarenko, J. Huck, M. Hagen, Using keyqueries to reduce misinformation in health-related search results, in: ROMCIR 2022: The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2022: the 44th European Conference on Information Retrieval, 2022.

[26] M. Fröbe, E. O. Schmidt, M. Hagen, Efficient Query Obfuscation with Keyqueries, in: 20th International IEEE/WIC/ACM Conference on Web Intelligence (WI-IAT 2021), ACM, 2021. doi:10.1145/3486622.3493950.

[27] A. Broder, A taxonomy of web search, SIGIR Forum 36 (2002).

[28] U. Lee, Z. Liu, J. Cho, Automatic identification of user goals in web search, in: WWW '05: Proceedings of the 14th international conference on World Wide Web, 2005, pp. 391–400. doi:10.1145/1060745.1060804.

[29] I.-h. Kang, G. Kim, Query type classification for web document retrieval, in: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2004. doi:10.1145/860435.860449.

[30] R. Baeza-Yates, L. Calderon-Benavides, C. González-Caro, The intention behind web queries, volume 4209, 2006, pp. 98–109. doi:10.1007/11880561_9.

[31] M. Khabsa, Z. Wu, C. L. Giles, Towards better understanding of academic search, in: Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries, 2016, pp. 111–114.

[32] S. Rohatgi, C. L. Giles, J. Wu, What were people searching for? a query log analysis of an academic search engine, in: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2021, pp. 342–343.

[33] C. Xiong, R. Power, J. Callan, Explicit semantic ranking for academic search via knowledge graph embedding, in: Proceedings of the 26th international conference on world wide web, 2017, pp.

1271–1279.

[34] J. Pickens, M. Cooper, G. Golovchinsky, Reverted indexing for feedback and expansion, in: J. X. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, A. An (Eds.), Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010, ACM, 2010, pp. 1049–1058. URL: https://doi.org/10.1145/1871437.1871571. doi:10.1145/1871437.1871571.

[35] M. Fröbe, H. Scells, T. Elstner, C. Akiki, L. Gienapp, J. H. Reimer, S. MacAvaney, B. Stein, M. Hagen, M. Potthast, Resources for Combining Teaching and Research in Information Retrieval Coursework, in: G. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, Y. Zhang (Eds.), 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024), ACM, 2024, pp. 1115–1125. doi:10.1145/3626772.3657886.

[36] M. Potthast, S. Günther, J. Bevendorff, J. P. Bittner, A. Bondarenko, M. Fröbe, C. Kahmann, A. Niekler, M. Völske, B. Stein, M. Hagen, The Information Retrieval Anthology, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), 44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021), ACM, 2021, pp. 2550–2555. doi:10.1145/3404835.3462798.

[37] 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. URL: https://doi.org/10.1145/3404835.3463254. doi:10.1145/3404835.3463254.

[38] J. Keller, M. Fröbe, G. Hendriksen, D. Alexander, M. Potthast, P. Schaer, Simplified longitudinal retrieval experiments: A case study on query expansion and document boosting, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2024, Madrid, Spain, September 9-12, 2025, Proceedings, Part I, Lecture Notes in Computer Science, Springer, 2025.

[39] T. Hagen, M. Fröbe, J. H. Merker, H. Scells, M. Hagen, M. Potthast, TIREx Tracker: The Information Retrieval Experiment Tracker, in: 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025), ACM, 2025.

[40] T. Breuer, J. Keller, P. Schaer, ir_metadata: An extensible metadata schema for IR experiments, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 3078–3089. URL: https://doi.org/10.1145/3477495.3531738. doi:10.1145/3477495.3531738.

[41] 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.

[42] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: H.-H. Chen, W. Duh, H.-H. Huang, M. Kato, J. Mothe, B. Poblete (Eds.), 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023, pp. 2826–2836. doi:10.1145/3539618.3591888.

[43] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. URL: https://aclanthology.org/2020.findings-emnlp.63. doi:10.18653/v1/2020.findings-emnlp.63.

[44] C. J. Van Rijsbergen, Information Retrieval, Butterworths, 1979.

[45] M. Kellar, C. Watters, M. Author, A field study characterizing web-based information seeking tasks, JASIST 58 (2007) 999–1018. doi:10.1002/asi.20590.

[46] D. M. Russell, D. Tang, M. Kellar, R. Jeffries, Task behaviors during web search: The difficulty of assigning labels, in: 2009 42nd Hawaii International Conference on System Sciences, IEEE, 2009, pp. 1–5.
[47] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, C. Ré, Snorkel: Rapid training data creation with weak supervision, in: Proceedings of the VLDB endowment. International conference on very large data bases, volume 11, 2017, p. 269.
[48] 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. URL: https://doi.org/10.1145/1277741.1277756. doi:10.1145/1277741.1277756.
[49] 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), volume 13980 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2023, pp. 313–329. doi:10.1007/978-3-031-28244-7_20.