

Evaluation of Temporal Change in IR Test Collections

ICTIR 2024

Jüri Keller, Timo Breuer, Philipp Schaer

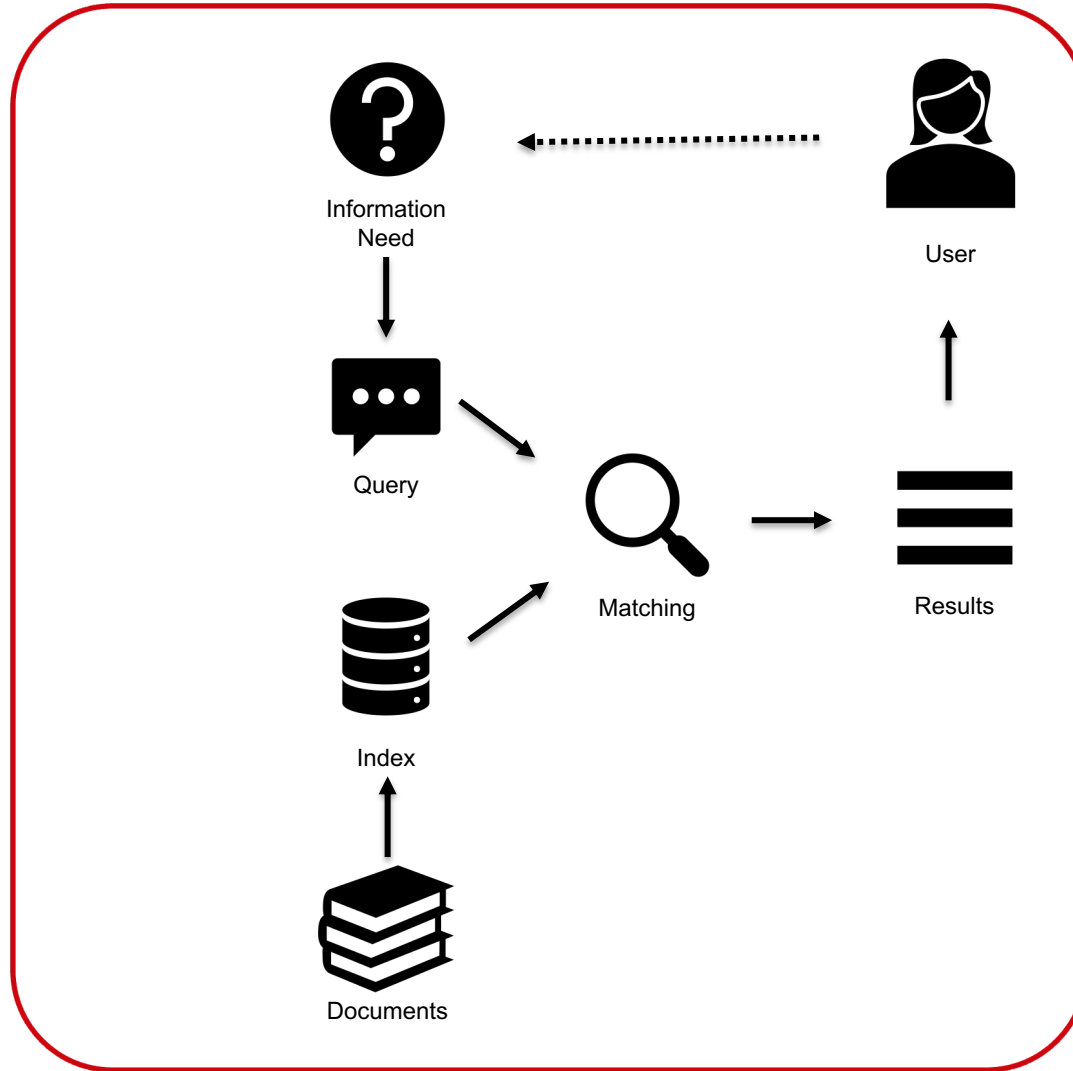
24-07-13 – Washington DC, USA
<https://ir.web.th-koeln.de>



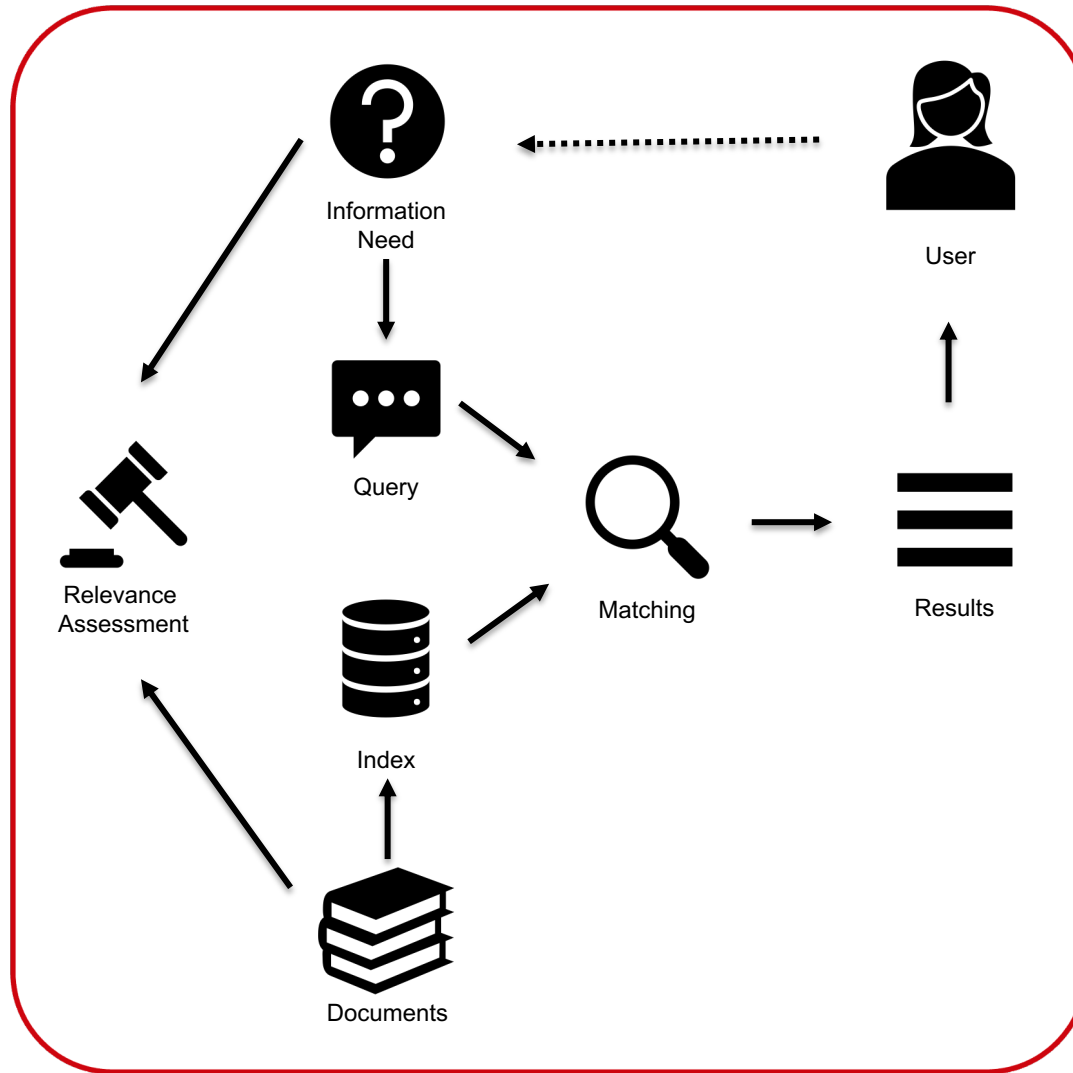
Technology
Arts Sciences
TH Köln



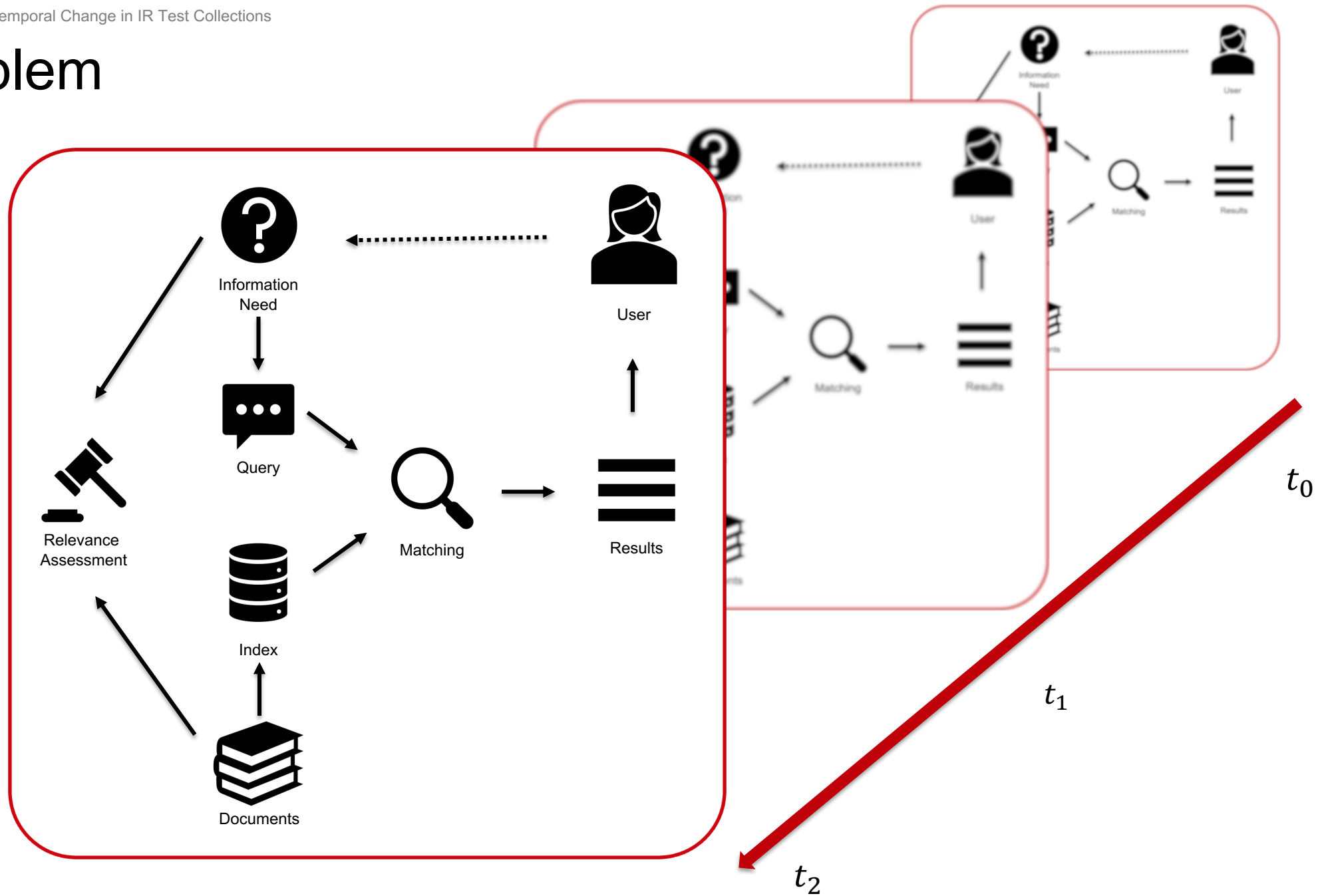
Problem



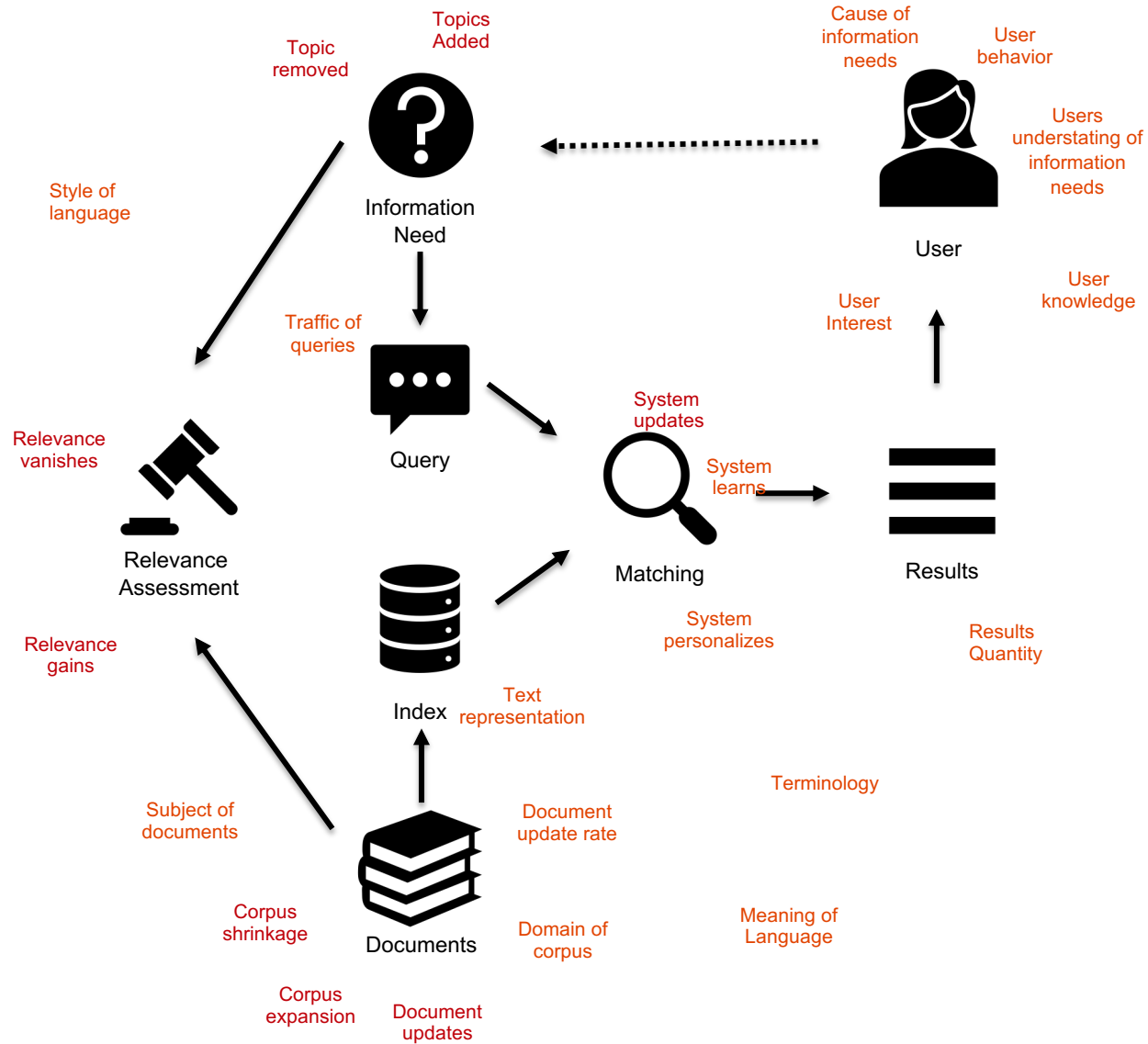
Problem



Problem

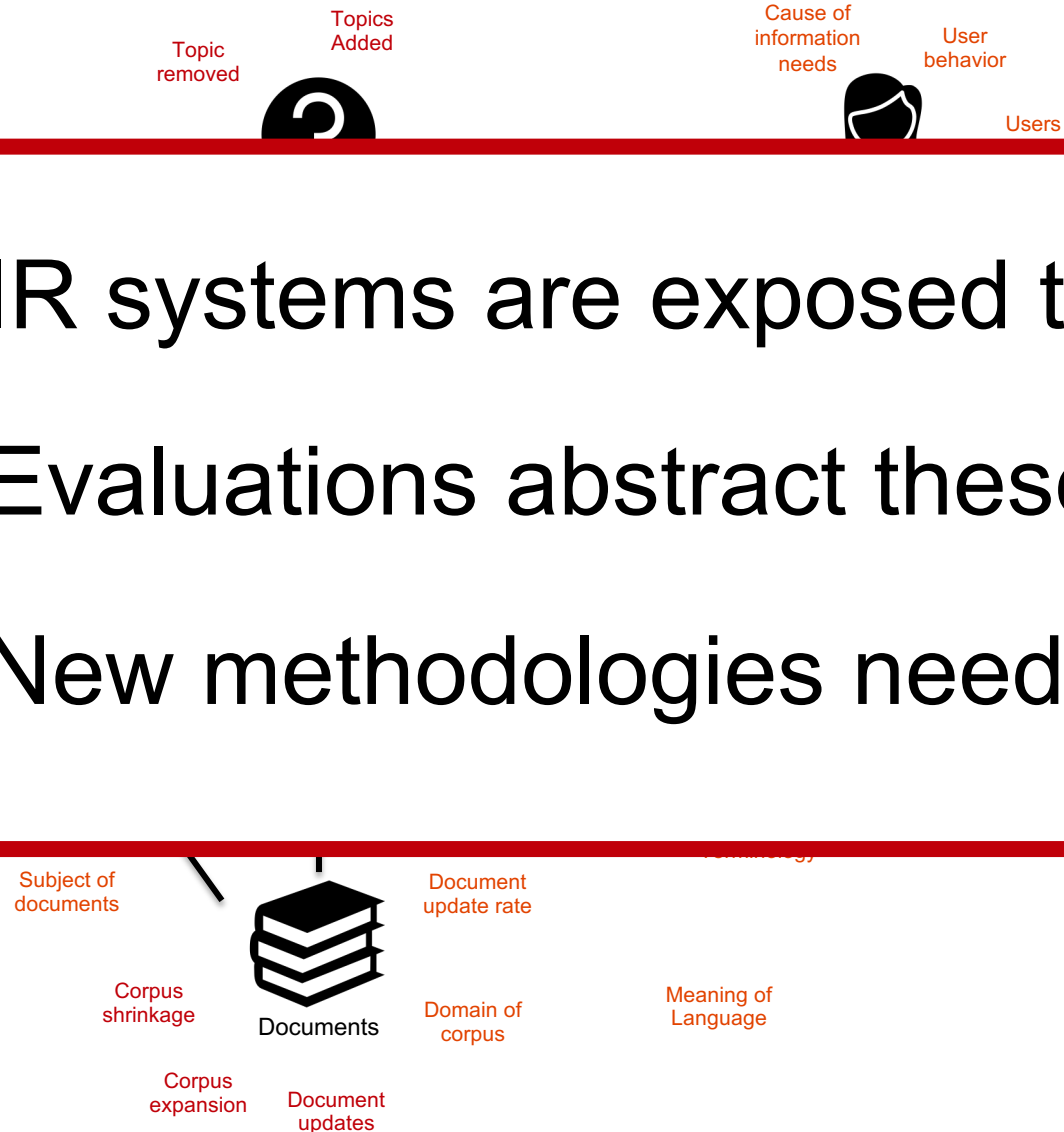


Problem



Problem

- IR systems are exposed to constant change
- Evaluations abstract these changes
- New methodologies needed to compare results



Goal

How can we quantify the impact of
changes in the evaluation setup
on the retrieval results?

Contributions

- C1: Definition of a classification schema to describe evolving retrieval scenarios
- C2: Propose measures to quantify change
- C3: Test methodology in an evaluation study
- C4: Discuss the results and challenges

Evaluation of Temporal Change in IR Test Collections

Jüri Keller
TH Köln
Cologne, Germany
jueri.keller@th-koeln.de

Timo Breuer
TH Köln
Cologne, Germany
timo.breuer@th-koeln.de

Philipp Schaefer
TH Köln
Cologne, Germany
philipp.schaefer@th-koeln.de

ABSTRACT

Information retrieval systems have been evaluated using the Cranfield paradigm for many years. This paradigm allows a systematic, fair, and reproducible evaluation of different retrieval methods in fixed experimental environments. However, real-world retrieval systems must cope with dynamic environments and temporal changes that affect the document collection, topical trends, and the individual user's perception of what is considered relevant. Yet, the temporal dimension in IR evaluations is still understudied.

To this end, this work investigates how the temporal generalizability of effectiveness evaluations can be assessed. As a conceptual model, we generalize Cranfield-type experiments to the temporal context by classifying the change in the essential components according to the create, update, and delete operations of persistent storage known from CRUD. From the different types of change different evaluation scenarios are derived and it is outlined what they imply. Based on these scenarios, renowned state-of-the-art retrieval systems are tested and it is investigated how the retrieval effectiveness changes on different levels of granularity.

We show that the proposed measures can be well adapted to describe the changes in the retrieval results. The experiments conducted confirm that the retrieval effectiveness strongly depends on the evaluation scenario investigated. We find that not only the average retrieval performance of single systems but also the relative system performance are strongly affected by the components that change and to what extent these components changed.

CCS CONCEPTS

• **Information systems** → **Retrieval effectiveness**; *Novelty in information retrieval*; Temporal data.

KEYWORDS

Longitudinal Evaluation, Continuous Evaluation, Reproducibility

ACM Reference Format:

Jüri Keller, Timo Breuer, and Philipp Schaefer. 2024. Evaluation of Temporal Change in IR Test Collections. In *Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '24)*, July 13, 2024, Washington, DC, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3664190.3672530>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICTIR '24, July 13, 2024, Washington, DC, USA
© 2024 Copyright held by the owner(s)/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0681-3/24/07
<https://doi.org/10.1145/3664190.3672530>

1 INTRODUCTION




Information Retrieval (IR) systems are exposed to constant change. The searched document collection evolves as new documents are added, removed, or updated [6, 22, 25], the users always encounter new information needs [18, 34, 45], and even the relevance is not static since information become outdated or opinions may change [13, 46]. In stark contrast, most IR experiments ignore the temporal dimension by only relying on snapshots or short time frames. By that, in test collection evaluations, all temporal changes are abstracted, and their influence on the effectiveness is minimized. Multiple sources suggest that IR experiments based on test collections are not temporally persistent [19, 25, 44]. Although there are some evolving dynamic test collections that span across more than one point in time, we identify the temporal dimension in IR evaluations as under-studied.

To investigate temporal dynamics in IR, we focus on the question: How can the impact of temporal changes in the evaluation setup on the retrieval results be quantified? Therefore, describing the changes in the retrieval setup and measuring their impact on the effectiveness are the primary concerns. We focus on test collection evaluations as a starting point and address changes in documents and relevance labels. We propose to classify the changes in the different components of Cranfield test collections by the CREATE, UPDATE and DELETE operation of persistent storage known as CRUD as high-level differentiation. Further, to investigate how changed effectiveness can be quantified different measures that are established in reproducibility evaluations are employed. To initially validate the proposed methodology, in the experimental evaluation, we repeatedly evaluate five state-of-the-art IR systems in controlled evolved experimental setups based on the three established test collections: TripClick [39], TREC-COVID [48], and LongEval [2]. These test collections cover a range of temporal changes as highlighted in Fig. 1. It is shown how the adapted measures describe how the initial effectiveness measured (at t_0) relates to the effectiveness measured at a later point in time (t_n). Different aspects of changing effectiveness, independent of relevance, on the topic level and with a focus on the system effect, are provided. This allows us to set the established systems into context so that new insights about them can be gained.

Since changes are unavoidable over time, we see great benefits in reintroducing temporal dynamics into test collection evaluations to learn about both, systems and test collections. Investigating temporal changes should help to improve the understanding of retrieval systems beyond their (relative) effectiveness by providing strategies to research how systems behave in specific situations. Investigating temporal changes can contribute to researching the reusability of test collections and emphasize the influence of the point of creation. Further, it can contribute to the field of test collection maintenance and to ensure reliably fair evaluations. Therefore,

Contributions

- C1: Definition of a classification schema to describe evolving retrieval scenarios
- C2: Propose measures to quantify change
- C3: Test methodology in an evaluation study
- C4: Discuss the results and challenges

	CREATE	UPDATE	DELETE
	Extension of document collection	Document content changed (e.g., online news articles, or websites)	Documents removed (e.g., due to licensing issues)
	New queries / topics (like current topics of interest)	Changed (head) queries from user logs (e.g., changed popularity)	Removed topics (due to missing interest or inappropriateness)
	Added new relevance labels (from old or new assessors)	Assessors changed their mind; new judgment guidelines	Relevance labels removed (due to low inter-rater agreement)

Contributions

- C1: Definition of a classification schema to describe evolving retrieval scenarios
- C2: Propose measures to quantify change
- C3: Test methodology in an evaluation study
- C4: Discuss the results and challenges

RBO

RMSE

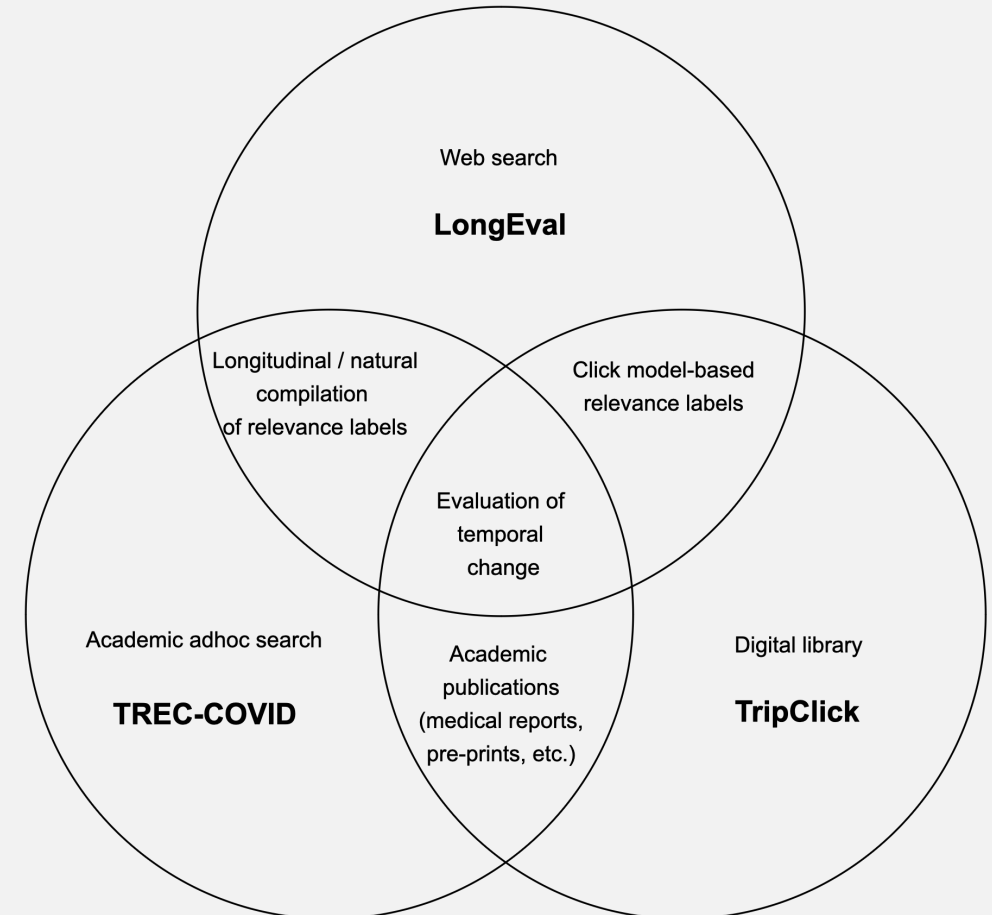
ΔRI

$\mathcal{R}_e\Delta$

ARP

Contributions

- C1: Definition of a classification schema to describe evolving retrieval scenarios
- C2: Propose measures to quantify change
- C3: Test methodology in an evaluation study
- C4: Discuss the results and challenges



Contributions

- C1: Definition of a classification schema to describe evolving retrieval scenarios
- C2: Propose measures to quantify change
- C3: Test methodology in an evaluation study
- C4: Discuss the results and challenges

Evolving Rankings

Comparing Effectiveness

Conclusions

Evolving Rankings

- Evolving test collection influence the rankings
 - Systems are static but rankings change
 - Similarity deteriorates over time
- Effectiveness fluctuates over time
 - Ranking of systems changes
 - No agreement between effectiveness and ranking similarity

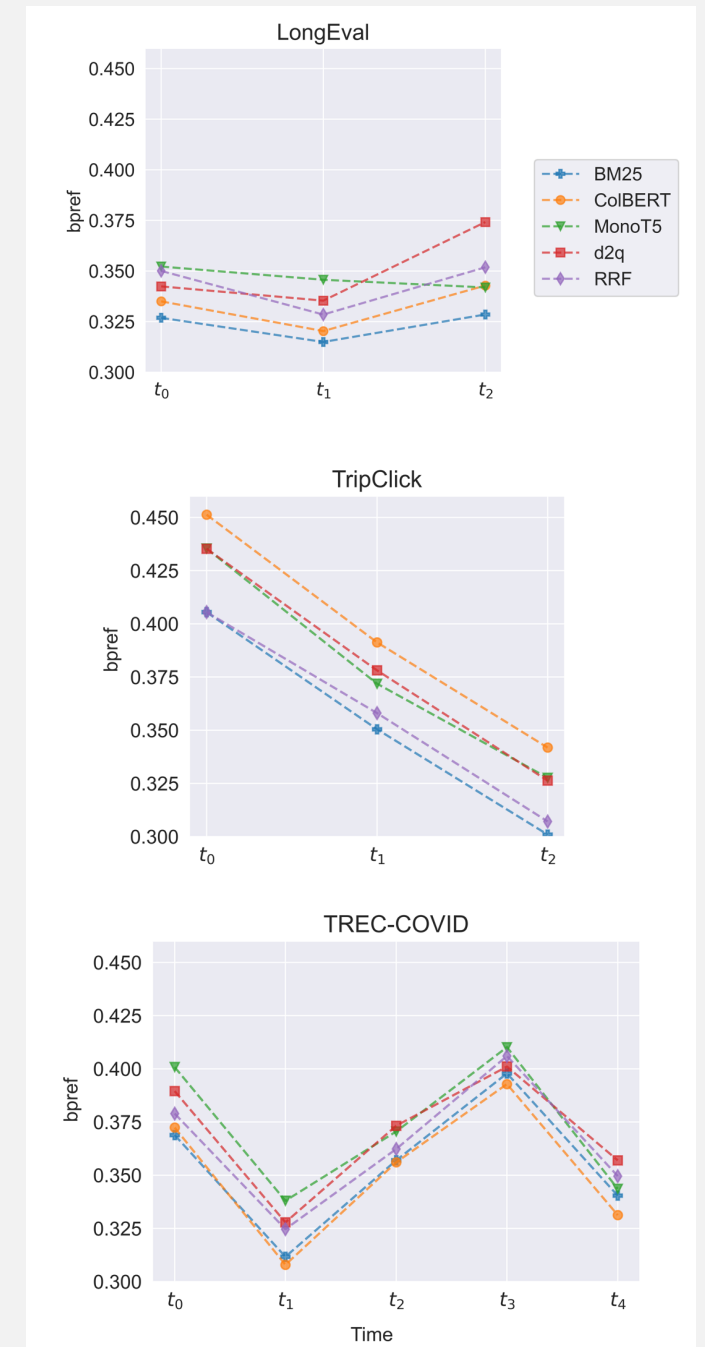
The measured effectiveness is not temporally reliable by default

TREC-COVID		RBO@100	
CREATE, UPDATE, DELETE	BM25	t_0	1
		t_1	0.761
		t_2	0.317
		t_3	0.207
		t_4	0.177
	ColBERT	t_0	1
		t_1	0.709
		t_2	0.235
		t_3	0.156
		t_4	0.136
	MonoT5	t_0	1
		t_1	0.761
		t_2	0.311
		t_3	0.190
		t_4	0.161
	RRF	t_0	1
		t_1	0.729
		t_2	0.309
		t_3	0.191
		t_4	0.139
d2q	t_0	1	
	t_1	0.502	
	t_2	0.128	
	t_3	0.083	
	t_4	0.069	

Evolving Rankings

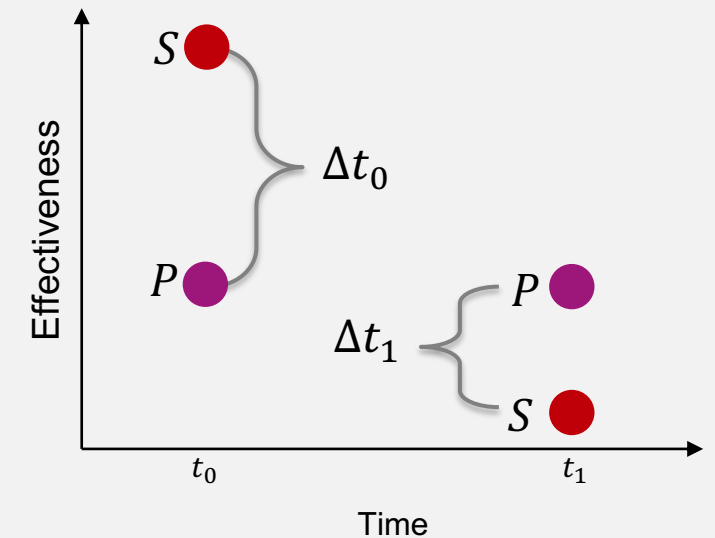
- Evolving test collection influence the rankings
 - Systems are static but rankings change
 - Similarity deteriorates over time
- Effectiveness fluctuates over time
 - Ranking of systems changes
 - No agreement between effectiveness and ranking similarity

The measured effectiveness is not temporally reliable by default

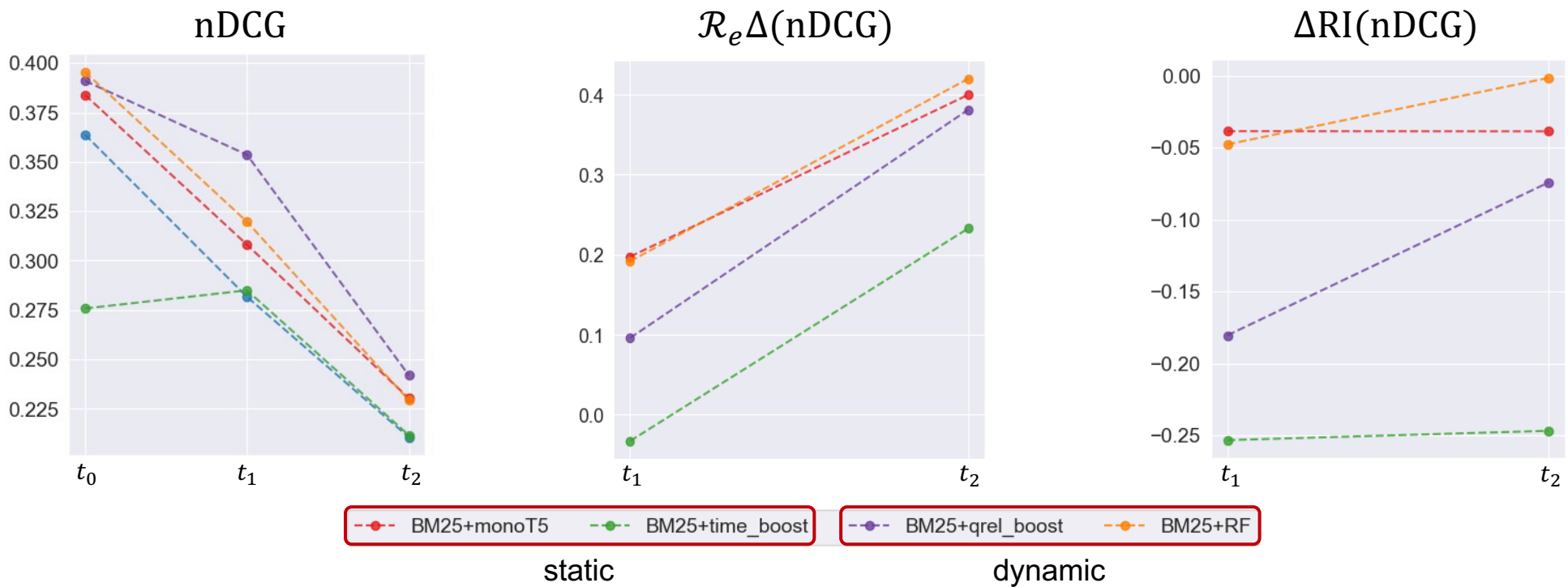


Comparing Effectiveness

- Direct comparison of effectiveness is not sufficient
 - Recall base changes
 - Mainly the environment effect is measured
- Advanced comparison strategy needed
 - Relate the **experimental system** to a **pivot system**
 - Compare the deltas



Comparing Effectiveness

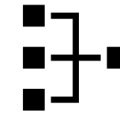


Pivot measures can compare results across time

Conclusions



Results depend on
the point in time



We need to account
for the temporal
dimension



Most effective
system is not the
most robust



Per topic differences
can amplify



How do users perceive
the changes?

Thank You!

