

CIR_cluster at LongEval 2025: Clustering Query Variants for Temporal Generalization

Notebook for the LongEval Lab at CLEF 2025

Arthur Muanza Ndiema¹, Jüri Keller^{1,*} and Philipp Schaer^{1,*}

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

Abstract

This paper details the participation of the CIR_cluster team in the CLEF 2025 LongEval for WebSearch task, for which we submitted four distinct runs. In longitudinal settings, approaches that leverage historical information—such as past relevance judgments—have demonstrated strong effectiveness. However, these methods are limited when no such information is available. For instance, relying on previous clicks is infeasible for queries that have never been issued before. We hypothesize that documents relevant to a given query are also relevant to its semantic variants. Based on this assumption, we cluster queries to identify query variants. This enables us to link previously unseen queries to the histories of its query variants. By that, the extended approaches can generalize not only to new and updated documents but also to new and updated queries.

Our experimental evaluation showed that clustering did not improve the average retrieval effectiveness. However, when query variants could be identified, the performance often polarizes—resulting in either substantial improvements or declines. Although our current approach did not yield overall performance improvements, we think that identifying query variants remains an interesting direction to generalization across queries for ranking approaches that employ prior relevance signals in longitudinal settings.

 https://github.com/irgroup/25-clef-cir_cluster

Keywords

Query Clustering, Relevance Feedback, Longitudinal Evaluation

1. Introduction

In this work, we describe the participation of the CIR_cluster team in the CLEF 2025 LongEval WebSearch task. This task aims to evaluate retrieval systems over time, focusing on how well they adapt to changes in the web and user behavior [1, 2, 3, 4]. The lab provides two dynamic test collections, consisting of multiple snapshots of the same search setting. Each snapshot describes an evolved state of the document corpus, query set, and relevance judgments (qrels). These settings provide a unique opportunity to evaluate the effectiveness of retrieval systems and likewise open up new opportunities to develop relevance signals from past information.

In previous works, we have shown that approaches that leverage historical information, such as past relevance judgments, can achieve strong effectiveness in longitudinal settings [5, 6]. As such, the Qrel Boost (QB) approach boosts relevant query-document pairs from previous snapshots, assuming that if a document was relevant for a query in the past, it is likely to be relevant for the same query in the future. In a similar manner, the Relevance Feedback (RF) approach extends the query with terms from previously relevant documents, assuming that these terms are still relevant for the current query.


Both methods are limited when no prior information is available. For example, relying on previous qrels is infeasible for queries that have never been issued before. In such cases, for unseen queries, the system falls back to the base ranker. Further, it was observed that many similar queries are captured in the test collections. Often, they even differ only on a lexical level, such as slight spelling variations

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

*Corresponding author.

 arthur-muanza.ndiema@smail.th-koeln.de (A. M. Ndiema); jueri.keller@th-koeln.de (J. Keller);

philipp.schaer@th-koeln.de (P. Schaer)

 0000-0002-9392-8646 (J. Keller); 0000-0002-8817-4632 (P. Schaer)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

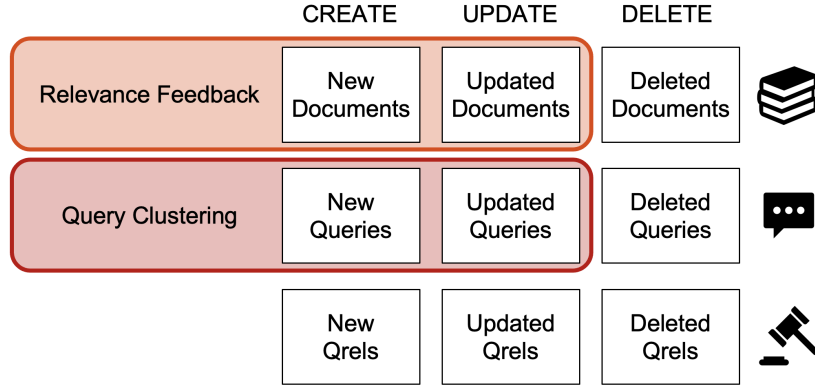


Figure 1: The schema displays how the components generalize in the change matrix as proposed in [9]. The QB systems do not generalize at all. The RF approach generalizes to new documents but not to new queries. By adding the query clustering component, both approaches can generalize to new and updated queries.

or changed word orders [6]. This means that for a supposedly new query, highly similar queries with known relevance judgments might exist, but the systems cannot leverage this information.

To address these limitations, we aim to identify query variants – queries that are related to the same information need but express it in different ways [7, 8]. We hypothesize that documents relevant to a given query are also relevant to its semantic variants. Based on this assumption, we cluster queries and link previously unseen queries to the history of their query variants.

In Figure 1, we illustrate how the approaches generalize across the components that change in the test bed of a dynamic test collection. In this work, we define generalization as the ability of our proposed approaches to handle previously unseen documents and queries. The QB approach does not generalize at all, as it directly relies on known query-document pairs. The RF approach can generalize to the documents dimension, as query reformulations can retrieve new or updated documents. The newly proposed query clustering extends both approaches so that they can be generalized to new and updated queries.

In short, the main contributions of this work are:

- Two clustering approaches to identify query variants,
- the extension of two retrieval approaches that rely on qrels from previous snapshots,
- a preliminary evaluation of the approaches on the LongEval WebSearch train collection.

We make all code, Docker images of the approaches, and cluster information publicly available.¹

2. Approaches

To source more relevance information from prior snapshots and to overcome the limitation that the approaches cannot generalize to new queries, we propose to identify query variants by clustering queries. These clusters are then used in two retrieval systems to produce the submitted rankings.

2.1. Query Clustering

All queries for all snapshots are grouped into clusters of similarity. This clustering was performed based on an early version of the LongEval WebRetrieval dataset, before queries were separated into individual sets per snapshot. In total, 54,658 queries were clustered. In a first step, all unique queries were transformed into 1024-dimensional vectors using sentence transformers and the Lajavaneusse/sentence-camambertlarge model [10, 11, 12].² This model was specifically trained for French texts.

¹https://github.com/irgroup/25-clef-cir_cluster

²<https://huggingface.co/Lajavaneusse/bilingual-embedding-large>

Subsequently, the embedded queries were clustered using k-means and DBSCAN [13, 14]. K-means clustering requires pre-defining the number of clusters beforehand [13]. Since the distribution of queries is unknown, we estimated it based on the results of different k values using the elbow method [14]. The DBSCAN algorithm does not require a target number of clusters and also identifies outliers [15]. This makes it theoretically well suited, as many independent queries can be expected and should be excluded from the clustering. Instead, the minimum points per cluster (MinP) and the maximum distance between two points (ϵ) need to be defined. Both parameters were estimated based on a grid search, the adjusted rand score, and the elbow method.

2.2. Retrieval Systems

The query clustering is used in the retrieval ranking by mapping the query to its cluster. The clusters can be understood as an abstraction of a retrieval topic containing different query variants. The approach is applied to two different retrieval systems initially proposed in the LongEval lab 2024 that were later further refined [5, 6].

Qrel Boost The first system, Qrels Boost (QB), directly boosts query document pairs that were previously relevant. The intuition is that if a document was relevant for a query in the past, it is likely to be relevant for the same query in the future. Initially, a BM25 ranking is created that is then reranked. Each query-document pair of the initial ranking is compared to the qrels of the previous snapshot. If the query-document pair was found, its ranking score is multiplied by a boost depending on the previous relevance label. This can be repeated for multiple previous snapshots. In the submitted runs, we used the default parameters of $\lambda = 0.7$ describing the strength of the boost and $\mu = 2$ as an additional factor for highly relevant documents. We used all available previous snapshots as history. This results in a history of eight snapshots for all the submitted test runs (2022-06 to 2023-09), and accordingly fewer for the training runs. At the first point in time (2022-06), no prior snapshots are available, and the approach falls back to the BM25 ranking.

Instead of boosting query-document pairs that were previously relevant, we boost all documents related to a query from the abstract topic. This means that for the query q_1 of the topic t we now also boost the document d although it only appeared in the qrels for the query q_2 of the same topic t .

Relevance Feedback The second system, Relevance Feedback (RF), similarly to the QB approach, uses previously relevant documents to extend the query. Therefore, the terms with the highest tf-idf scores are extracted from the documents that were previously relevant to a query. The extended query is issued to a BM25 retriever. If no previously relevant documents are available, the system only uses the original query. We used the default parameters of 10 feedback terms from the top 3 documents. For this approach, we also used all available history.

Instead of using only the previously relevant documents of the query, we use all previously relevant documents of the topic. This means that for the query q_1 of some topic t we now also use document d although it only appeared in the qrels for the query q_2 of topic t .

3. Experimental Evaluation

First the clustering methods evaluated in different settings and then the retrieval systems employing the clusters are validated in the eight training snapshots of the LongEval WebSearch dataset.

3.1. Query Clustering

The query clusters are created based on the initial version of the LongEval WebSearch dataset from 2025. At this stage only one query set for all snapshots were published. While this was later reverted and many queries were removed, the query IDs remained the same and the cluster are still valid.

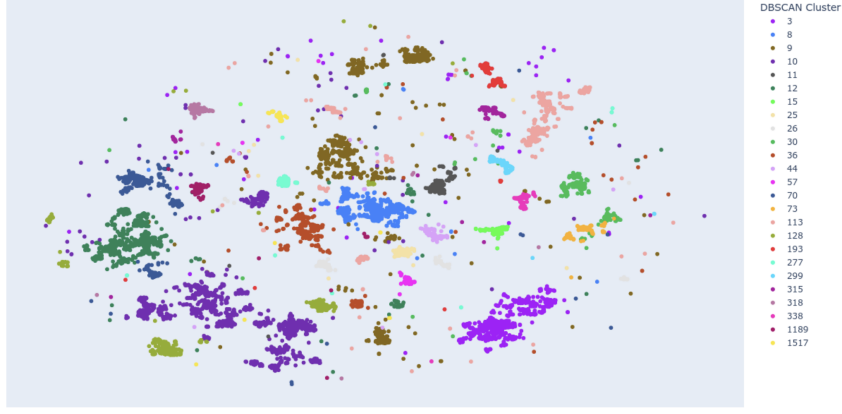


Figure 2: Selected query clusters based on DBSCAN and the selected parameter $\epsilon = 0.2$ and $MinP = 2$ visualized by t-SNE [16].

Table 1

Average nDCG@10 for each system and snapshot. The best results are highlighted in **bold**. Since no history is available for 2022-06, all approaches rely on the first stage BM25 results.

snapshot	2022-06	2022-07	2022-08	2022-09	2022-10	2022-11	2022-12	2023-01	2023-02
bm25	0.194	0.192	0.193	0.168	0.24	0.243	0.254	0.258	0.258
QB	0.193	0.339	0.379	0.314	0.335	0.388	0.427	0.443	0.504
QB-kmeans	0.193	0.343	0.383	0.322	0.322	0.36	0.38	0.387	0.447
QB-dbscan	0.193	0.331	0.362	0.264	0.281	0.319	0.339	0.349	0.392
RF	0.193	0.352	0.371	0.304	0.323	0.368	0.381	0.387	0.41
RF-kmeans	0.193	0.335	0.344	0.23	0.223	0.225	0.227	0.231	0.259
RF-dbscan	0.193	0.316	0.324	0.189	0.226	0.251	0.264	0.272	0.281

Based on the results, for the k-means clustering $k = 5000$ clusters were derived as optimal values. For DBSCAN a maximum distance of $\epsilon = 0.2$ and $MinP = 2$ was chosen. This results in 5334 clusters, a similar value to k-means. 27.9% of the 54,864 queries are classified as outliers. This means that no variants for those queries could be identified, and the approaches can only rely on the original query. Most clusters consist of only two queries that often differ only in spelling or minor variations. For example the query *chateau de villiers-le-mahieu* (75386) and *château de villiers-le-mahieu* (74083) both belong to the cluster 5327. Other clusters differ more strongly, for example, the cluster 5245 with the queries *loi militaire* and *projet de loi militaire*. Bigger clusters capture whole categories or sub-domains. For example, the biggest cluster 10 consists of 3067 food-related queries like *pomme de terre*, *gateau*, or *recette*. Figure 2 visualizes the cluster with DBSCAN. In comparison, the k-means clusters are much smaller with 139 queries per cluster at most. Only 212 clusters contain only two queries. This yields clusters of similar sizes.

3.2. Retrieval Experiments

For an initial validation, we tested both approaches with both clustering methods on all previous training snapshots. Additionally, we compared the results to the official QB and RF baselines and also BM25. The results on nDCG@10 are reported in Table 1 and Figure 3. Only the k-means clustering improves the effectiveness for the QB system at the snapshots 2022-08 and 2022-08, and only by a little. The official baselines outperform all other combinations. Figure 3 indicates a drop in effectiveness at 2022-09 for all systems. After that snapshot, the effectiveness increases again, for BM25 only slightly, and especially for the QB approaches more strongly. While the official RB baseline shows a clearly better effectiveness, its clustering extensions are mostly comparable to BM25. Notably, the RF approaches maintain the delta to BM25 over the snapshots, while the QB approaches further expand it.

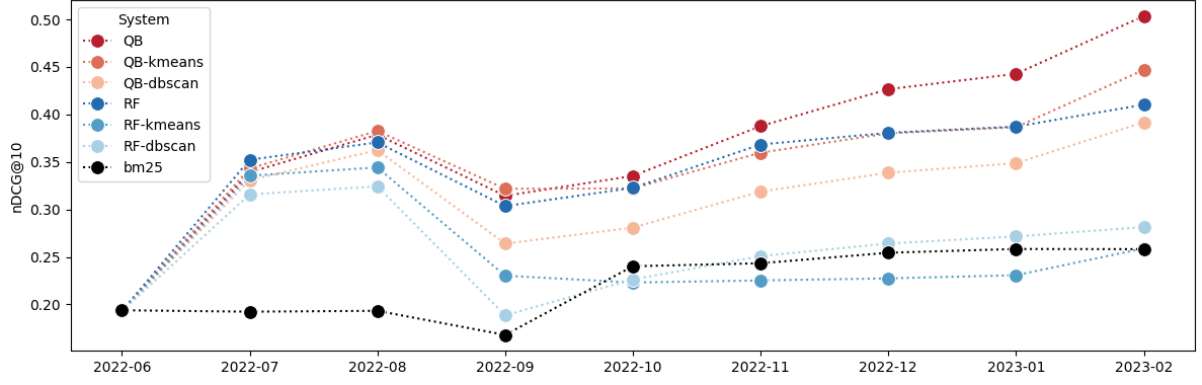


Figure 3: Line plot of the NDCG@10 scores for the different approaches at different snapshots. QB approaches are shown in red and RF approaches in blue. Except for the 2022-08 and 2022-09, the query clustering impairs the average effectiveness, sometimes drastically.

4. Discussion

Foundational to the proposed approaches is the query clustering. The results showed that the different clustering methods could identify some query variants. However, these good clusters are overshadowed by too large and diverse clusters. In the context of the whole query distribution, queries from those clusters are well related, for example, the queries *pomme de terre* and *gateau*, but clearly too different for the same documents to be assumed to be relevant. Better parameters could improve the clustering, but finding them is a challenging endeavor, especially in the context of a continuous query stream. Additionally, other features, such as the clicked or relevant documents for a query, could be used as additional features.

Regarding the implementation of the retrieval approaches, the clustering could be replaced with a similarity function that finds similar queries up to a certain threshold. The approaches utilize all previous snapshots. This means that for later snapshots, many more prior qrels are available. This could support the observation that the QB system over time diverges from BM25. Regarding the RF systems, the effect remains unclear, and more tests with different histories are needed. Since the tf-idf scores of expansion terms are compared across all snapshots, outliers that only appeared once could strongly influence the results.

The proposed approaches are limited in different ways. Both clustering approaches do not differentiate between the original query and query variants of the cluster. This means that a query variant from the same topic, focusing on a specific aspect, can introduce highly specific terms for query expansion, although they may only be relevant to some aspects of the topic. More sophisticated methods are needed that differentiate between core queries and more distantly related query variants. For example, RM3 weights the terms individually. A similar approach can be implemented based on the similarity between the original query and its variations.

5. Conclusion

In this paper, we presented our participation in the CLEF 2025 LongEval WebSearch task. We proposed a query clustering approach to identify query variants and applied it to two retrieval systems: Query-Based (QB) and Relevance Feedback (RF). The approaches were initially evaluated on the training collection of the task. Unfortunately, the results did not show improvements in retrieval effectiveness compared to the baselines. However, we observed that when meaningful query variants were identified, the performance can improve.

Acknowledgments

We gratefully acknowledge the support of the German Research Foundation (DFG) through project grant No. 407518790.

Declaration on Generative AI

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

References

- [1] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.
- [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] 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, Overview of the CLEF 2024 longeval lab on longitudinal evaluation of model performance, in: CLEF (2), volume 14959 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 208–230.
- [4] 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: ECIR (5), volume 15576 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 382–388.
- [5] J. Keller, T. Breuer, P. Schaer, Leveraging prior relevance signals in web search, in: CLEF (Working Notes), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2396–2406.
- [6] J. Keller, M. Fröbe, G. Hendriksen, D. Alexander, M. Potthast, M. Hagen, P. Schaer, Counterfactual query rewriting to use historical relevance feedback, in: ECIR (3), volume 15574 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 138–147.
- [7] P. Bailey, A. Moffat, F. Scholer, P. Thomas, User variability and IR system evaluation, in: SIGIR, ACM, 2015, pp. 625–634.
- [8] P. Bailey, A. Moffat, F. Scholer, P. Thomas, UQV100: A test collection with query variability, in: R. Perego, F. Sebastiani, J. A. Aslam, I. Ruthven, J. Zobel (Eds.), Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17–21, 2016, ACM, 2016, pp. 725–728. doi:10.1145/2911451.2914671.
- [9] J. Keller, T. Breuer, P. Schaer, Evaluation of temporal change in IR test collections, in: ICTIR, ACM, 2024, pp. 3–13.
- [10] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: EMNLP/IJCNLP (1), Association for Computational Linguistics, 2019, pp. 3980–3990.
- [11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: ACL, Association for Computational Linguistics, 2020, pp. 8440–8451.
- [12] N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, Augmented SBERT: data augmentation method

- for improving bi-encoders for pairwise sentence scoring tasks, in: NAACL-HLT, Association for Computational Linguistics, 2021, pp. 296–310.
- [13] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, University of California Press, 1967, pp. 281–297.
 - [14] R. L. Thorndike, Who belongs in the family?, *Psychometrika* 18 (1953) 267–276. doi:10.1007/BF02289263.
 - [15] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: KDD, AAAI Press, 1996, pp. 226–231.
 - [16] L. van der Maaten, G. Hinton, Visualizing data using t-sne, *Journal of Machine Learning Research* 9 (2008) 2579–2605. URL: <http://jmlr.org/papers/v9/vandermaaten08a.html>.