Toward Evaluating the Reproducibility of Information Retrieval Systems with Simulated Users

ACM Conference on Reproducibility and Replicability (ACM REP '24)
June 18–20, 2024, Rennes, France

Timo Breuer[1] **Maria Maistro**[2]

[1]TH Köln - University of Applied Sciences, Germany

[2]University of Copenhagen, Denmark

June 20, 2024

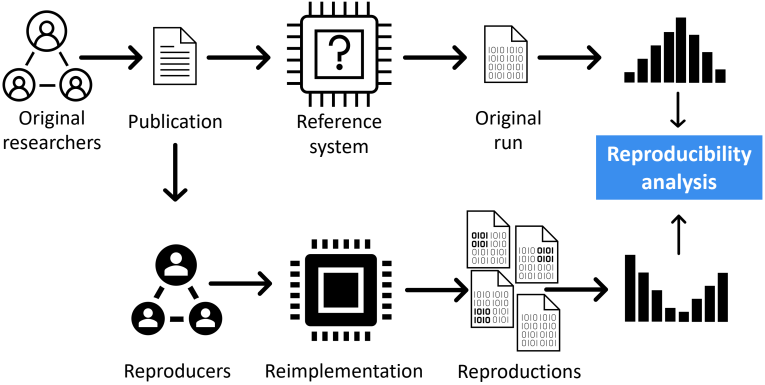
Motivation

Issue: Gap between system- and user-oriented evaluations of retrieval systems.

Question: When does a user consider a retrieval system as reproduced?

Solution: Bridge the gap with user simulations for large-scale evaluations!

Reactive reproducibility analysis

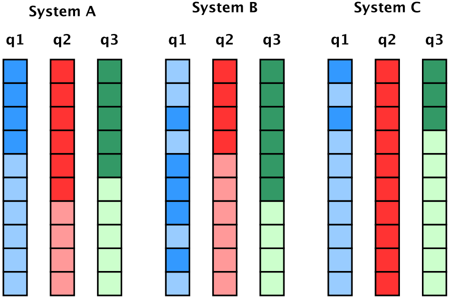


`repro_eval`: A Python Interface to Reproducibility Measures of System-Oriented IR Experiments, Breuer et al., ECIR'21

System-oriented information retrieval experiments

Experimental setup of the Cranfield paradigm [1]:

- Document collection
- Topics / queries
- Relevance labels

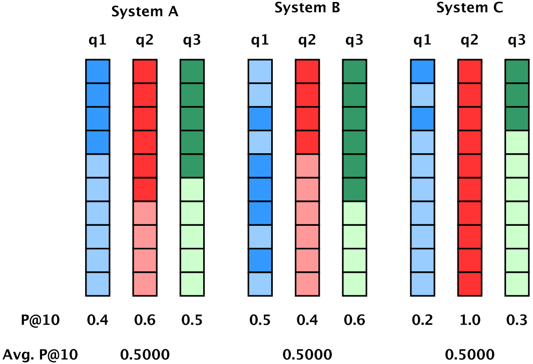


How to Measure the Reproducibility of System-oriented IR Experiments, Breuer et al., SIGIR'20
An in-depth investigation on the behavior of measures to quantify reproducibility, Maistro et al., IPM'23

Retrieval effectiveness

Precision

$$P = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$



How to Measure the Reproducibility of System-oriented IR Experiments, Breuer et al., SIGIR'20
An in-depth investigation on the behavior of measures to quantify reproducibility, Maistro et al., IPM'23

Retrieval effectiveness error

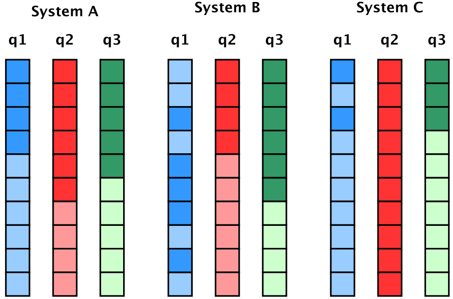
Root mean square error

$$\mathrm{RMSE}\left(M(r), M(r')\right) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(M_j(r) - M_j(r')\right)^2}$$

r, r' Original and reproduced runs

n Total number of queries

$M(r)$ Vector where each component is the score of an evaluation measure M with respect to the j-th query



P@10	0.4	0.6	0.5	0.5	0.4	0.6	0.2	1.0	0.3
Avg. P@10		0.5000			0.5000			0.5000	
RMSE		0.0000			0.1414			0.2828	

How to Measure the Reproducibility of System-oriented IR Experiments, Breuer et al., SIGIR'20
An in-depth investigation on the behavior of measures to quantify reproducibility, Maistro et al., IPM'23

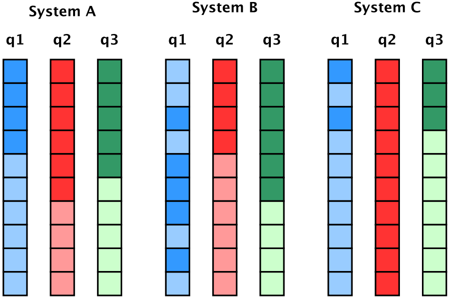
Rank correlation

Kendall's τ

$$\tau_j(r_j, r'_j) = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}}$$

$$\bar{\tau}(r, r') = \frac{1}{n}\sum_{j=1}^{n} \tau_j(r_j, r'_j)$$

r_j, r'_j	Original and reproduced rankings
n	Total number of queries
P, Q	Number of dis-/concordant pairs
U, V	Number of ties in r_j and r'_j



P@10	0.4	0.6	0.5	0.5	0.4	0.6	0.2	1.0	0.3
Avg. P@10	0.5000			0.5000			0.5000		
RMSE	0.0000			0.1414			0.2828		
KTU	1.000			0.4370			0.0815		

How to Measure the Reproducibility of System-oriented IR Experiments, Breuer et al., SIGIR'20
An in-depth investigation on the behavior of measures to quantify reproducibility, Maistro et al., IPM'23

Retrieval method

BM25 [2] is a strong and common baseline, also implemented in many industrial applications.

$$s(d, q) = \sum_{t \in q} \log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}}$$

d	Document $d \in D$
q	Query
t	Term contained in query q
N	Number of all documents in collection D
tf_{td}	Term frequency of term t in document d
df_t	Document frequency of term t in collection D
L_d	Length of the document d
L_{avg}	Average length of the documents in collection D
b	**controls the impact of document length normalization**
k_1	**controls the saturation point of term frequency normalization**

Experimental setup

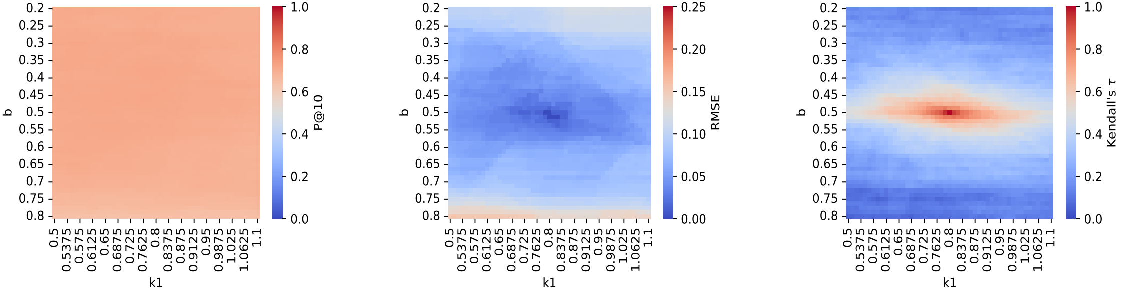
Simulating system reproductions:

- **Original system:** BM25 with $b = 0.5$ and $k_1 = 0.8$ as the reference system

- **Reproduced systems:** A total of 2,400 “reproduced” BM25-based systems with different b and k_1 parameters simulating reimplementations of the reference system

- **Dataset:** TREC-COVID [4] test collection comprising 191,175 documents, 50 topics, and 69,318 relevance judgments

Experimental results



(a) Retrieval effectiveness

(b) Retrieval effectiveness error

(c) Rank correlation

Figure: Heatmaps showing the average retrieval effectiveness in terms of P@10, the error between topic scores in terms of RMSE, and the correlation between the rankings in terms of Kendall's τ. Each patch is averaged over 50 queries given by the TREC-COVID collection and based on BM25 rankings with different b and k_1 parameters. RMSE and Kendall's τ are determined wrt. the center patch with $b = 0.5$ and $k_1 = 0.8$.

What is a successful reproduction?

👍 **Same retrieval effectiveness:**
On average, the retrieval effectiveness in terms of P@10 is the same.

👎 **Users would see different rankings as in the original experiment:**
What does a different ranking mean for the user in terms of reproducibility?

❓ **Users can compensate for changes or deteriorations in the rankings** [5]:
Is bit- or listwise reproducibility really a hard requirement?

User simulations

Real-life user studies are too costly and time-intensive:
Think of 2,401 systems x 50 queries = 120,050 experiments!

User simulations are a cost- and time-efficient solution:

- No participants need to be recruited,
- User behavior is controllable,
- Reproducible interactions with the system,
- No (cognitive) biases or learning effects,
- Simulate different devices (e.g., mobile vs. desktop).

Validation of query variations and click behavior:
Validating Simulations of User Query Variants, Breuer et al., ECIR'22
Validating Synthetic Usage Data in Living Lab Environments, Breuer et al., JDIQ'24

In a nutshell

- Always document and report (hyper-)parameters as detailed as possible!

- Reproducibility can be quantified at different levels, from different perspectives.

- User simulations for estimating the implications of reproduced real-life applications.

- Future work needs to evaluate the fidelity of the simulations and the user model.

Thank you!

Thank you for your attention.
Questions?

https://github.com/breuert/acmrep24

https://doi.org/10.5281/zenodo.10931438

References I

[1] D. Harman, *Information Retrieval Evaluation* (Synthesis Lectures on Information Concepts, Retrieval, and Services). Morgan & Claypool Publishers, 2011.

[2] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *TREC*, ser. NIST Special Publication, vol. 500-225, National Institute of Standards and Technology (NIST), 1994, pp. 109–126.

[3] M. G. Kendall, *Rank Correlation Methods*. Oxford, England: Griffin, 1948.

[4] E. M. Voorhees, T. Alam, S. Bedrick, *et al.*, "TREC-COVID: constructing a pandemic information retrieval test collection," *SIGIR Forum*, vol. 54, no. 1, 1:1–1:12, 2020.

[5] W. R. Hersh, A. Turpin, S. Price, *et al.*, "Do batch and user evaluation give the same results?" In *SIGIR*, ACM, 2000, pp. 17–24.