Counterfactual Query Rewriting to Use Historical Relevance Feedback

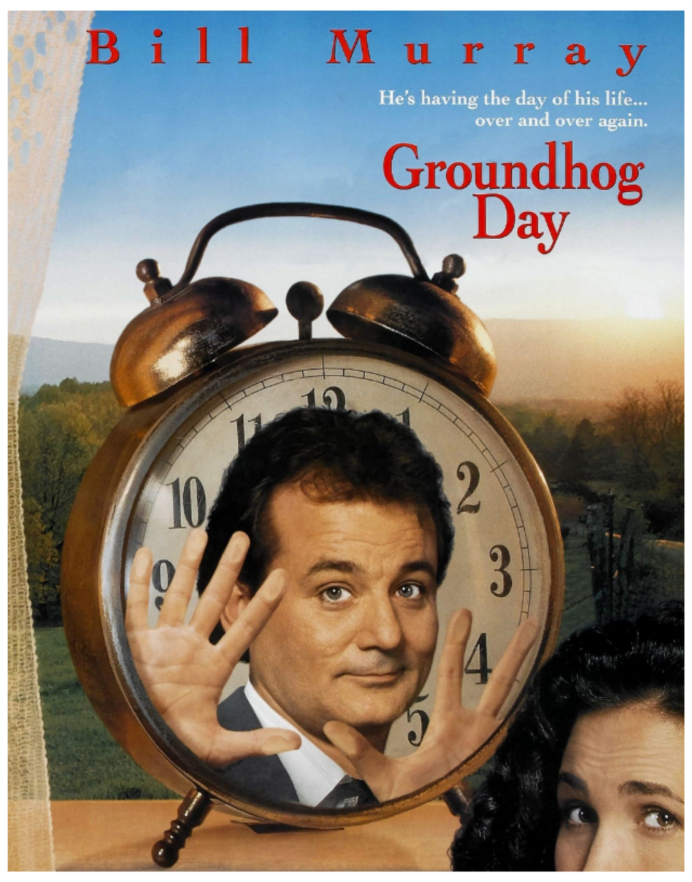
Jüri Keller, Maik Fröbe, Gijs Hendriksen, Daria Alexander, Martin Potthast, Matthias Hagen, Philipp Schaer

The same query… …over and over again.

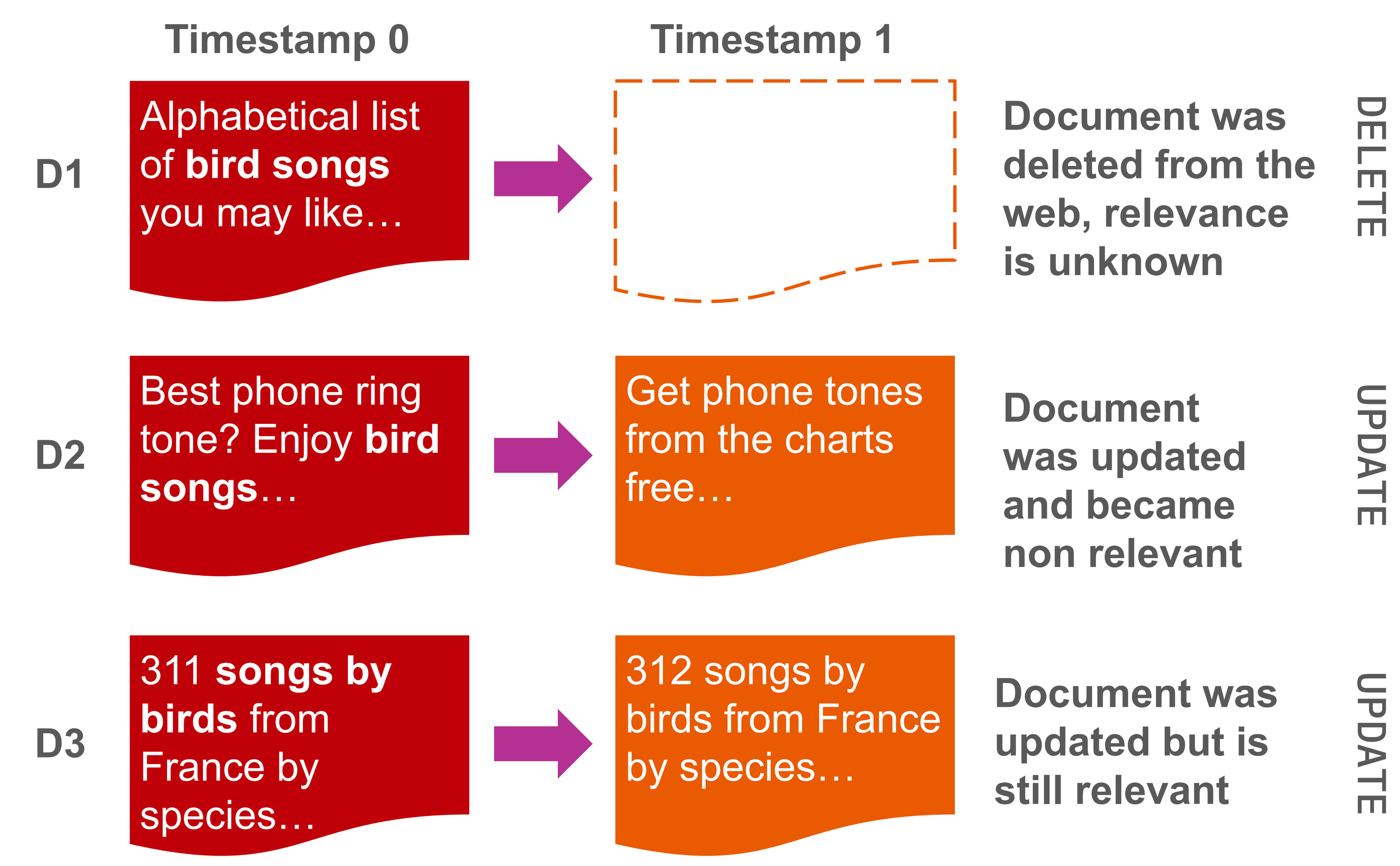
Many queries a search engine receives are not new!

How should this change the behaviour of the search engine?

- User interactions as relevance indicators
- We **counterfactually** assume that these pseudo-relevance labels are still valid, even if the documents **changed**
- These pseudo-relevance labels can be used to rewrite users' queries



Historic Relevance Feedback



1 Boosting

- Boost known documents based on their historic relevance feedback
- Can be repeated over time

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

- **Can not generalize** beyond known query-document pairs

2 Relevance Feedback

- Expand users' queries with terms from previously relevant docs
- Terms with top-k tf-idf scores
- Can be calculated offline as soon as relevance feedback exists
- RM3 as equivalent for new queries
- Generalizes to new and updated documents

3 Keyqueries

Perfect query for target docs

- Rewrite user query into a **keyquery**
- Based on the previously relevant documents as **target documents**

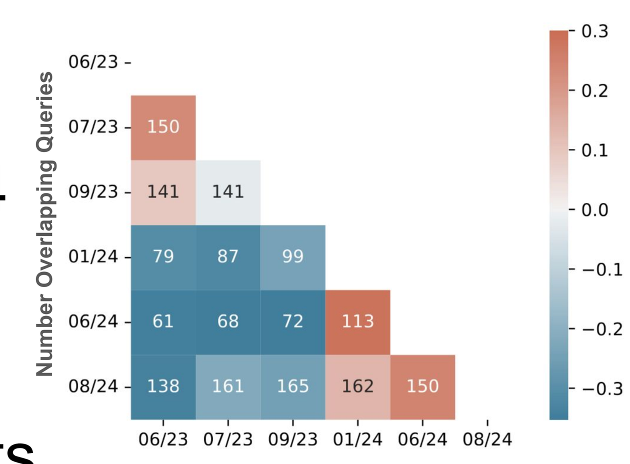
Query q is a **keyquery** for a set D of **target documents** against a search engine iff:

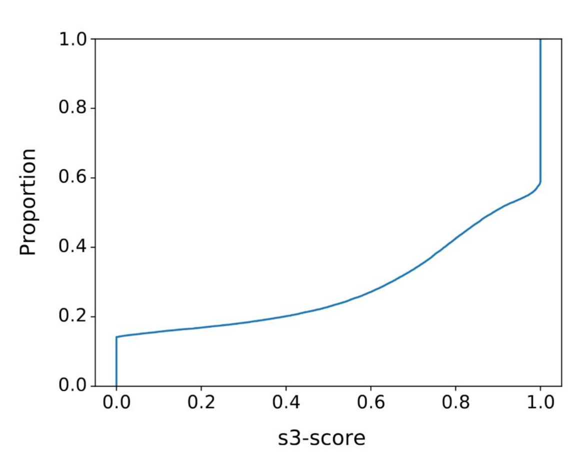
1. Every $d \in D$ is in the top-k results. (specificity)
2. Query q has at least l results. (generality)
3. No subquery $q' \subset q$ satisfies the above. (minimality)

- Prevents over and under fitting of the ranking to the target documents

Experiments

- Evaluated on six sub-collections between June 2023 and August 2024 of the LongEval Web collection
- Ablation study investigates how the systems generalize to new documents





Results

- ColBERT, List-in-T5, and monoT5 outperform the BM25 (+RM3) baselines
- Our three approaches substantially outperform all five baselines!
- Keyqueries perform the best and generalize well to new documents

System	nDCG@10 Normalized Discounted Cumulative Gain (nDCG)					nDCG@10′ nDCG on judged documents only				
	07/22	09/22	01/23	06/23	08/23	07/22	09/22	01/23	06/23	08/23
BM25	.155	.184	.172	.175	.134	.471	.492‡	.516‡	.486‡	.379‡
BM25_{RM3}	.147‡	.181	.163	.174	.134	.478‡	.490‡	.524‡	.492‡	.388‡
ColBERT	.198	.207	.201	.184	.151	.402†	.409†	.420†	.408†	.315†
List-in-T5	.203	.204	.202	.198	.161	.401†	.413†	.425†	.413†	.317†
monoT5	.202	.219	.197	.202	.154	.405	.410†	.415†	.411†	.314†
1 BM25_{Boost}	**.355**†‡	.372†‡	**.287**†‡	**.364**†‡	**.271**†‡	.529‡	.546‡	.541‡	.540‡	.412‡
2 BM25_{RF}	.303†‡	.332†‡	.241†	.262†‡	.191†‡	.606†‡	.611†‡	**.590**†‡	.552†‡	**.426**†‡
3 $\text{BM25}_{keyquery}$	.350†‡	**.391**†‡	.233	.262	.185	**.642**	**.655**‡	.574†	**.554**	.422‡

Conclusion

1. The advanced approaches **generalize** beyond known documents
2. Few feedback docs already substantially **improve** the retrieval effectiveness
3. Systems outperform expensive transformer-based models at a much **lower cost**

TH Köln - University of Applied Sciences, Cologne

Friedrich-Schiller-Universität Jena

Radboud Universiteit Nijmegen

University of Kassel, hessian.AI, ScaDS.AI

Webis