# Counterfactual Query Rewriting for Historical Relevance Feedback

First Author[1], Maik's Mic[2,3], and Third Author[3]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com

**Abstract.** Search engines have seen many of the submitted queries before, which might provide valuable relevance feedback for recurring queries that did not change their intent. However, this relevance feedback can often not be applied directly, as documents that were previously relevant to a query might not exist anymore or might have substantially changed content. By counterfactually assuming that the previously relevant document still exists, we can formulate so-called keyqueries so that the previously relevant documents, if they would still exist, would be retrieved at the top positions of the ranking. Our evaluations in the LongEval scenario with varying time gaps between 1 month to 1 year and naturally and simulated removed/changed documents show that XYZ.

**Keywords:** Query Rewriting · Longtitutal evaluation · Another keyword.

## 1 Introduction

Many of the queries a search engine receives have been seen before [CITATION]. By analyzing how users interact with the displayed results for these queries, valuable relevance information can be gathered. Such signals can be directly leveraged to improve the search engine's effectiveness. For example, based on the documents users click for a query, click models can be constructed that synthesize a relevance indicator. Based on these created labels, documents can be boosted or rerankers trained. While relevance feedback has shown to be effective, exploiting it as a boosting feature directly has a limited application since the relevance signal does not naturally generalize to new topics and documents and may lead to unwanted biases.

Time plays a crucial role since the queries, documents, and even the relevance may evolve over time. Documents may change their content, queries their intent, and even if both are static, the relevance may change because of external factors.

To address these challenges, we explore how previous relevance labels can be used for query expansion in a temporal setting. Therefore, we first describe different classes of temporal changes in the web search setting and propose to create key queries based on the previously relevant documents.

[key query introduction]

**Table 1.** TODO: This table is from CLEF22 [5], I would like to abstract from the framework by Keller et al. [9] to show simple examples for the classes that we look into. Examples of differences between versions of positive training instances from MS MARCO Version 1 (crawled in 2018) and Version 2 (2021). Text fragments highlighted in blue italics indicate relevance (erroneous versions have no blue italics).

| **Query** | **Relevant Document** | | **Comment** |
|---|---|---|---|
| | Version 1 (2018) | Version 2 (2021) | |
| what are deposit solutions banking | Oops! There was a problem! We had an unexpected problem processing your request. | *Deposit Solutions* Crunchbase *Company Profile* . . . | Crawling error in V1 |
| what are yellow roses mean | Meaning Of A Yellow Rose . . . a yellow rose *stands for joy and happiness* . . . | 20 Best Knockout Roses To Make Your Garden Outstanding | Redirect in V2 |
| how much magnesium in kidney beans | Kidney Beans . . . *a cup of kidney beans contains 70 mg of magnesium* . . . | Magnesium Grocery List. Bring this list to the store to . . . | Content change in V2 |
| Intent Change | | | |
| soccer em (queried in 2024 vs 2020 | | | |

In an initial experimental evaluation we compare the key query approach to different query expansion and pseudo relevance feedback baselines on the evolving LongEval test collection. The results indicate that through these methods, we can exploit relevance feedback beyond known query document pairs to new documents.

## 2 Related Work

**Table 2.** Taxonomy reproduced and adapted from [9].

| | `CREATE` | `UPDATE` | `DELETE` |
|---|---|---|---|
| Document | New unseen documents | Known URL but changed website | URL not in subcollection anymore |
| Topic | New query | Fixed typos, translation or changed accents marks | Query not in subcollection anymore |
| Qrel | New relevance label | Known document query pair but changed relevance | Relevance became unknown |

Changes in evolving test collections are classified on a high level along theire main components, the documents, topics, and qrels, and the create, update, and delete operation of persistent storage by Keller et al. in [9]. Based on this classification schema different retrieval scenarios emerge that can be quantified on different levels and directly related to challenges IR systems face.

While the LongEval dataset contains changes in all components of all types, the investigated approaches are mainly concerned with changing documents that may also affect the relevance label.

Things that sound related but are not yet checked/incorporated are appended to this sentence [11].

Integrate prior work into the topology by Keller et al. [9]. E.g., transfer of relevance labels to (near-)duplicate documents [6], documents that change their content and therefore might be not relevant anymore [5], etc (Harrie Oosterhuis likely has much on this).

## 3 XY for Previously seen Queries

Our approach.

We noticed that queries overlap over different time slots, and in case their intent stays the same, we aim to transfer their relevance information to the new time slots. Consecutively, for those queries we know what documents were clicked a few months ago. We decided to use this feedback and query expansion with the BO1 model [3] to create keyqueries and use the same approach as [7]. Thereby, we use BO1 to obtain candidate terms for query terms, as pilot experiments showed that BO1 expansion terms yield higher effectiveness than RM3 [1] expansions. We inserted the clicked documents into the current corpus and reformulated the queries with the BO1 model until those documents were in the top positions. After that we removed old documents from the ranking. This implementation of the keyquery concept is not the most effective one, more effective approaches that leverage a generate-and-test paradigm [8] exist and are interesting directions for future work (i.e., explicitly generating many variants and selecting the variants that are highly effective).

Short analysis: how much did documents change? We use CopyCat [6], as it was previously used to deduplicate web crawls, e.g., the ClueWebs, in default settings. ToDo: add motivation for measure and containment conceptually implemented by the S3 score.

## 4 Evaluation

The LongEval test collections capture dynamic evaluation scenarios with different changes in all components. It evolves naturally over time, and a natural overlap between documents, queries, and qrels occurs. Since the explored approaches systematically exploit this overlap, the evaluation is challenging. Since the overlap occurs naturally in the test collection, evaluating the systems in this

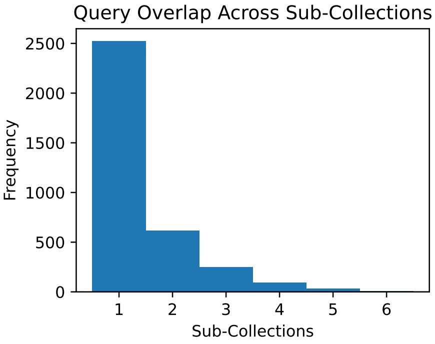

**Fig. 1.** Frequency of queries and points in time.

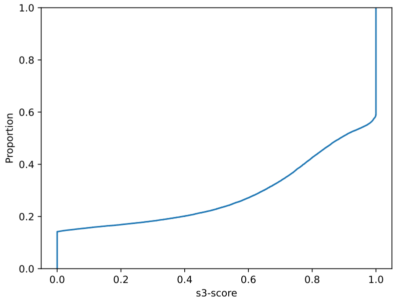

**Fig. 2.** $S_3$ Similarities of documents with overlapping URLs as eCDF plot.

life-like scenario with redundancies is valid. However, we can hardly make any assumptions about how well the approaches generalize. Therefore, we include a second evaluation scenario that excludes any redundancies in the relevant documents.

### 4.1 Baselines

The keyqueries approach is compared to further baselines. The qrel boost method directly boosts query document pairs of BM25 [12] that are known to be relevant from past points in time up based on a weighting factor $\lambda^2$ and all known and not relevant query document pairs down by $(1-\lambda)^2$. Additionally, since the LongEval test collection has graded relevance labels, the score of all highly relevant query document pairs is additionally multiplied by $\mu$. If more than one previous point in time is used for the boosting, the score of a query document pair is repeatedly multiplied by the boost. While this approach appears to be highly effective [2, 10], it can not generalize to new documents and is therefore not effective for the cross-validation and only included as an upper bound naive baseline.

Improving on this baseline, a tf-idf query expansion based on previously relevant documents wes tested. Instead of directly boosting known relevant query document pairs, the top 10 tf-idf terms from the relevant documents are used to expand the original query. The expanded query is then used to query the corpus with BM25. Like the qrel boost approach before, the tf-idf query expansion also only affects topics that are already known but can generalize to new documents.

### 4.2 Natural Evolving Test Collection

In this evaluation scenario, the effectiveness is assessed in a life-like setting where the overlap naturally occurs. Thus, this setting can describe the effectiveness as it may occur in a web search scenario. Currently, the test collection covers six pints in time with, on average, 1.77 million documents. For each point in time between 407 and 1518 queries are logged and on average, 12874 qrels are constructed through the simplified Dynamic Bayesian Network (sDBN) Click Model [4]. The topic overlap between points in time is displayed in Figure 1, indicating that less than a third of the queries are logged repeatedly.

### 4.3 Temporal Per Topic Cross Validation

To emphasize the actual effect of the approaches and how well they generalize to new documents, we evaluated them in a cross-validation setting. For each topic at each point in time, the relevant documents are split into $k$ folds. For each fold, the test, all documents except the ones in the train split are indexed and retrieved. The different approaches can then use the documents in the train split if they were already present at the previous points in time. Since some documents can be relevant for multiple topics, the train test splits must be updated after creation to avoid duplicates. This makes the splits less random and differently large but should be neglectable for this initial evaluation.

## 5 Conclusion and Future Work

TBD.

## References

1. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: Umass at trec 2004: Novelty and hard. Computer Science Department Faculty Publication Series p. 189 (2004)
2. Alkhalifa, R., Borkakoty, H., Deveaud, R., El-Ebshihy, A., Anke, L.E., Fink, T., Galuscáková, P., Sáez, G.G., Goeuriot, L., Iommi, D., Liakata, M., Madabushi, H.T., Medina-Alias, P., Mulhem, P., Piroi, F., Popel, M., Zubiaga, A.: Extended overview of the CLEF 2024 longeval lab on longitudinal evaluation of model performance. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024. CEUR Workshop Proceedings, vol. 3740, pp. 2267–2289. CEUR-WS.org (2024), https://ceur-ws.org/Vol-3740/paper-213.pdf
3. Amati, G.: Divergence from randomness. (2003)
4. Chapelle, O., Zhang, Y.: A dynamic bayesian network click model for web search ranking. In: Quemada, J., León, G., Maarek, Y.S., Nejdl, W. (eds.) Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009. pp. 1–10. ACM (2009). https://doi.org/10.1145/1526709.1526711, https://doi.org/10.1145/1526709.1526711
5. Fröbe, M., Akiki, C., Potthast, M., Hagen, M.: Noise-reduction for automatically transferred relevance judgments. In: Barrón-Cedeño, A., Da San Martino, G., Esposti, M.D., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF

2022). Lecture Notes in Computer Science, vol. 13390, pp. 48–61. Springer, Berlin Heidelberg New York (Sep 2022). https://doi.org/10.1007/978-3-031-13643-6_4

6. Fröbe, M., Bevendorff, J., Gienapp, L., Völske, M., Stein, B., Potthast, M., Hagen, M.: CopyCat: Near-Duplicates within and between the ClueWeb and the Common Crawl. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) 44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021). pp. 2398–2404. ACM (Jul 2021). https://doi.org/10.1145/3404835.3463246
7. Fröbe, M., Günther, S., Bondarenko, A., Huck, J., Hagen, M.: Using keyqueries to reduce misinformation in health-related search results. In: ROMCIR@ ECIR. pp. 1–10 (2022)
8. Fröbe, M., Schmidt, E.O., Hagen, M.: Efficient query obfuscation with keyqueries. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. pp. 154–161 (2021)
9. Keller, J., Breuer, T., Schaer, P.: Evaluation of temporal change in IR test collections. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024. pp. 3–13. ACM (2024). https://doi.org/10.1145/3664190.3672530, https://doi.org/10.1145/3664190.3672530
10. Keller, J., Breuer, T., Schaer, P.: Leveraging prior relevance signals in web search. In: Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S. (eds.) Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024. CEUR Workshop Proceedings, vol. 3740, pp. 2396–2406. CEUR-WS.org (2024)
11. Li, S., Lv, F., Jin, T., Li, G., Zheng, Y., Zhuang, T., Liu, Q., Zeng, X., Kwok, J.T., Ma, Q.: Query rewriting in taobao search. In: Hasan, M.A., Xiong, L. (eds.) Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022. pp. 3262–3271. ACM (2022). https://doi.org/10.1145/3511808.3557068, https://doi.org/10.1145/3511808.3557068
12. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. (Jan 1994)