# Continuous Evaluation in Information Retrieval

Master thesis submitted for the degree:

*Master of Science* in the course of study Digital Sciences - Data and Information Science

at the faculty for Information Science and Communication Studies

of TH Köln - Cologne University of Applied Sciences


Presented by: Jüri Keller


Submitted to: Prof. Dr. Philipp Schaer
Second assessor: Dr. Timo Breuer

Cologne, 10.07.2023

# Abstract

As the information era progresses, the sheer volume of information calls for sophisticated retrieval systems. Evaluating them holds the key to ensuring the reliability and relevance of retrieved information. If evaluated with renowned methods, the measured quality is generally presumed to be dependable. That said, it is often forgotten that most evaluations are only snapshots in time and the reliability might be only valid for a short moment. Further, each evaluation method makes assumptions about the circumstances of a search and thereby has different characteristics. Achieving reliable evaluation is critical to retain the aspired quality of an IR system and maintain the confidence of the users. Therefore, we investigate how the evaluation environment (EE) evolves over time and how this might affect the effectiveness of retrieval systems. Further, attention is paid to the differences in the evaluation methods and how they work together in a continuous evaluation framework.

A literature review was conducted to investigate changing components which are then modeled in an extended EE. Exemplarily, the effect of document and qrel updates on the effectiveness of IR systems is investigated through reproducibility experiments in the LongEval shared task. As a result, 11 changing components together with initial measures to quantify how they change are identified, the temporal consistency of five IR systems could precisely be quantified through reproducibility and replicability measures and the findings were integrated into a continuous evaluation framework. Ultimately, this work contributes to more holistic evaluations in IR.

**Keywords:** Continuous Evaluation, Longitudinal Evaluation, Evaluation Environment

# Kurzfassung

Das fortschreitendende Informationszeitalter und die damit einhergehende Menge an Informationen erfordern fortschrittliche Retrieval-Systeme. Um sicherzustellen, dass diese relevante Ergebnisse finden und somit zuverlässig funktionieren, ist eine Evaluation dieser Systeme unerlässlich. Gängige Evaluationsmethoden gelten hierzu als verlässlich. Da sie aber oft nur auf Momentaufnahmen basieren, könnte ihre Geltungsdauer begrenzt sein. Zudem trifft jede Evaluationsmethode unterschiedliche Annahmen über die Umstände einer Suche und kann daher auch entsprechend nur bestimmte Aspekte eines Retrieval-Systems zuverlässig bemessen. Verlässliche Evaluationen sind aber entscheidend, um die angestrebte Qualität des Retrieval-Systems zu erhalten und das Vertrauen der Nutzenden zu bewahren. Um diesem Problem zu begegnen, untersucht diese Arbeit, wie sich die Evaluation Environment (EE) im Laufe der Zeit entwickelt und inwiefern sich diese Entwicklung auf die Effektivität von Retrieval-Systemen auswirken könnte. Darüber hinaus werden die verschiedenen Evaluationsmethoden sowie deren mögliche Kombinationen im Rahmen von Continuous Evaluation in den Blick genommen.

Durch eine umfassende Literaturrecherche wurden zunächst sich verändernde Komponenten identifiziert und die EE darauf basierend erweitert. Außerdem wurde mithilfe von Reproduzierbarkeitsexperimenten exemplarisch die Auswirkung von Dokument- und Qrel-Updates auf Retrieval-Systeme im Rahmen des Long-Eval Shared Task untersucht.

Hierbei konnten 11 sich verändernde Komponenten sowie erste Maße zur Quantifizierung ihrer Veränderungen identifiziert werden. Weitergehend wurde die zeitliche Stabilität von fünf Retrieval-Systemen durch Reproduzierbarkeits- und Replizierbarkeitsmaße präzise bemessen. Die Ergebnisse wurden abschließend in ein Continuous Evaluation Framework integriert. So leistet diese Arbeit einen Beitrag zur ganzheitlichen Evaluation im Information Retrieval.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

With a steadily increasing volume of information in the information era, differentiating between relevant and irrelevant information is an increasing challenge for Information Retrieval (IR) administers. IR finds application from large-scale web search to private search on individual level. Beyond ad-hoc search, IR systems are also foundational components in recommender systems and chatbots for question-answering. Increasingly sophisticated systems assist users to process information, not solely restricted to text but across all media types. Thereby, over the last decades, IR systems became more contextual and personal and increasingly depend on these factors for relevance ranking (Hofmann et al., 2016). Given that, it is not surprising that we interact with an IR system of any kind on a daily basis (Manning et al., 2008). These systems are often a significant influence on how we consume, how we are entertained or how we educate ourselves.

We naturally rely, sometimes almost blind on the IR system to retrieve the best results possible. To ensure the quality of an IR system, the effectiveness is evaluated in experiments. These experiments focus either on the absolute quality of a system, in terms of the utility for the users or on the relative quality in comparison to other IR systems (Balog & Zhai, 2023). Achieving reliable evaluations is critical, because false assumptions might be made, leading to misinvestments of time and money. Further, the choice of the system influences which information is presented and is accessible to users. This directly contributes to their understanding of the world. Ultimately, a flawed system may lead to situations in which critical information may be withheld.

Evaluations in IR are often based on short snapshots in time. This raises the question to what extent the results are generalizable for longer periods. Or, in other words, how temporal reliable the predominant evaluation methods actually are. To investigate that, evaluations based on multiple points in time need to be put in context. However, it is observed that many evaluation methods are not necessarily repeatable and a direct comparison is hardly possible (Balog & Zhai, 2023; Soboroff, 2006; Tan et al., 2017). To achieve temporaly reliable evaluations, it needs to be investigated how the evaluation environment changes over time and how that affects the IR system.

## 1.1 Research Questions and Methodology

Based on the described problems and the related work, the following research questions are formulated and answered in this work.

**RQ1** *How is the environment of an IR system evolving?* Many factors influence an IR system and potentially affect the evaluation. These factors are further assessed in the three sub-questions:

- **RQ1.1** *What are the evolving components of an IR environment?* The environment of an IR system comprises of various components of different kinds. To precisely differentiate and locate the effects, the individual components need to be identified.

**RQ1.2** *How do the components in an IR evaluation environment evolve?* The different components may evolve over time and do so differently. Identifying these changes is necessary to understand and distinguish them.

**RQ1.3** *How can this evolution be measured?* Measururing how the components evolve is paramount for comparing the changes and their influence on the effectiveness evaluation.

By answering these research questions, the EE can be specified better. This is a necessary prerequisite for factoring changes during continuous evaluations and ultimately achieving temporaly persistent results. Research question one is addressed in Section 3, where the three sub-questions are answered and in Section 4, where the EE is further specified.

**RQ2** *How can the evolving environment of an IR system be considered in practice during effectiveness evaluation?* To achieve temporally reliable effectiveness evaluations the environment of an IR system needs to be factored. Research question two is addressed in Section 5.

Methodologically, the research questions are answered by means of a systematic literature review and practical participation in the LongEval shared task (Alkhalifa et al., 2023). The literature review is founded on the methods of Silva and Neiva (2016). Based on initial literature and keywords, literature databases are searched for relevant publications. During the process, the search is refined and a review protocol is created. Further, inclusion criteria are determined. While the primary focus lies on continuous evaluation techniques in IR, adjacent diciplins are included.

The LongEval shared task is used as a testbed to test initial methods for evaluating temporal persistence. The first sub-task asks for systems that maintain constant performance over time. To evaluate this, the provided dataset contains related sub-collections from different points in time.

## 1.2 Contributions and Outline

To answer the raised research questions, a systematic literature review is conducted. 118 relevant publications are assessed and 11 components are derived, that undergo frequent changes and are part of the EE. The changes are identified and characterized and measures are gathered that quantify them. Based on the findings, the EE, initially proposed by Sáez, Goeuriot, et al. (2021), is generalized to the general IR problem and defined more explicitly. By that, it can guide future studies, that investigate the effect of different changes on the effectiveness of IR systems.

Further, we conduct initial investigations how changes in the EE influence the effectiveness of IR systems as part of the LongEval shared task. We adapt replicability and reproducibility measures to isolate document and qrel changes. Additionally, five state-of-the-art systems are submitted. To contextualize the results, an outlook is compiled in which synergies between evaluation methods are discussed.

In short, the main contributions of this work are:

- **Specification of the EE** based on changes observed in a systematic literature review,
- the collection of techniques and measures that **quantify changes in the EE**,
- the submission of **five state-of-the-art systems** to the LongEval shared task for longitudinal evaluation,
- the adaption of replicability and reproducibility experiments to measure **isolated influences** of the EE on the effectiveness of IR systems.

All resources created in this work, including the assessments of the publications from the literature review and the systems submitted to the LongEval shared task are made publicly available on Zenodo (Keller, 2023). Further, the open-source release of the experimental setup for the LongEval shared task can be found on GitHub.[1]

The structure of this work is oriented on the research questions. After this introduction, a short theoretical foundation is provided about information retrieval and evaluations in this field, by example of test collections in Section 2. In Section 3, the components and changes of the EE are described, based on the literature review. Section 4 is dedicated to defining the EE. Then, in Section 5, the temporal persistence is evaluated in practice through reproducibility and replicability measures, as a contribution to the Long Eval shared task. In Section 6, the findings are brought together as part of the continuous evaluation framework. Finally, this work concludes in Section 7 with attention to the initially formulated research question.

[1] https://github.com/irgroup/CLEF2023-LongEval-IRC

# 2 Theoretical Foundation

Continuous evaluation in the field of IR deals with the evaluation of IR systems, an essential task in IR. This work investigates respective components. To set a foundation, IR and the problem of matching Information needs with results is characterized. Further, a short overview of IR evaluations is given, using the example of static test collections and related measures.

## 2.1 Information Retrieval

Various sources defined IR mostly as a computational approach to answering questions. Manning et al. (2008) provide a good definition by describing IR as:

> "Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."

In other words, IR describes the process of providing relevant resources to user information needs. The information need of a user is verbalized as a query. On the contrary, the document corpus is represented through an index. A retrieval model functions as the connective link between those, by selecting the documents supposably relevant to the information need. The result is often represented as a ranked list of documents, ordered by their modeled relevance (Baeza-Yates & Ribeiro-Neto, 2011).

In this process, different key concepts are involved which are schematically visualized in Figure 1. On the user side, an *information need* is decisive for the search. It represents the gap between the user's current state of knowledge or understanding and the desired state of knowledge. Therefore, the information need may not yet be completely understood by the user, and can evolve during the search. The need arises, when a user requires specific information to fulfill a particular purpose or to answer a question (Baeza-Yates & Ribeiro-Neto, 2011).

The user expresses the information need in a *query*. While the information need can be understood as the general topic, the query is the search string that is passed to the system. It consists of the keywords or phrases, the user thinks are relevant for the search and reflects on the current understanding of the information need. During the search, the query might be reformulated (Manning et al., 2008). While previously mainly keywords were used, with improving retrieval models, often queries can be expressed as phrases in natural language.

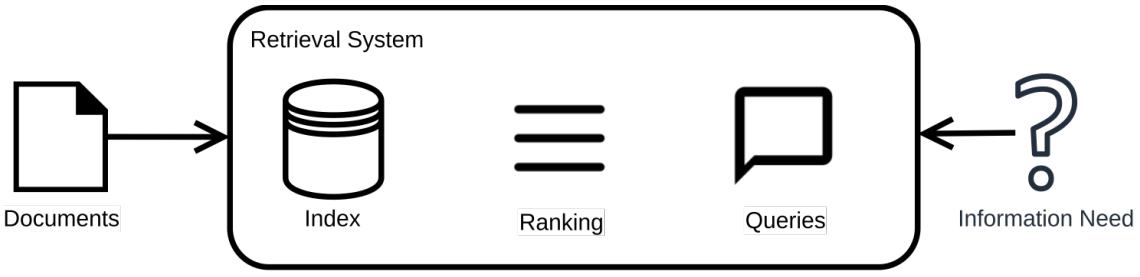


Figure 1: Schematic visualization of the IR problem.

Answers are supposably contained in the units of information which are called *documents* in this work. While often actual documents are searched, like scientific publications, documents can represent all kinds of data, for example, E-Mails, books or social media posts. Even beyond textual media, audio or video content can be represented as documents and searched. A group of documents forms the *corpus*, which is the entire body of information, available to be searched or analyzed.

The *index* is the data structure that represents the corpus for search. Since using whole documents is impractical, representations meaningful for the retrieval model are constructed. To create the index, the documents are analyzed and processed to increase both efficiency and effectiveness during search (Manning et al., 2008).

At the core, a retrieval model determines the relevance of a document to a given query. By determining the relevance for all documents in the corpus, the documents can be ranked and returned as results to fulfill the information need. Common retrieval models include the boolean model, vector space models like TF-IDF, and probabilistic models like BM25. Recent trends further employ language models or neural networks to the IR problem (Baeza-Yates & Ribeiro-Neto, 2011).

## 2.2 Evaluation in Information Retrieval

Evaluation of IR systems is a central and important task to quantify the capabilities of IR systems during development and beyond. Typically, IR systems are evaluated according to two dimensions: effectiveness and efficiency. While efficiency measures the system's resource consumption, effectiveness describes the quality of the system output. To assess the effectiveness of an IR system, the intent of the system, a measure of how well the intent is met, a measurement technique to measure that and an estimated measurement error is needed (Büttcher et al., 2010).

These components are provided by reusable test collections consisting of information needs and relevance judgments of documents and various evaluation measures. Due to the easy availability and cost-efficient re-usability after initial creation, test collection-based evaluation experiments are the de-facto standard in academic IR evaluations (Büttcher et al., 2010; Croft et al., 2009; Manning et al., 2008). Beyond that, further evaluation methods are employed, that provide a more user centered evaluation. User studies (Kelly, 2009) and online evaluations (Hofmann et al., 2016) evaluate the utility of an IR system for the users. While these methods provide great insights based on real users, they have difficulties to be repeatable. Therefore, simulations gain popularity as they also focus on the user but can be repeated (Balog & Zhai, 2023). While this work is not dedicated solely to test collection evaluations, they are exemplarily used to describe the principles of an IR evaluation in this section.

Voorhees (2019) provides a great overview of test collections following the traditional Cranfield paradigm (Cleverdon, 1997). Fixed information needs are captured in *topics* which contain the queries used for searching and often additional information that explains the information need further. The relevance of document query pairs is captured in *qrels* which are the ground truth for the evaluation. Given these components and an effectiveness measure, a system can be evaluated. Therefore, the top most relevant $k$ documents are

retrieved from a system for each topic and saved in a *run*. The effectiveness is then determined per topic or averaged over all topics, which yields an effectiveness score for a system, allowing to compare multiple systems in a ranking.

Test collections abstract the IR problem to a static level. That means that neither the document corpus, the test collection nor the relevance assessments change. Only by that, a direct comparison of systems can be done.

Since assessing every document in a corpus is not feasible due to the high assessment cost, mostly the corpus of a test collection is only partially assessed. The documents that are assessed are called pools and are mainly determined by runs from retrieval systems. By that, hypothetically, the most relevant documents to be judged should be found (Sanderson, 2010).

Different effectiveness measures exist to determine the performance of a system by the quality of the results. Assuming relevance is expressed as binary labels, *Precision* and *recall* are among the most popular ones. The precision is the number of relevant documents retrieved in relation to all retrieved documents. Intuitively, this measure describes the quality of the ranking. Therefore, let $R$ be the set of relevant documents and $A$ the set of documents that are retrieved by the system that is evaluated. Then the precision is defined by Baeza-Yates and Ribeiro-Neto (2011) as:

$$Precision = p = \frac{|R \cap A|}{|A|}. \tag{1}$$

Similarly, the *Recall* describes the completeness of the results, as the fraction of the relevant documents retrieved. Baeza-Yates and Ribeiro-Neto (2011) define it as:

$$Recall = r = \frac{|R \cap A|}{|R|}. \tag{2}$$

However, since both measures depends on the length of the result set $A$, the set is limited to a fixed length.

The Mean Average Precision (MAP) provides a single value measure. Intuitively, it describes the precision at every rank a relevant document is retrieved. Therefore, $R_j$ denotes the set of relevant documents for topic $j$ and $p(R_j[k])$ is the precision for the ranking, limited to the length until the $k$-th relevant document is observed in the retrieved result set. If the $k$-th document is not in the ranking, the precision is assumed to be 0. The MAP for a topic $i$ is then defined by Baeza-Yates and Ribeiro-Neto (2011) as:

$$MAP_i = \frac{1}{|R_i|} \sum_{k=1}^{|R_i|} p(R_i[k]) \tag{3}$$

and can be averaged over all topics:

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} MAP. \tag{4}$$

The sum of all MAPs for all topics is therefore divided by the total number of topics $|Q|$ in the test collection.

Since these measures assume every document in a test collection to be annotated, which is most often not the case, Buckley and Voorhees (2004) propose the *bpref* measure. The measurement determines the quality of a ranking by the number of judged and not relevant documents that are ranked higher than the first relevant one. It is defined as:

$$bpref = \frac{1}{|R|} \sum_{r} 1 - \frac{|n\,ranked\,higher\,than\,r|}{|R|}. \tag{5}$$

Instead of binary relevance labels, graded labels can differentiate between more and less relevant documents. The Discounted Cumulated Gain (DCG) factors the different grading levels and additionally incorporates the assumption that the first ranking positions are more important than the later ones.

$$DCG(D) = \sum_{d \in D, i=1} \frac{rel(j,d)}{log_2(i+1)} \tag{6}$$

$R(j,d)$ denotes the relevance label for the query $j$ and the document $d$ which is reduced by the discount factor in $log_2(i+1)$ based on the rank $i$.

The DCG is further normalized to the nDCG by relating it to the perfect DCG achievable per topic:

$$nDCG(Q) = \frac{1}{|Q|} \sum_{j \in Q} \frac{DCG(j)}{DCG(sorted(R_j))}. \tag{7}$$

Here, $R_j$ denotes all relevant documents for topic $j$ which are sorted to achieve a perfect ranking.

## 2.3 Summary

A short high-level overview of IR and the related challenges is given to set the foundation. The key concepts involved in the IR problem are characterized as the user with an information need on the one side and a collection, consisting of documents on the other. Through an IR system, the query and the index are linked to create results, often in the form of a ranked list. Various methods exist that try to estimate the effectiveness of IR systems. All of them make different assumptions about the search like equal or graded relevance. This makes them suitable to evaluate different aspects of the problem. Test collections are the de facto standard method in academia and abstract the evaluation environment so that it can be reused. Based on that, different measures quantify the effectiveness of IR systems.

# 3 Continuity Over Time

The results of conventional IR evaluations are assumed to be generalizable over time. If this assumption holds, an IR system would continuously achieve the measured effectiveness if the evaluation is repeated over time. However, this temporal dimension is rarely considered during evaluations at all. In contrast, the environment of an IR system, e.g. the components involved in the IR problem, can change. With increasingly contextual IR systems, the generalizability may decrease (Hofmann et al., 2016). To investigate if and how IR systems are affected by such changes, it is an important prerequisite to gain a deeper understanding of the evaluation environment. Therefore, in this chapter the research question *How is the environment of an IR system evolving?* is investigated. The question is further subdivided into the three questions:

**RQ1.1** *What are the evolving components of an IR environment?*

**RQ1.2** *How do the components in an IR evaluation environment evolve?*

**RQ1.3** *How can this evolution be measured?*

Methodologically, an extensive literature review is conducted to survey the previous work. Over 2000 articles from four relevant databases are examined which yield in total 118 relevant publications for review. The relevant publications are first quantitatively analyzed and then qualitatively set into context.

First, we describe the compiled review protocol, which specifies the methodology in 3.1. Then we report the results of the review, initially as a meta-analysis in Section 3.2 and further quantitatively in Section 3.3. The results are analyzed, ordered and contextualized qualitatively in Section 3.4. We summarize the findings in Section 3.5 and provide an online results table.[2]

## 3.1 Review Protocol

The goal of the review is to examine the general concern that the temporal reliability of cumbersome and expensive IR evaluations may be limited. By surveying existing work, an extensive overview of the different facets of the problem is gathered, to discriminate between beliefs and facts. By that, a starting point for addressing the problem is given in a structured way.

Beforehand, the ACM Computing Reviews journal[3] was searched with the simple keyword *Information Retrieval* and the resulting 194 reviews were examined. One of the initial publications by Tonon et al. (2015) was re-discovered. Also, the temporal dimension as query elements in the field of Temporal Information Retrieval (TIR) was found (Kim et al., 2013). Other than that, no relevant reviews could be found, which strengthens the need for this literature review.

Methodologically, Silva and Neiva (2016) propose a guideline for systematic literature reviews in the domain of computer science. They characterize the process in twelve steps. This is mostly in line with the system Kitchenham and Charters (2007) proposed for the

[2] https://th-koeln.sciebo.de/s/A5t4ATMWzOTOZ73
[3] https://libraries.acm.org/digital-library/acm-computing-reviews

domain of software engineering. They describe the process in greater detail and separate it into three general phases with multiple sub-steps. Rather than describing the process as a linear list of tasks, it is characterized as an iterative process, going back and forth to refine parts of it after gaining a better understanding of the problem. The three main phases are planning, conducting and reporting the review. Based on these, a review protocol is created comprised of research questions, exclusion criteria and quality assessments.

The research questions were formulated in a structured way to clearly identify the different elements they ask for. Therefore, the PICOC (Population, Intervention, Comparison, Outcome and Context) elements were identified from them (Petticrew & Roberts, 2006). While these originate from medical research, Kitchenham and Charters (2007) transferred them to the domain of software engineering, making them applicable to research in computer science. The population originally asks for the Who in a research question. This is transferred to the area or parts that are affected. The intervention asks for What or how a change is introduced. This maps to the method applied in contrast to the Comparison system. The Outcome describes the goal that should be accomplished. The last element is the Context, describing the circumstances of a research question. Together, these describe the domain the question focuses on.

In the described research questions the PICOC elements are mainly the same, except for the different intervention methods asked for. The population consists of the IR systems that are investigated.

The first research question *What are the evolving components of an IR environment?* asks for the concrete components involved during evaluation that change over time. Answers to this question need to identify concrete components. The PICOC elements are:

***Population:*** The components that change.

***Intervention:*** The time that affects the components.

***Comparison:*** The components in a conventional IR evaluation.

***Outcome:*** A list of components that can change.

***Context:*** IR systems in general.

The second research question *How do the components in an IR evaluation environment evolve?* asks for the change that is observed in distinct components of the EE. Answers to this question need to identify the change of the components. The PICOC elements are:

***Population:*** The components that change.

***Intervention:*** The time that affects the components.

***Comparison:*** The components in a conventional IR evaluation.

***Outcome:*** A list of components that can change.

***Context:*** IR systems in general.

The third research question *How can this evolution be measured?* asks for the measures and methods used to identify changes in the EE that may influence the effectiveness of IR systems. Preliminary research showed not much attention is given to this research area

in IR, therefore the question is not only focussed on this area but includes also adjacent disciplines like classification for example. The PICOC elements are:

***Population:*** The changing components and how they change.

***Intervention:*** The measures that quantify the change.

***Comparison:*** Conventional evaluations.

***Outcome:*** A list of measures and methods.

***Context:*** General IR systems and similar models from other disciplines.

### 3.1.1 Collecting Data

After clarifying the goal of the review through the research questions, the keywords for the systematic search are gathered. They are determined from the research questions, the PICOC elements and a preliminary unstructured search. The research area appears to be dynamic and not clearly defined. Therefore no fixed terminology is established yet. To capture the broad topics, comparably many keywords and synonyms were used for the search, also including generic ones like `change` or `temporal`. All keywords are listed in Table 3.1.1 and the full search strings for the different databases can be found in the appendix Section 7.2.

| **Query terms from questions:** |
|---|
| `evolving environment, temporal, temporal persistence, measure, change, changing artifacts, changing parts, changing components` |
| **Query terms from preliminary search:** |
| `continuous evaluation, temporal shift, longitudinal evaluation, dynamic test collection, evolving test collection, evolving dataset, temporal generalisability, temporal decay, temporal evolution, evolution, delta, evaluation environment, time-evolving` |
| **Scoping terms:** |
| `Information Retrieval, IR` |

Table 1: shows the keywords that were used to search the literature databases. They were extracted from the research questions and gathered during the preliminary research.

These keywords are used to search the four literature databases *IEEE Xplore*[4], *ACM-DL*[5], *Semantic Scholar*[6] and *DBLP*[7], which are often used for research in the computer science domain. Further, the database *Scopus*[8] was considered but was not used in the end, due to unsatisfactory results. The databases *IEEE Xplore* and *ACM-DL* support advanced queries with boolean operators like `AND` and `OR`. This allows concatenating the keywords

[4] www.ieeexplore.com
[5] https://dl.acm.org/
[6] www.semanticscholar.org
[7] www.dblp.org
[8] www.scopus.com

and additionally adding the scoping keywords `Information Retrieval` and `IR` so that any keyword must occur in conjunction with a scoping keyword. The search considered the title, abstract and keywords.

The databases *Semantic Scholar* and *DBLP* only support keyword searches. Therefore, multiple searches were conducted by combining the keywords individually with `Information Retrieval`. These searches yield more results, with potentially many more irrelevant articles. To restrict the results found, only the top 100 articles per search were considered further. While *Semantic Scholar* conducts a semantic search, the *DBLP* only allows keyword search over the titles of publications.

### 3.1.2 First Screening Stage

The reviewing process comprises three stages. In the first screening stage, the titles and if available the abstracts of the publications are assessed for general relevance concerning the research questions of the temporal review. During this stage, the inclusion and exclusion criteria are created.

The inclusion criteria are that a study needs to be in English or German and should be a primary study. Also, a clear relevance to the field of Information Retrieval or at least an adjacent discipline like classification or clustering in the context of computer science is needed for inclusion. The method needs to focus on textual data or be applicable.

Exclusion criteria are studies that propose systems that use temporal aspects of the retrieved media only as features. For example, systems that promote recent documents, extract latent temporality from queries or exploit the temporal aspects of sequential media like video and audio. These include most TIR and Temporal Information Extraction (TIE) studies as well as many time series analysis studies. Retrieval from databases, for example in the context of schema evolution, are excluded as well. Further, if the change occurs mainly on the system side, in algorithms or methods, the study is excluded if not explicitly evaluated for IR. This excludes evolutionary algorithms or evolving neural networks.

### 3.1.3 Second Screening Stage

These inclusion and exclusion criteria are applied in the second screening phase. This time, the introductions and conclusions are assessed. Each publication is graded on a scale of zero to three. Studies with a zero assigned are considered irrelevant and are excluded. The remaining labels are: 1 partially relevant, 2 relevant and 3 highly relevant. The lowest rating zero is assigned to all studies that do not fit the pre-defined inclusion criteria or are simply not retrievable as full text (or slide/recording considering presentations and talks). For an error analysis, the reason for the rejection is noted. Partially relevant publications should contain some relevant information and contribute valuable answers to one research question. Also, they might contain relevant information but the overall focus is not on the IR or the core topic. An example of a partially relevant publication focus on the change of language by analyzing an archive. Relevant publications contribute to more than one research question. Highly relevant publications contribute to multiple research questions of a high quality or focus specifically on the core topic.

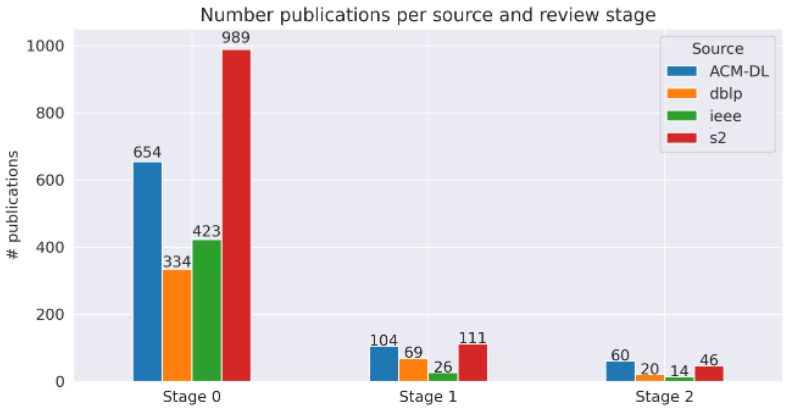


Figure 2: Over the two annotation stages, more documents were found not relevant.

### 3.1.4 Analysis Stage

For the final analysis, during this review stage, the relevant publications are initially systematically captured by gathering their findings in relation to the research questions. Therefore, the following categories are defined:

- Domain
- Timeframe
- Changing components (Q1.1)
- Observed change in the components (Q1.2)
- Measures used to quantify change (Q1.3)

The first two categories *Domain* and *Timeframe* collect metadata about the studies. The further categories cover the main aspects of the research questions and are the foundation for the quantitative analysis.

Finally, in the last review stage, the relevant publications are assessed in more detail for the final analysis. Thereby the focus lies mainly on highly relevant publications. During this process, studies are often reassessed, categories are evolved and findings are re-evaluated. For the final analysis, the results are first summarized quantitatively and then qualitatively described.

## 3.2 Meta-Analysis

Through the initial search, 2448 publications are found. Due to the keyword-based search in *DBLP* and *Semantic Scholar* the results contain duplicates in these sources. Further duplicates are found across sources and were removed. This yields in total 1943 unique publications for evaluation in the first stage. After this stage, where titles and abstracts are assessed, only 246 publications are left. The yield of less than 10 % can be explained through the relatively generic keywords used for the search as well as the long result lists from *DBLP* and *Semantic Scholar*. The main errors are publications not related to IR from the field of computer science, studies not focussing on effectiveness and publication

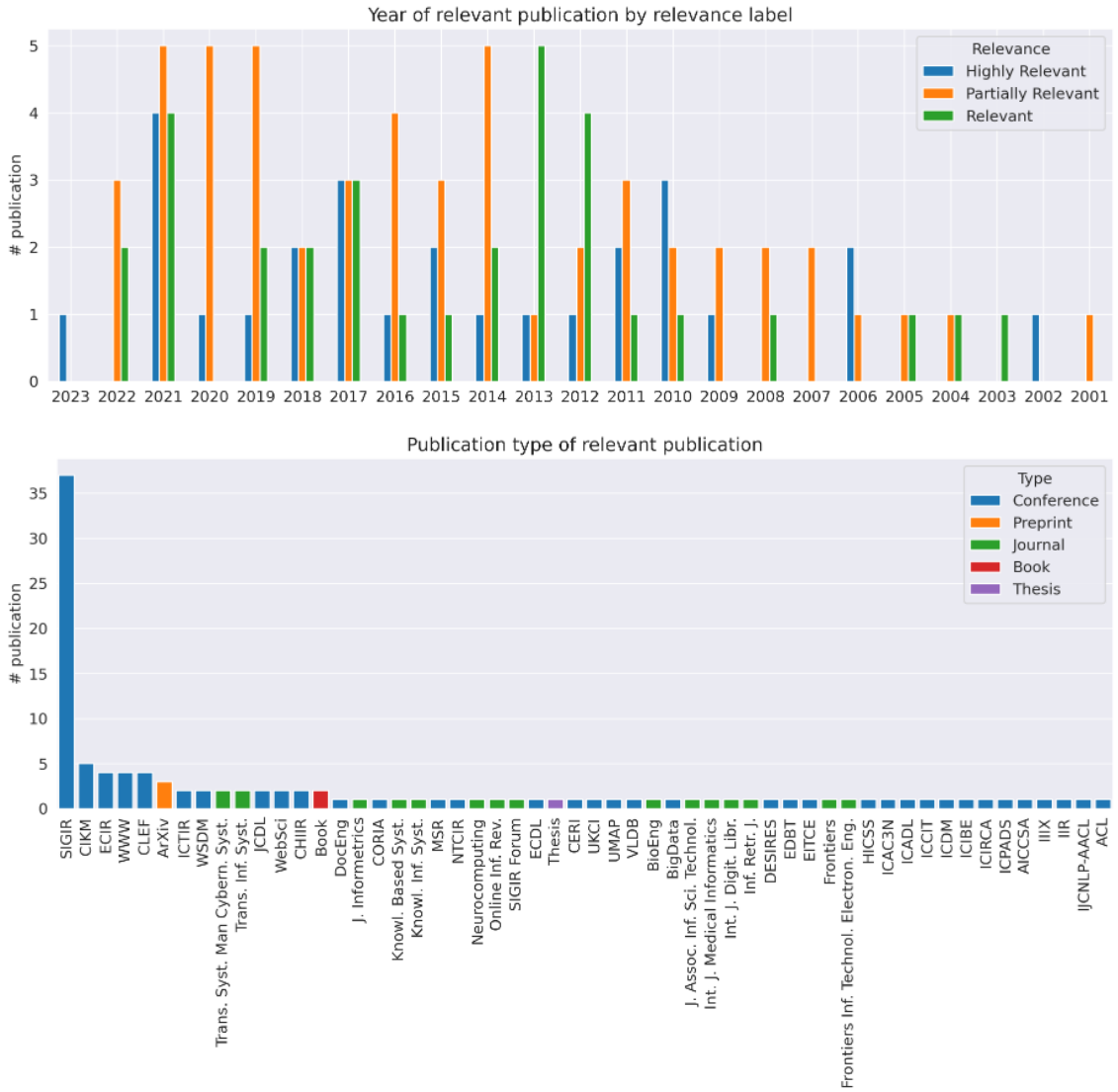


Figure 3: Relevant publication, separated into partially relevant, relevant and highly relevant over time (top). The sources of relevant publication form a long tail distribution (bottom).

from the fields TIE or TIR, which exploit temporal aspects for improved effectiveness but do not consider change or consistency over time at all.

In the second evaluation stage, the introduction and conclusion sections are assessed. In total, 128 further publications are considered not relevant, leaving 118 relevant publications. Among the three quality classes, 27 publications are found highly relevant, 32 relevant and 53 partially relevant.

Figure 2 shows how the different relevant publications distribute across the four databases they were retrieved from, separated by evaluation stages. Since often the same publications were received from multiple sources, these numbers do not correspond to the absolute ones reported before. Instead, an overview of the different literature databases is given. The ACM-DL yields the most relevant publications relative to the total, followed by dblp.

The earliest publications assessed were published in 1962. The publications considered at least relevant are published over the last 22 years, beginning in 2001. Without any gaps, relevant publications are found in every year. A trend towards the recent years can

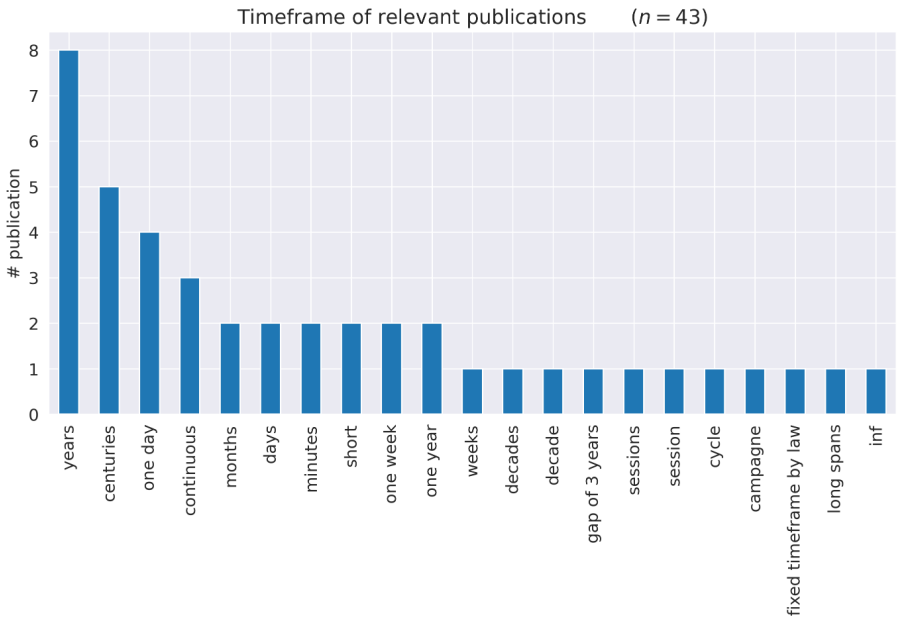


Figure 4: The timeframe mentioned in relevant publications.

be observed in Figure 3. However, considering the graded relevance criteria, no period of time stands out.

Considering the type of publication, conferences are the main source, followed by some informal publications from ArXiv and occasionally journals. Figure 3 shows how relevant publications distribute across the publication form. It is a strong Pareto distribution, by far the most relevant publications are published in the SIGIR conference, followed by the conferences CIKM and ECIR. The top venues are also the ones considered most relevant before the analysis. The really strong Pareto distribution with over 35 SIGIR papers in first place compared to five CICKM papers in second place, shows the strength of the effect. While a full assessment was important for completeness, restricting the search to the common venues yields the most relevant publications and reduces the effort significantly.

### 3.3 Quantitative Analysis

First, the observed timeframes and the domains are described. Further, the different fields regarding the measurements, changes and systems are reported.

While not particularly in the focus of the research questions, the different timeframes that are considered by the studies show the diversity and granularity change is investigated on. The timeframe that is investigated by different studies is heterogeneous, as shown by the 21 different timeframes found. In total, only 43 studies explicitly describe the timeframe and most timeframes are only investigated by one study. Few studies investigate multiple timeframes or compare results between timeframes. With eight occurrences, studies investigate change over the course of a year, followed by whole centuries and one day as the timeframe. The distribution of timeframes is visualized in Figure 4.

The distribution of the different domains is skewed even stronger. In total, 11 different domains are found in 43 studies. Most prominent, with 16 occurrences is the *web* domain, followed by *news* with 8 occurrences. The long tail contains *Wikipedia*, *literature*, *medical*

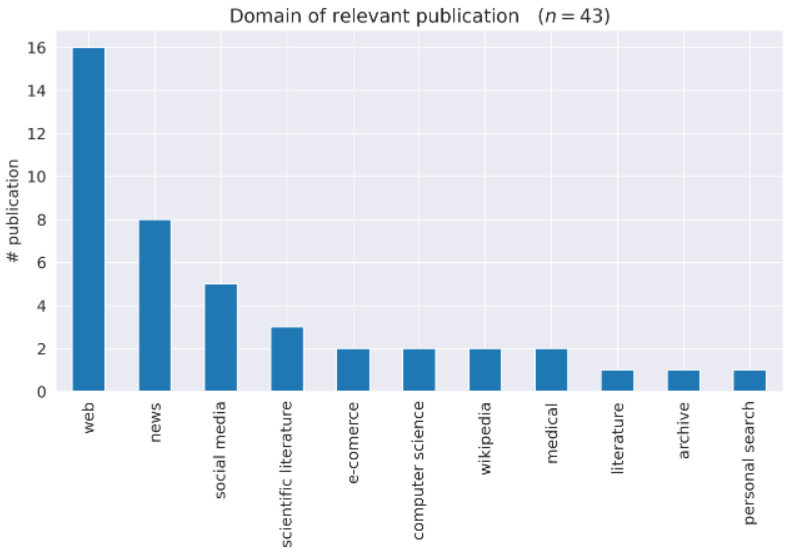


Figure 5: Domains mentioned in relevant publications.

and others. Most studies focus on a single domain only, while fewer test or compare in multiple domains such as *e-commerce* and *news* or *microblog* and *medical*.

In total, 28 measures are found that are intended to report on changes or consistency of any kind over time. These measures are found in 22 studies, with as many as four measures named. The overlap is small, only four studies use the same measures.

The observed components that change are highly skewed. In total 31 different components from 154 studies were found. By far the most investigated component is the document on an individual level, followed by the corpus of documents as the entire dataset that is searched, and the language. 14 components are the focus of only a single study. Changes on the document are investigated by 29 studies.

A system architecture could be obtained from 26 studies only. With five occurrences, TF-IDF is the architecture most often mentioned, followed by Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) models. In total, 14 different architecture categories are observed. Most studies report an improved performance compared to a baseline.

## 3.4 Qualitative Analysis

After surveying the different components quantitatively, they are analyzed qualitatively to answer the research questions. It quickly becomes apparent that the observations can be further grouped based on their characteristics. For example, most measures quantify differences and some focus on similarities or on how many changing components are related to the components in an IR test collection. Further, by contrasting the categories and observations, similarities and differences can be observed, which are described in the following.

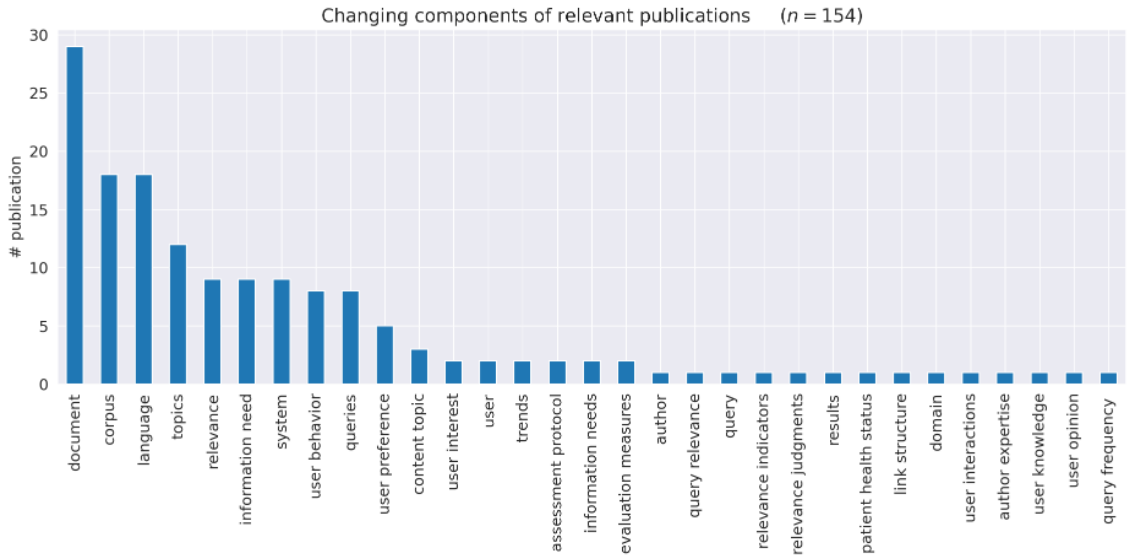


Figure 6: Changing components mentioned in relevant publications.

### 3.4.1 Timeframe

Despite time appearing inherently linear and continuous, this is most often impractical for investigations and instead, discrete timestamps are considered.

The size of the different timeframes draws attention to vastly different components. The longest spans observed are multiple centuries long and the shortest spans are investigating timeframes shorter than a minute. Since the granularity vastly differs, a comparison is often impossible. Longer intervals are primarily chosen to investigate changes in the language, in the corpus or on the document level. Shorter intervals are used for investigations more oriented towards the user, as for example for the user behavior or queries. Besides these correlations, many exceptions occur and in general, the chosen timeframe is defined by the component investigated, and on the longer end by the resources available.

Although time is considered continuous, for an investigation, measurements are made often at specified points, like daily (Forman, 2006) or in contrast as a comparison with multiple years inbetween (Altingövde et al., 2011). These points in time can be understood as the resolution of the investigation correspond to the length of the investigation dictated by the use case.

In contrast to timeframes that can be measured as time, sometimes the timeframe is understood as a repeating cycle of actions. Here, the focus lies on the recurring actions, like seasonally recurring queries (Mansouri et al., 2017). Instead of a fixed interval, the timeframe might also be variable.

### 3.4.2 Domain

In general, the analyzed studies show a diverse field of domains where change and consistency across time are investigated. These domains range from specific, like *computer science questions* to generic like *English literature*. Most often, changes are observed in the *web* domain, followed by *news*. These domains are known to be fast-changing and there-

fore well suited to investigate change. Further, the domains correlate with the landscape of available datasets in IR.

### 3.4.3 Components

In total, 11 changing components were identified and are described in the following. Table 2 provides a broad overview.

**Language** changes over time. This is most directly observed in a change of the terminology used as investigated by many sources (Duan et al., 2021; Holzmann & Risse, 2014b; Jatowt & Tanaka, 2012; Kaluarachchi et al., 2011; Kaluarachchi et al., 2010; Kanhabua, 2009; Leibscher, 2004; Ma et al., 2022; B. Wang et al., 2021; Whiting et al., 2012; Wu et al., 2018; Yoon et al., 2018). The terminology change describes a general change in the words used. New words appear and others are disappearing over time. These changes are measured through the word frequency, which increases or decreases over time. Often terms still retain their meaning but newer terms are used more frequently to describe the same concept and by that replace the old ones. This in general is a change observed over longer periods of time, however strongly depending on the domain. For example, terminology changes in social media are faster due to community languages and the quick content evolution (Stowe & Gurevych, 2021). While often this word migration appears over time and unnoticed, sometimes an event leads to terminology changes. This is especially the case for named entities (Chen et al., 2017; Holzmann & Risse, 2014a, 2014b). For example, the Indian city *Mumbai* was previously called *Bombay*, after her marriage *Meghan Markle* is referred to as *Meghan Duchess of Sussex* and the formerly known company *Google* was renamed *Alphabet*.

Language changes are investigated with the help of embeddings (Stowe & Gurevych, 2021). Therefore, the dataset is split into sub-collections based on the age of the documents and an LM is trained on each sub-collection. The change is made visible by comparing word embeddings for the same terms from different LMs. Further, to actively address the terminology change, the time can be modeled explicitly into the embeddings (Duan et al., 2021; B. Wang et al., 2021).

While the terminology change mainly refers to concepts that change their name, the opposite is true for the so-called concept drift. Concept drifts describe a change in meaning while the terminology stays the same (Duan et al., 2021; Efron, 2013; Forman, 2006; Irfan et al., 2018; Ma et al., 2022; Nishida et al., 2012; Wei & Dong, 2001; H. Zhou et al., 2017). This is often the case if topics evolve and the state of knowledge progresses. Thereby, the older meaning of a concept is overwritten by newer ones.

Beyond the words and their meaning, Delasalles et al. (2021) and van Dam and Hauff (2014) investigate the style of documents and conclude that the style is more similar if the documents were written closer together. Delasalles et al. (2019) observe that the changes are dependent on the author or the community of the authors.

All these observations inevitably lead to a vocabulary mismatch (Furnas et al., 1987). For example, a query and a document may contain the same concepts but no match occurs because the concept is described in different words. A change in language affects all other changing components and therefore can be understood as a super category of change

| Component | Change | References |
|---|---|---|
| Language | Terminology | Chen et al. (2017), Duan et al. (2021), Furnas et al. (1987), Holzmann and Risse (2014b), Jatowt and Tanaka (2012), Kaluarachchi et al. (2011), Kaluarachchi et al. (2010), Kanhabua (2009), Leibscher (2004), Ma et al. (2022), Stowe and Gurevych (2021), B. Wang et al. (2021), Whiting et al. (2012), and Wu et al. (2018), Yoon et al. (2018) |
| | Embedding | Duan et al. (2021), Stowe and Gurevych (2021), and B. Wang et al. (2021) |
| | Meaning | Duan et al. (2021), Efron (2013), Forman (2006), Irfan et al. (2018), Ma et al. (2022), Nishida et al. (2012), Wei and Dong (2001), and H. Zhou et al. (2017) |
| | Style | Delasalles et al. (2019, 2021) and van Dam and Hauff (2014) |
| Corpus | Expansion | Deveaud et al. (2023), Dumais (2010), Frieder and Jensen (2006), Holubová et al. (2019), Hopfgartner et al. (2019), Ibrahim and Landa-Silva (2014), Qian et al. (2016), Radinsky et al. (2013), Roberts et al. (2020), Ryu et al. (2008), Sáez, Goeuriot, et al. (2021), Sánchez et al. (2018), and Strötgen et al. (2012) |
| | Shrinkage | Bar-Ilan (2002), Dumais (2010), Frieder and Jensen (2006), and Sáez, Goeuriot, et al. (2021) |
| | Domain | Thakur et al. (2021) |
| Document | Updates | Adar et al. (2009), Altingövde et al. (2011), Bar-Ilan (2002), Dai and Davison (2010), Deveaud et al. (2023), Radinsky et al. (2013), Ryu et al. (2008), and Sáez, Goeuriot, et al. (2021) |
| | Update rate | Adar et al. (2009), Joho et al. (2014), Nunes et al. (2010), and Sisman and Kak (2012) |
| Content Topic | Arise | Delasalles et al. (2021), Tsevas and Iakovidis (2011), Wei and Chang (2007), Wei and Dong (2001), C. C. Yang et al. (2009), X. Zhang et al. (2019), and H. Zhou et al. (2017) |
| | Vanish | Chen et al. (2017) and X. Zhang et al. (2019) |
| Topic | Added | Deveaud et al. (2023), Hopfgartner et al. (2019), Kleinberg (2016), Sáez, Goeuriot, et al. (2021), Takeda et al. (2017), and L. L. Wang et al. (2020) |
| | Removed | Deveaud et al. (2023), Kleinberg (2016), Roberts et al. (2020), and Sáez, Goeuriot, et al. (2021) |
| Information Need | Understanding | Aliannejadi et al. (2020), Dumais (2014), Golovchinsky et al. (2012), and Zein (2021) |
| | Causes | Alonso (2013), Diaz and Jones (2004), Dumais (2014), Joho et al. (2014), Ma et al. (2022), Mansouri et al. (2017), Qian et al. (2016), Radinsky et al. (2013), Takeda et al. (2017), and Vouzoukidou (2015) |
| Queries | Traffic | Dumais (2014), Radinsky et al. (2013), and Sun et al. (2018) |
| | Type | Adar et al. (2009), Cheng et al. (2013), Dai et al. (2011), Dumais (2010), Ren et al. (2017), Sun et al. (2018), and Svore et al. (2012) |
| | Dependence | Carterette et al. (2015) and Dumais (2014) |
| User | Behaviour | Dumais (2014), Radinsky et al. (2013), Sahraoui and Faiz (2017), and Sun et al. (2018) |
| | Interrest | Yadav et al. (2021) and Zheng et al. (2020) |
| | Engagement | Aggarwal et al. (2020) |
| | Knowledge | Aggarwal et al. (2020) and Zein (2021) |
| Relevance | Vanish | Clarke et al. (2008), Deveaud et al. (2023), Salles et al. (2010), Tikhonov et al. (2013), and Uehara et al. (2005) |
| | Age | Clarke et al. (2008), Uehara et al. (2005), and Yeniterzi and Callan (2014) |
| | Realized | Tonon et al. (2015) |
| | Restricted | Kaluarachchi et al. (2010), Kanhabua and Anand (2016), Whiting et al. (2012), and Whiting et al. (2011), L. Zhang et al. (2022) |
| Results | Quantity | Altingövde et al. (2011) |
| System | Weighting | Perkiö et al. (2005) and Y. Yang and Kisiel (2003) |
| | Learn | Cohen (2021), Kuang and Clement H.C. (2019), and A. J. Zhou et al. (2015) |
| | Translate queries | Efron (2013), Jatowt and Tanaka (2012), and Kaluarachchi et al. (2010) |

Table 2: Overview of the different components, how they change and who investigated the change.

**Corpus** is the core component of every search system, as it contains all documents that can be searched. The most obvious change to the corpus is expansion. Over time, new documents are created and are added to the corpus (Deveaud et al., 2023; Dumais, 2010; Frieder & Jensen, 2006; Holubová et al., 2019; Hopfgartner et al., 2019; Ibrahim & Landa-Silva, 2014; Qian et al., 2016; Radinsky et al., 2013; Roberts et al., 2020; Ryu et al., 2008; Sáez, Goeuriot, et al., 2021; Sánchez et al., 2018; Strötgen et al., 2012). The frequency with which new documents are added depends for example on the domain. In a social network, where every user creates their own content, the corpus expands faster as in a corpus of patents. The frequency itself can carry value for the system, for example, to be used as an indicator of popularity in the case of product reviews. The more reviews are added to a product, the more popular it might be (Strötgen et al., 2012). Further, an expanding corpus holds implications for the system as the corpus statistics are changing over time. Considering added documents as the only change to the corpus in an append-only corpus, this can be used beneficially, as some corpus statistics can be estimated instead of calculated as shown by Mohammed et al. (2017).

But often documents are not only added but also removed (Bar-Ilan, 2002; Dumais, 2010; Frieder & Jensen, 2006; Sáez, Goeuriot, et al., 2021). For example in a web corpus, websites are shut down and not available anymore, therefore they also need to be removed from the corpus.

In contrast to a static test collection following the Cranfield methodology (Cleverdon, 1997), like the TREC collections, real-live corpora are dynamic. It has been shown that these corpora are not temporally reliable and outdate quickly after construction (Frieder & Jensen, 2006; Hashemi & Kamps, 2017; Soboroff, 2006).

A more drastic change in the corpus regards its domain. To reliably evaluate the performance of a system it is of interest how it performs on different domains. Therefore, the BEIR[9] benchmark contains datasets for different IR tasks on different domains (Thakur et al., 2021).

**Documents** are the individual parts of the corpus. The most prominent change of a document is the update (Bar-Ilan, 2002; Dai & Davison, 2010; Deveaud et al., 2023; Radinsky et al., 2013; Ryu et al., 2008; Sáez, Goeuriot, et al., 2021). In the web search domain, a document represents a website. Websites are not static but are often maintained such that new content is added or existing content is removed or changed. Altingövde et al. (2011) observed based on the AOL query log Pass et al. (2006), that after 3 years, the titles and URLs of webpages shrunk and the snipped size increased. Adar et al. (2009) showed that most often not entire webpages change but rather parts of them are changing while the rest remains steady. Continuing the webpage example, the outline, of a webpage, like general information, is often static, while its content changes.

How fast the content is changing, depends on the domain (Joho et al., 2014; Sisman & Kak, 2012). For example, news pages are updated more than hourly (Adar et al., 2009). In contrast, Wikipedia pages are more static, at least considering larger changes. This is investigated by Nunes et al. (2010) who compared the term frequency distributions of multiple document versions. Systems exploit the velocity of change as a feature, for

[9] https://github.com/beir-cellar/beir

example as a popularity indicator in social networks (Joho et al., 2014) or in the search for software bugs by identifying fast-changing code sections (Sisman & Kak, 2012).

Sometimes, a newly added document is actually only an evolved version of an old document. Considering web pages, this can lead to near duplicates on the result page, which should be avoided. If documents are publications, finding such documents can help to identify plagiarism (Cho et al., 2017).

**Content Topic** refers to the topic of a document in this section. The uncommon terminology is chosen to differentiate it from the topic in IR test collections as discussed later. With new documents added to the corpus, also new content topics arise (Delasalles et al., 2021; Wei & Dong, 2001; X. Zhang et al., 2019; H. Zhou et al., 2017). In a wider sense, other documents might contribute to an existing category. An example is the news domain where a news event sets the topic for a news story and further reports about the event continue to contribute to the topic (C. C. Yang et al., 2009). Another example is medical records of a patient. Every new record contributes to the patient's health history, which can be understood as a topic (Tsevas & Iakovidis, 2011; Wei & Chang, 2007).

In contrast to emerging categories, older categories vanish if no new documents are added that continue the topic (X. Zhang et al., 2019). Across documents, over time, topics can also split into further topics if they evolve in different directions. On the contrary, topics can merge as well to form a topic together (Chen et al., 2017).

The name of a topic is not necessarily fixed as it is subject to the language dynamics as described earlier. Therefore, it might change over time according to advancements in the field. This change is strongly connected to the concept drift. Therefore, an effort is put into tracking a topic and the related documents across time and evolving labels (Irfan et al., 2018; B. Li et al., 2006). If modeling the topic, the temporal order of topics in a stream might be important. For example in social media a post with the topic *smoking* followed by *respiration problems* might be a valuable insight (Sidana et al., 2016).

**Topics** capture the information need and queries in test collections. To make test collections less static, additional topics can be added (Deveaud et al., 2023; Hopfgartner et al., 2019; Kleinberg, 2016; Sáez, Goeuriot, et al., 2021; Takeda et al., 2017; L. L. Wang et al., 2020). Similarly, topics can be removed or updated (Deveaud et al., 2023; Kleinberg, 2016; Roberts et al., 2020; Sáez, Goeuriot, et al., 2021). Such actions have implications for the other components in a test collection, especially the relevance judgments. Theoretically, a topic could contain multiple and evolving queries (Deveaud et al., 2023).

**Information need** lies on the other side of the retrieval schema to the corpus. The information need changes over longer periods of time or even in a session. During one session for example, retrieved documents reveal answers that might lead to further questions (Aliannejadi et al., 2020; Dumais, 2014). By this, the understanding of the information need improves for the user (Golovchinsky et al., 2012; Zein, 2021). While the information need can change quickly, others remain steady for a long time, for example, the general interest in a music band. Such standing information needs can be fulfilled by recommending systems (Qian et al., 2016).

Changes in long-term information needs of a user are caused by different factors. Some change appear seasonally (Alonso, 2013; Joho et al., 2014; Mansouri et al., 2017) or follow

other trends (Diaz & Jones, 2004; Dumais, 2014; Vouzoukidou, 2015). Others are triggered by certain events, which can be grouped into personal or collective events. For example, marriage or pregnancy are personal events that change the information need to prepare for the upcoming life changes (Takeda et al., 2017). Collective events in contrast are for example news that trigger information needs related to gaining a better understanding of the incidents (Ma et al., 2022; Radinsky et al., 2013).

As the information need is mainly investigated by a query, its immediate verbalization, it will be further discussed later.

**Queries** are investigated for change mostly on a collective level, like the number of queries received by a search engine (Dumais, 2014; Radinsky et al., 2013; Sun et al., 2018). They are a direct, although incomplete indicator of the information need of the users.

Ren et al. (2017) identify four different patterns of long-term query frequencies. They categorize them into *stable queries*, *one-time burst queries*, *periodic multi-time burst queries*, and *periodic multi-time burst queries*. Stable queries have no major spikes, so they are timeless and are not bound to events (Cheng et al., 2013; Ren et al., 2017). One-time burst queries experience explosive popularity for a short time but are generally low in frequency. Multi-time burst queries have multiple, smaller popularity spikes often triggered by irregular events that occur more often. Lastly, the *periodic multi-time burst queries* are queries that have a clear repeating pattern bound to events that occur frequently. (Dai et al., 2011) further decompose such traffic time series into components and thereby separate trend traffic from seasonal traffic. Svore et al. (2012) show that trending queries benefit from longer snippets than general queries.

Sun et al. (2018) investigate queries on an individual level and differentiate them into a static interest and a dynamic interest. Static queries contribute to the long-term information need and represent a general interest. The dynamic interest accounts for the remaining queries. Further, it is investigated if, and how queries in one session depend on each other (Carterette et al., 2015; Dumais, 2014)

Dumais (2010) investigate navigational queries and find four patterns of revisitation queries. They show that 60-80 % of queries are to re-visit webpages. Adar et al. (2009) relate the re-visitation patterns to changing website content.

Ren et al. (2017) and L. Zhang et al. (2018) find that query changes are dependent on time and location. For example the query *fashionable haircut* asks for a recent style. Since it is fashionable it, is highly temporal.

**Users** change collectively or individually (Dumais, 2014; Radinsky et al., 2013; Sun et al., 2018). This is similar to the change observed in the queries. On the other dimension, a change between long-term and short-term changes can be differentiated (Sahraoui & Faiz, 2017). Qualitatively, the interest of users change (Yadav et al., 2021; Zheng et al., 2020) and related to that the engagement (Aggarwal et al., 2020) changes as well. The knowledge of a user changes, sometimes during a session if the user learns from the results, which is directly reflected in the queries (Aggarwal et al., 2020; Zein, 2021).

**Relevance** of documents is not static. Previously relevant documents can quickly become not relevant anymore (Clarke et al., 2008; Deveaud et al., 2023; Tikhonov et al., 2013; Uehara et al., 2005). For example, web pages are outdated if not updated sufficiently.

Therefore, especially in web search, the novelty of a document, described as freshness, is an important relevance indicator (Clarke et al., 2008; Uehara et al., 2005). Additionally, the documents ranked in a session depend on each other. Documents that are investigated before but did not fulfill the information need might lose relevance. Further reasons for a change in relevance are that new and more relevant documents are retrieved (Salles et al., 2010), or that the assessment protocol evolved (Sáez, Goeuriot, et al., 2021). The relevance assessment is further problematic, since it relies on the pools of the contributing systems and is costly to create. Therefore, Tonon et al. (2015) proposes to distribute the effort across time and contributors, so that new topics can be added and assessed. Frieder and Jensen (2006) find that to judge live systems, more queries need to be judged compared to test collections.

To exploit that the relevance is not static, temporal relevance profiles are modeled for documents (Kaluarachchi et al., 2010; Kanhabua & Anand, 2016; L. Zhang et al., 2022) and topics (Whiting et al., 2012; Whiting et al., 2011). Such profiles specify a time range in which the most relevant results can be achieved.

Yeniterzi and Callan (2014) state that aggregating relevance feedback is temporally biased and needs to be normalized since older documents have more time to gather feedback.

**Results** evolution is investigated by Altingövde et al. (2011). They find that for queries with already many results, even more are found at a later point in time, while for queries with fewer results, relatively fewer additional documents are found. This is known as the Mathews effect and contributes to the Pareto distribution of query results. They investigated this by revisiting the AOL query log dataset (Pass et al., 2006) for the web domain.

**Systems** can be considered static in most evaluations. Exceptions are systems that explicitly use temporal features to adapt the ranking, like favoring newer documents over older ones. However, such TIR systems were not the focus of this analysis.

Still, some systems change beyond only adapting to temporal features. How to adapt to temporal features appears to be difficult because newer is not always better. However, trends are an important indicator for relevance (Perkiö et al., 2005). Mohammed et al. (2017) found that in append-only collections, like archives, some document statistics scale uniformly and therefore can be estimated precisely, rendering some recalculations unnecessary. More prominently, adaptive filtering modifies a threshold for relevance over time to focus on recent user interrests (Y. Yang & Kisiel, 2003). Further systems actually learn the ranking function over time. These systems are based on reinforcement learning (Cohen, 2021; Kuang & Clement H.C., 2019; A. J. Zhou et al., 2015). The relevance of documents for a query is newly estimated based on the past rankings or recommendations produced and the user feedback, for example on clicks. This can be done on a session level or beyond through a user profile. As a conclusion to the observed changes in language, many efforts are done to adapt a query to linguistic changes (Efron, 2013; Jatowt & Tanaka, 2012; Kaluarachchi et al., 2010). This may be to expand queries with similar terms or categories of different points in time or to boost documents according to the temporal pattern of a query.

### 3.4.4 Measures

Persistence is measured rarely. The opposite of persistence change is more often the focus of the studies. By measuring change, persistence can be estimated as the absence of change. How change is measured depends on the type of change. As a starting point, change can often simply be counted, for example in the case of traffic or documents added. More precisely, this describes the delta between two points in time. With additional measurements of further points in time, a time series emerges. Advanced measurements, set measures in relations to better differentiate between the origin of effects. Changes in the language used are diversety.

Jatowt and Tanaka (2012) calculate summarization statistics of language change over decades. They use the measures: unique words per decade, change in word length, temporal entropy and temporal kurtosis of word distributions. These measures give a general overview. To measure more nuanced changes, Stowe and Gurevych (2021) investigate how rapidly language changes over short time periods affect classification. They determine a normalized accuracy delta between datasets from different points in time as Temporal Rigidity (TR) defined as:

$$TR = \frac{1}{N} \sum i \neq j \frac{|M_s(EE_i, EE_j) - M_s(EE_i, EE_i)|}{|i - j|} \tag{8}$$

where $M_s$ is the measured performance (F1 in this case) of system $s$ trained on samples of the sub-collection $EE_i$ and evaluated on samples of $EE_i$ or $EE_j$ respectively. Further, they measure the correlation of results to an underlying temporal structure, like the sequence of events.

Delasalles et al. (2019) use a complex approach to investigate and model the language of author communities. They use the cosine similarity of embeddings that incorporate temporal features and measure the perplexity gain.

User focussed change on an individual level is observed by Zein (2021) who estimates the knowledge gain of users. The knowledge is represented through the known documents represented with Language Models (LM) and also through a reference benchmark. In this benchmark, users were asked questions to test their knowledge and then were allowed to research the web before re-answering the questions. The per-page knowledge gain was calculated with the help of a linear regression over all visited pages. The regression coefficient then represents the page knowledge gain.

Changes in the corpus are investigated by Bar-Ilan (2002) who quantify changes in web corpora based on ten measures. The freshness of documents and their stability over time is of their concern. The freshness factors are the percentage of broken links, the percentage of pages updated and the number of new pages. Further variations are possible if data from more points in time are available. Then, differentiations between new compared to the last point in time and totally new compared to all points in time are possible. The same is true for the measurements based on broken links. These measurements capture the stability by comparing how often websites were available or not available.

Tikhonov et al. (2013) explore temporal patterns of web pages using web traffic. They classify the time series of user traffic for websites with a limited lifespan into four categories. Based on the measured web traffic, a gradient-boosted decision tree is used for classification.

On a document level, Dumais (2010) describes different measures to quantify change and plots to visualize them. As a summary, the number of web pages changed overall or on a user level, as the number of visited pages changed are described. The amount of changes is captured by the average Dice coefficient which can be visualized as a curve with the x-axes as time over different timeframes. Also, the time between changes is a further measure. On a document level, changes can be measured up to a term-level difference.

Sánchez et al. (2018) are interested in measuring the temporal novelty of recommender system results. They measure how much new items are favored by a system. Therefore, a novelty function is defined as the first or last appearance of an item or the average or median interaction of an item. To enable comparisons with other items and allow aggregation, the item novelty is normalized. Novelty is then defined as and reproduced from Sánchez and Bellogín (2018):

$$nov^{f,n}(i|\theta_t) = n(f(\theta_t(i)), \theta_t) \tag{9}$$

where the novelty of an item $i$ and its temporal representation $\theta_t(i)$ is measured, based on a novelty function $f$ and a normalization method $n$.

Closser related to temporal effects on the effectiveness of a system, Dai et al. (2011) formulate a hybrid nDCG which additionally incorporates freshness into nDCG and allows to weight the importance between relevance and freshness which is equivalent to novelty. They define the hybrid nDCG as:

$$hybrid\,nDCG(n) = Z_n \sum_{j=1}^{n} \frac{2^{(\gamma y_R + (1-\gamma) y_F)} - 1}{log_2(j+1)} \tag{10}$$

$Z_n$ is the oracle (best possible) nDCG that bounds the hybrid nDCG between 0 and 1. $\gamma$ is the weighting factor between the relevance $y_R$ measured as DCG and the freshness $y_F$. Additionally to the novelty measures described before, Dai et al. (2011) propose to use page maintenance as a reference point for freshness. This measure allows for evaluating a system's performance with regard to temporality.

X. Zhang et al. (2019) measure word co-occurrences of keywords in scientific publications over time to evaluate the research field evolution. A word co-occurrence takes place if two keywords appear in the same paper. By connecting these keywords, a network emerges and the edges are weighted by the number of co-occurrences, forming the bases for advanced network analysis.

Thakur et al. (2021) measure the pairwise weighted Jaccad similarity for different datasets. They show correlations between the Jaccard similarity and the domain of the datasets. Further, they evaluate the effectiveness of IR systems on multiple datasets from different domains. They use nDCG@10 as the only measure and BM25 as pivot system. Results are also reported as the number of systems performing better than the pivot system.

Advancing inter-system and dataset evaluations, Sáez, Goeuriot, et al. (2021) propose to use Result Deltas denoted as $\mathcal{R}\Delta$ to measure change between systems on the same environment, different environments and the combination of both. The $\mathcal{R}\Delta$ is defined

with respect to the system as:

$$\mathcal{R}_s\Delta = M(S_1, EE_1) - M(S_2, EE_1). \tag{11}$$

It describes the difference between the results of the systems $S_1$ and $S_2$ achieved by measure $M$ on test collection $EE_1$. Likewise with respect to different EEs as,

$$\mathcal{R}_e\Delta = M(S_1, EE_1) - M(S_1, EE_2) \tag{12}$$

describes the difference between the system $S_1$ on the two different test collections $EE_1$ and $EE_2$ measured by $M$.

$\mathcal{R}_{se}\Delta$ tries to capture the difference between two systems, evaluated on two different test collections. Since all components are changing, this is hard to measure. For estimation, a pivot system is introduced and the performance of a system $S_1$ is measured in relation to the pivot system on $EE_1$ and then compared to the difference between the performance of system $S_2$ and the pivot system on $EE_2$. This is formalized by Sáez, Mulhem, et al. (2021):

$$\mathcal{R}_s\Delta(Pivot, S_1, EE_1) = \frac{M(S_1, EE_1) - M(Pivot, EE_1)}{M(Pivot, EE_1)} \tag{13}$$

and respectively for all other systems compared. In this constellation, a consistent and correct pivot system is crucial. The same measures are planned to be used as relative nDCG in the LongEval shared task (Deveaud et al., 2023).

Tonon et al. (2015) propose two measures that describe how fair systems are evaluated based on expanding test collections. As they propose to further expand test collections with new systems entering a ranking, retrieval results are not directly comparable anymore. To measure that a Fairness Score and the opportunistic number of relevant documents are proposed. Evaluations with pooling-based test collections assume not judged documents to be not relevant. With relevance judgments added and systems evaluated simultaneously, systems that are evaluated later and with more judgments are more precise. To measure that, the Fairnes Score (FS) is measured based on the AP and formulated as:

$$FS(run) = \frac{\sum k = 1^{k=1} JudCov(k) \cdot J(k)}{n}. \tag{14}$$

$JudCov(k)$ is the fraction of judged documents in the top $k$ ranking positions of the results for a topic and $J(k)$ is 1 if the $k$-th document is judged and 0 if not. The measure is 0 if no retrieved documents are judged and 1 if all are judged.

Further, they define the relative difference between a new and the current best system which is equivalent to eqasion 3.4.4, just that the best system instead of a stable pivot system is used. In the continuous evaluation framework as described by Tonon et al. (2015) this can be used to calculate an upper and lower bound of an expected performance, taking the documents that would need to be judged into account. Likewise, the opportunistic number of relevant documents is calculated as the minimum number of documents needed to reliably show that a new system outperforms the current best system. This helps to estimate the cost of an evaluation.

## 3.5 Summary

An extensive literature review was conducted to investigate persistence over time. By that the research question *How is the environment of an IR system evolving?* is answered. The question is further divided into three sub-questions which are first quantitatively and then qualitatively assessed. Concerning the sub-question *What are the evolving components of an IR environment?*, 11 components are identified. They can be organized partially in a hierarchy but sometimes also in parallel. At the core, they resemble an IR test collection. Further components are more related to the user or the system. Closely connected to the components, the change is investigated in the sub-question *How do the components in an IR evaluation environment evolve?* The changes observed often affect multiple components and cannot clearly be discriminated. Often a duality can be observed between individual change and collective change or short-term and long-term change. For example, a document changes but also the corpus as the entirety of documents changes as well, or the interest of a user changes and collectively this forms trends. All in all, the environment of an IR system is highly dynamic with most components evolving over time.

The third question *How can this evolution be measured?* is difficult to answer, as no method or framework is established yet as state of the art. Rather different measures are used to quantify the various changes in the evaluation environment. To approximate persistence, few or no changes can be interpreted as persistence. On a high level, change is measured as the delta between, mostly two, points in time. Long-term investigations, comparing multiple points in time are rare, except to investigate language change. Only a few attempts could be found that actually investigate how changes in the evaluation environment affect retrieval effectiveness. More specific measures are missing, which makes a nuanced comparison difficult. As shown earlier, results are influenced by many changes making comparisons difficult again. Often it is not clear to what components the observed changes should be attributed to. This also shows the importance of using multiple measures to gain a more complete understanding.

An open research question remains about how resilient different IR systems are against evolving environments. Only little research explicitly focuses on this question, but with the LongEval shared task, a first testbed is created to specifically address this question. Besides testbeds that are capable of isolating different components to evaluate their impact on the effectiveness of a system, dedicated measures are needed to quantify that. While the proposed result deltas are a first step in this direction, more precise measures are yet to be found. A further open question regards how to file an improving performance over time. While persistence is not given if better results are obtained over time, the users benefit from better results.

Methodologically, this study was only partially successful. With less than 10 % relevant publications found, the gained findings in relation to the effort seem not in balance. Especially the strong correlation between relevant publications and primarily considered IR venues suggest focusing on these sources. Almost all literature from the preliminary search was re-discovered in the review. However, additional highly relevant literature was also found after the review. This shows that the keywords that were used could not extensively capture all aspects of the topic.

In conclusion, to consider the evolving environment during effectiveness evaluation, many components need to be considered with differentiating and specific measures. While many effects can be measured, they often overlay and it remains difficult to differentiate the effects to precisely attribute them to the system or the EE. The diversity and number of results can be interpreted as a general concern about temporal persistence. However, the lack of specificity also shows a general research gap.

# 4 The Evaluation Environment

The effectiveness of a retrieval system does not solely depend on the system's capabilities. Rather it is influenced by various other components that it is exposed to or relies on. In a Cranfield test collection (Cleverdon, 1997), these components are fixed. In other evaluation cases, they are not. Sáez, Goeuriot, et al. (2021) define these components as the Evaluation Environment (EE). They include the entirety of resources that are needed to evaluate an IR system. Focussing on test collections, they name the topics, the corpus and the qrels as well as the evaluation measures and the pooling strategy. Based on the components and changes we identified in the literature review, we extend the EE beyond static test collection experiments to capture the general IR problem as described in Section 2.1. Therefore, the core components, the query, the corpus and the results function as general categories, to which the other components are vaguely related too. Since each category contains different changes and these changes may happen on different levels, the inevitable emerging hierarchy does not necessarily fit. Likewise, rather different changes may be grouped in a category, just because they are related to the same component. This is difficult to avoid, considering the versatility of the different components and changes.

We comprised this in a schematic visualization shown in Figure 7, which provides a visual overview and helps to navigate the different components in continuous experiments. The system is located in the center and the other components which influence it are located in layers around the system. First, the three core components form broader categories followed by the more granular components. The language surrounds the EE as an aura due to its ubiquitous effect.

For navigational purposes, the different components which are assessed in an experiment can be highlighted as shown in Figure 8. The EEs of the three experiments which will be described in the next section are visualized. As indicated by the highlight of multiple components, a change in more than one component is often investigated. This is due to the difficulties of isolating changes. Therefore, the highlights need to be understood as the focal point of an experiment and not as the solely changing components. For example, when investigating changes in document updates, like by using reproducibility measures on two sub-collections, these sub-collections may be affected by an evolving language.

In the following, the three main components, the query, the corpus and the results are further described. Then, a first set of measures is proposed to quantify changes in the EE and the findings are summarized.

## 4.1 Query

Queries underlie different changes. For example on a collective level as the traffic of a search engine that forms trends, or on an individual scale as changing expression of an information need in a session or topic. The user, information need and topic are directly related to the query. The query is a direct manifestation of the information need from a user, which is captured in a topic. User related changes are often observed in relation to their interests, engagement and knowledge. This occurs over different timeframes, from a session to longer periods. The information need of a user can as well change during a

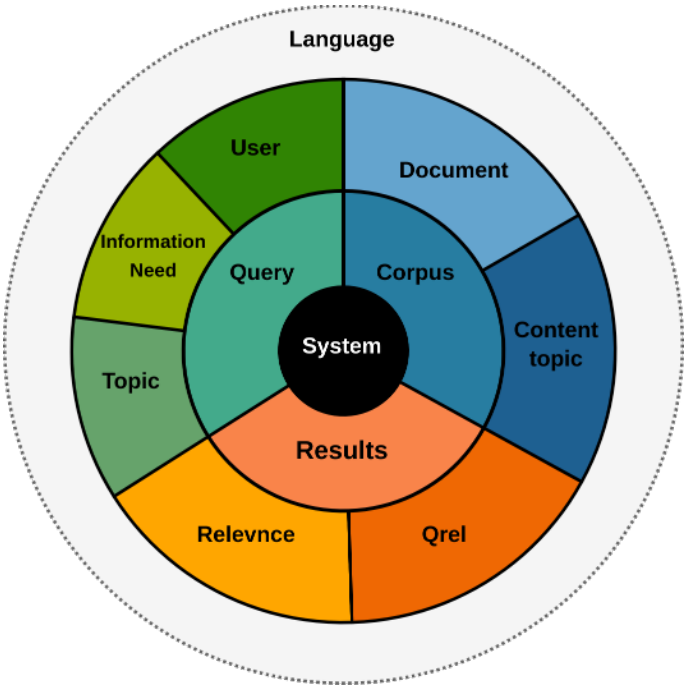


Figure 7: Schematica visualization of the Evaluation Environment (EE).

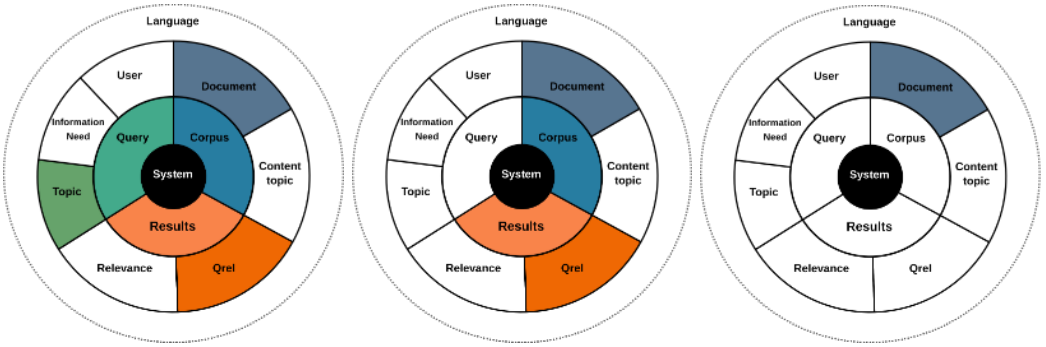


Figure 8: Schematic visualization of the EEs for the three experiments, Result Delta (left), replicability (center), reproducibility (right) as detailed in Section 5.

session with an improved understanding of the topic. Longer-term changes have different triggers like seasons or events. These components are abstracted and fixated in topics containing a fixed set of queries and ideally further description of theire information need. By variating these topics, changes can be simulated in a controlled setting. Nevertheless, in connection with the other components in a test collection, they still underley these changes. This shows how these components are strongly interconnected.

## 4.2 Corpus

The corpus consists of documents that are related to domains and categories. Changes in the corpus can be partially aligned with the CRUD operations, create, update and delete. Either documents are added or removed from the corpus or documents are updated. More precisely, this happens on the document level, which indirectly influences the corpus. Through these changes, the collection changes its semantic content. The content

topics evolve over time, which can be modeled and tracked. Thereby, new topics get more attention and others less. Topics also split into different ones over time or can merge. This is strongly connected with concept drifts and form the domain on a global level. These changes may directly influence the effectiveness of an IR system if the properties of the corpus diverge. While the system was initially adapted to a corpus, or may only perform in a certain way on a corpus during assessment, these properties may not be given anymore. Therefore, the true effectiveness of the system on the diverging corpus is unknown.

## 4.3 Results

The results are the final artifact of the retrieval effort, combining the query and the corpus. Ideally, they are relevant. The results may change in quality based on the corpora and queries, but also in quantity over time. The relevance is dynamic since documents become outdated or are overshadowed by newer and more relevant documents. Further, facts can change. The relevance corresponding to query document pairs is preserved in the qrels. These assessments are often not available for all combinations, so that additional relevance assessments can be added later. Further, changes in the assessment strategy influence the documents that are assessed or the guidelines for the assessments. Since the effectiveness of a system is directly evaluated by the results, they have an immediate influence, independent from the type of experiment.

## 4.4 Measures and Methods

The systems and measures are not understood as part of the EE. This is because they are exchangeable in this context. The measures are connected with the different components of the EE as they quantify the change. These might be for example to measure how documents have evolved, how the traffic is composed over time or how this influences the effectiveness of the systems.

In addition to the EE, a set of measures and methods is proposed. Since the components group different changes, and not all changes may be assessed, the list is not necessarily exhaustive. It rather shows a first attempt to quantify the changes in the EE. Further, these measures and methods are not necessarily bound to specific components but often can be applied to different or multiple ones. Especially the text-based components like documents and queries can often be assessed with the same measures. Therefore, the group of textual measures is introduced. Further, user-centered measures and methods quantify changes like traffic evolution or user knowledge and the relevance measures quantify the effect on the system.

**Textual**:

- Relative text length
- Relative number of changes
- Novelty
- Jaccard similarity
- Dice coefficient

- Word co-occurrences
- Topic modeling measures
- Unique words
- Entropy of word distribution
- Kurtosis of word distribution
- Embedding similarity
- Perplexity gain

**User**:

- Number reformulations in a session
- Traffic evolution
- Traffic per query type
- Knowledge tests
- Clicks and click model indicators

**Relevance**:

- Relevance profiles of documents and queries
- Evaluation fairness score
- Result deltas
- Cross-domain evaluation

## 4.5 Summary

Defining the EE is a foundational step necessary to address changes that influence the effectiveness of IR systems. To differentiate precisely between effects, locating them is a necessity. Therefore, the different components need to be investigated on their own and also the effect on the effectiveness of the systems.

All main components of traditional experiment types can be found in the EE. For example the corpus, topics and qrels a static test collection is composed of, or A/B tests and living labs with the corpus, queries and user interactions. By expanding the EE beyond the traditional test collection, it gets more applicable to all types of experiments and also incorporates additional components, with their changes, which previously could not be mapped.

Related to the EE, different measures and methods are gathered, that are related to the different components. While this list is certainly not exhaustive and further specifications and investigations are needed, it may be used as a foundation for future work. The magnitude of change that can be observed in relation to the EE shows how continuous evaluation of an IR system is not possible without taking the changing EE into account.

# 5 Evaluating Temporal Persistence Using Replicability Measures

In real-world Information Retrieval (IR) experiments, the Evaluation Environment (EE) is exposed to constant change. Documents are added, removed or updated, and the information need and the search behavior of users is evolving. Simultaneously, IR systems are expected to retain a consistent quality. The LongEval Lab seeks to investigate the longitudinal persistence of IR systems, and in this work, we describe our participation. We submitted runs of five advanced retrieval systems, namely a reciprocal rank fusion approach, ColBERT, monoT5, Doc2Query and E5, to both sub-tasks. Further, we cast the longitudinal evaluation as a replicability study to better understand the temporal change observed. As a result, we quantify the persistence of the submitted runs and see great potential in this evaluation method.

## 5.1 Introduction

This paper describes our contribution to the CLEF 2023 LongEval Lab.[10] The lab seeks to investigate the temporal persistence of retrieval systems and therefore provided a first-of-its-kind web retrieval collection with three sub-collections from different points in time (Deveaud et al., 2023). We participated in the retrieval task by providing runs of five systems to both sub-task.

A retrieval system's Evaluation Environment (EE) is under constant change. Not only but especially web retrieval systems are exposed to this due to the dynamic nature of the web. Documents, websites in this case, get created, updated or created (Bar-Ilan, 2002; Dumais, 2010). But besides the evolving collection, all other aspects of an EE underlay change as well, from the information need and search behavior of the users (Adar et al., 2009) all the way to the evolving language itself (Jatowt & Tanaka, 2012). These changes raise questions about the persistence and generalizability of IR system effectiveness evaluations.

By requiring a temporarily reliable system to perform consistently over time, evaluating this can be understood as a replicability task. Oriented at the ACM definition of replicability[11], the goal is to achieve the same measurements in a different experimental setup, in this case, at a proceeded point in time.

To investigate temporal persistence, we submitted runs of five advanced retrieval systems to both sub-tasks of the LongEval Lab. The systems are not specifically adapted to change or the LongEval dataset to form an idea of how temporal reliably system-oriented IR evaluations following the Cranfield paradigm are. Further, as a proof of concept, we use the replicability measures Delta Relative Improvement ($\Delta$ RI) and the Effect Ratio (ER) Breuer et al., 2020 to investigate the temporal persistence. In short, the contributions of this work are:

- Descriptions of **five state-of-the-art systems** submitted to both retrieval sub-tasks,

[10] https://clef-longeval.github.io
[11] https://www.acm.org/publications/policies/artifact-review-and-badging-current

- an **extensive evaluation** of retrieval effectiveness,
- an **adaptation of replicability measures** to evaluate temporal persistence,
- an **open-source release** of the experimental setup.

The remainder of this paper is structured as follows. Section 5.2 contains an analysis of the LongEval dataset. The five retrieval systems are described in Section 5.3. Further, Section 5.4 provides the results on the train slice and a preliminary evaluation of the results. In Section 5.5, we describe the replicability efforts. This paper concludes with a short discussion and some future work in Section 5.7. The code is publicly available on GitHub.[12]

## 5.2 LongEval Dataset

To our knowledge, the LongEval dataset (Deveaud et al., 2023) is the first dataset specifically designed to investigate temporal changes in IR. On a high level, the collection consists of three sub-collections from different points in time. Each collection contains topics and qrels. The documents as well as the topics and qrels originate from the French, privacy-focused search engine Qwant.[13] For this work, we entirely rely on the English automatic translations of the dataset. The documents contain the cleaned content of websites. They are filtered for adult and spam content, but no further processing was done, sometimes leaving unconnected phrases, keywords or code artifacts in the documents.

The topics are selected according to *"popularity, stability, generality, and diversity"* (Deveaud et al., 2023). For these topics, queries are selected from the Qwant search engine logs if they contain the topic as a sub-string. The qrels for the shared task are simulated based on the Cascade Click Model (Chapelle & Zhang, 2009; Craswell et al., 2008). Documents are assessed as not relevant, relevant and highly relevant. Further, human-assessed gold labels are announced for September (2023). More details can be found in the original publication (Deveaud et al., 2023).

The sub-collections are sequential snapshots of an evolving search environment for temporal comparison. The topics are constructed once, but the queries are partially changing across sub-collections. The documents, i.e., the websites identified by the URL, are also mainly static across sub-collections but the content of the documents changes.

The collections are organized into a WT, ST and LT sub-collection. The WT (within time) sub-collection was created in June 2022. The ST (short-term) sub-collection was created in July 2022, immediately after the WT collection. The third sub-collection, LT (long term), contains more distant data as it was created with a one-month gap from ST in September 2022. Table 3 gives an overview of the sub-collections.

The LongEval dataset contains over 1.5 million documents. Not every document is present in every sub-collection, but most documents do. The core document collection contains 1,011,613 documents. They are present in every sub-collection but do not necessarily contain exactly the same content. The documents evolve over time, meaning that the content of one website might change over time. To capture this change on a general

[12] https://github.com/irgroup/CLEF2023-LongEval-IRC
[13] https://www.qwant.com/

| | WT | ST | LT | Intersection |
|---|---|---|---|---|
| Timeframe | June 2022 | July 2022 | September 2022 | |
| Number documents | 1,570,734 | 1,593,376 | 1,081,334 | 1,011,613 |
| Mean document length | 794.11 | 793.96 | 807.28 | |
| Min document length | 0 | 0 | 1 | |
| Max document length | 7065 | 12210 | 7255 | |
| Number queries | 753 | 860 | 910 | 124 |
| Mean query length | 2.73 | 2.71 | 2.52 | |
| Min query length | 1 | 1 | 1 | |
| Max query length | 6 | 11 | 9 | |

Table 3: LongEval subcollection statistics. The length of documents and queries are measured in tokens, split on white spaces. The query WT q062213307 and ST q072211861 is excluded as an outlier since it only contains the token *leg* 108 and 110 times.

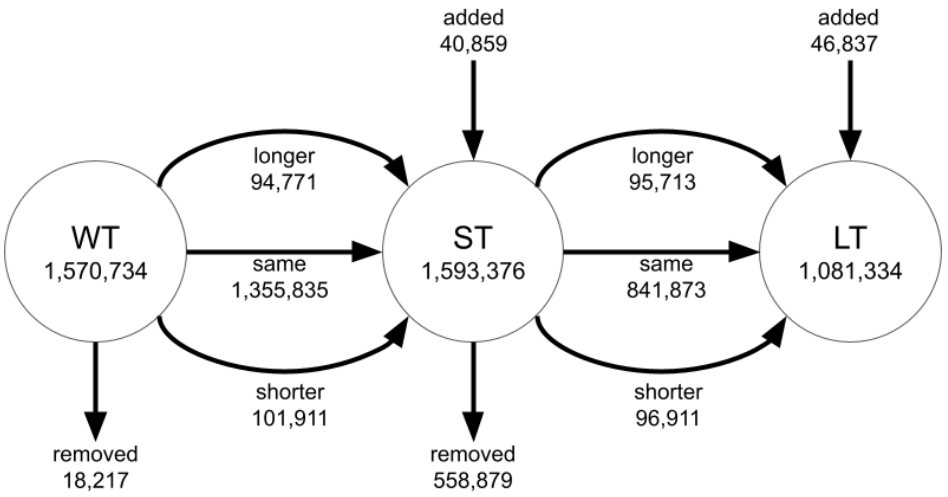


Figure 9: The evolution of the LongEval dataset documents across the three sub-collections. Transitioning from one sub-collection to the next, documents are added, removed or updated.

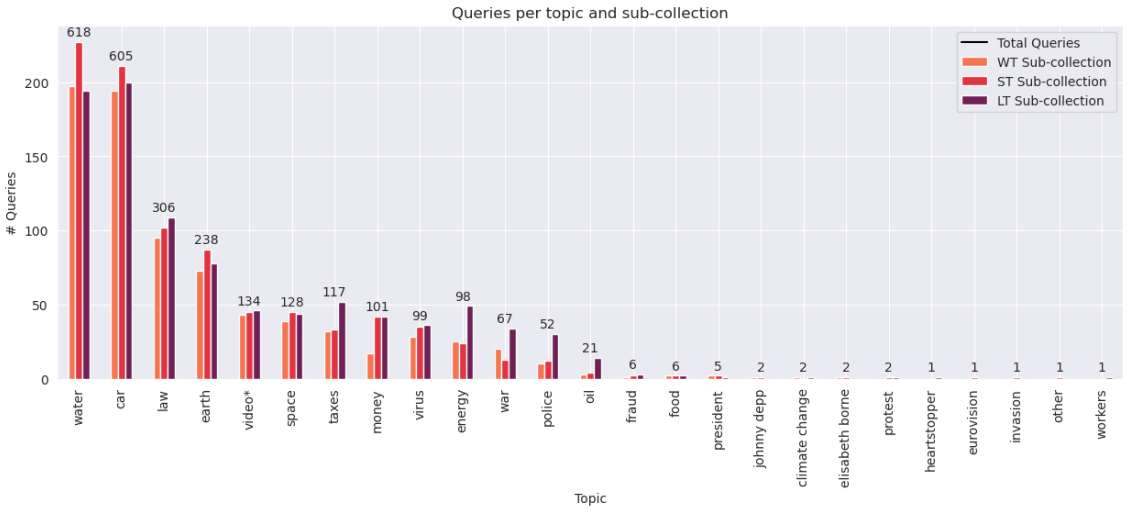


Figure 10: Distribution of queries over the topics per sub-collection and total. The plot shows the long-tail distribution of topics.

level, Figure 9 shows how many documents increase or decrease in character length and how many documents are added, deleted or stay the same in length.

Like the documents, the queries change over time as well. However, relatively fewer core queries that appear in all sub-collections exist. In total, only 124 unique query strings appear in all collections. Considering the query ids, the overlap is larger. This is due to duplicate queries, probably caused by the automatic translations.

Following the method described by (Deveaud et al., 2023), the queries can be attributed back to the original topics. Therefore, each query gets all topics assigned where the query contains the topic as a substring. To match the original procedure, the mapping is done based on the French topics and queries and the English names are only used for reporting at the end. Unluckily, this leaves many queries without a topic. Three exceptions are introduced to assign these queries a topic. First queries that contain the string *taxe* are assigned the topic *impots*(taxes). Then, an additional topic *video* is introduced which can be assigned to 134 queries. Finally, the query *amber heard*[14] is assigned to the topic *johnny depp*.

The number of queries per topic variates highly between only one query up to 618 queries in total. Figure 10 shows the full distribution of queries per topic and sub-collection. The distribution appears to be skewed with a long tail of topics that only appear less than 50 times. Seven of the 24 topics even don't appear in all three sub-collections.

Investigating the queries of the largest topics *eau*(water), many topics that were falsely related to the topic appear. For example the query *aeroport bord**eau**x*[15] contains the topic *eau* as a sub-string, but has nothing to do with the topic. This may be the case for further topics and impedes per-topic evaluations without further assessment.

The qrels classify documents on a scale of not relevant, relevant and highly relevant. In general, the dataset has few assessed documents per topic. While the mean number of qrels is 14 per topic, the absolute number fluctuates between 2 and 59. Figure 11 shows the distribution of all qrels per query. Most of the documents are marked as not relevant,

[14]LongEval ST qid: q072287
[15]LongEval WT qid: q062228

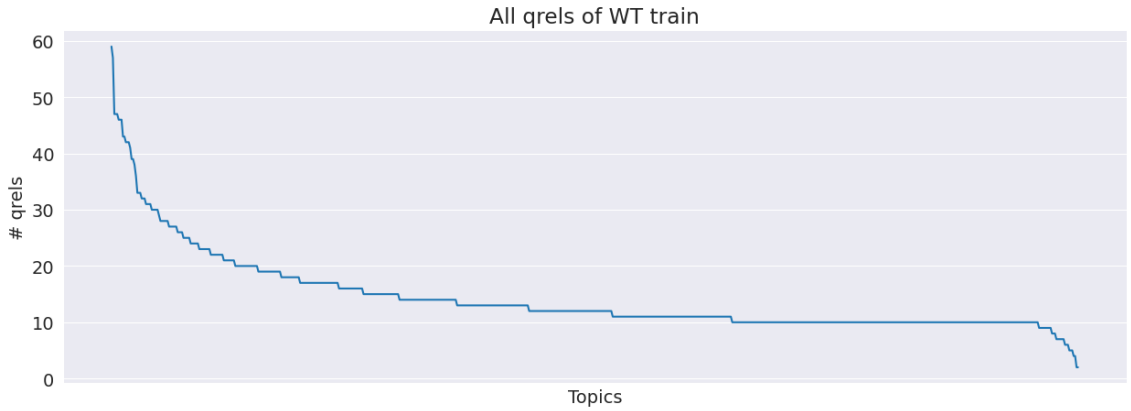


Figure 11: Distribution of qrels per query for the WT train sub-collection.

and the distribution of relevant and highly relevant qrels is skewed as well. Especially the highly relevant qrels are rare, with a maximum of only four and a mean of only one highly relevant document per topic. These documents carry a specially high weight because they are so rare and may change a ranking drastically. While relevant qrels are generally rare, 16 queries don't have a single relevant document.

## 5.3 Approaches and Implementations

We compared different ranking functions and multi-stage retrieval systems on the WT train slice of the LongEval dataset. The systems were chosen as they represent state-of-the-art, off-the-shelf methods that are used in many evaluations. Therefore, it is especially interesting how these systems behave over time without being specifically adapted to a changing environment.

### 5.3.1 Statistical Ranking Functions

Different ranking functions were used as baselines in their default configurations. Special attention was given to the BM25 (Robertson et al., 1994) ranking function as it has been proven to be a robust, efficient and often hard-to-beat baseline. We use this run to compare advanced systems to it. Since we use the PyTerrier (Macdonald & Tonellotto, 2020) framework for experiments, the default parameters $k_1 = 1.2$ and $b = 0.75$ were kept. Further we included PL2 (Amati, 2003), TF-IDF and XSqrA_M (Amati, 2006).

For advancing the ranking functions, two query expansion methods are employed. Namely, RM3 (Jaleel et al., 2004) and Bo1 (Amati, 2003) are used to extend the queries through pseudo-relevance feedback. The default PyTerrier parameters are also kept here; three feedback documents were used to gather ten feedback terms.

### 5.3.2 Rank Fusion

Multiple runs were combined into a single ranking to profit from the diversity of multiple ranking functions. First, BM25, XSqrA_M and PL2 are fused through Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) with the `ranx` Python library (Bassani & Romelli, 2022). Further runs are created by using the pseudo-relevance-feedback methods on top of

BM25. The default parameters $min_k = 10$, $max_k = 100$ and $step = 10$ were used for the RRF.

### 5.3.3 ColBERT

ColBERT (Khattab & Zaharia, 2020) applies the BERT (Devlin et al., 2019) Language Model (LM) to overcome the lexical gap (Furnas et al., 1987) by creating semantic representations of queries and documents as embeddings. In contrast to traditional BERT-based approaches like cross encoders, the interaction mechanism used to calculate the similarity between a document and a query is detached from the embedding creation process. However, in contrast to bi-encoder systems, nuanced similarities can be calculated. To do so, semantic representations for a query or a document are calculated as a set of token embeddings. The relevance score between a query and a document is then calculated as the sum of the max of the cosine similarity or the L2 distance between all embeddings for the query and the document.

By separating the scoring from the embedding process, the efficiency at run time can be greatly improved as all document embeddings can be calculated beforehand offline. ColBERT can also be used in a later retrieval stage as a reranker. The PyTerrier version of ColBERT [16] was used in a zero-shot fashion. Besides using ColBERT as a first-stage retriever, where the whole corpus is converted to embeddings, ColBERT was also used to rerank the top 1000 BM25 results.

### 5.3.4 monoT5

The potential of sequence-to-sequence models can be fostered for the ranking task by providing a query and a document as input and asking the model to decide if the document is relevant for this query by generating "true" or "false." The softmax of the generated token probability is then used as confidence for the predicted class to compute the final relevance of the document (Nogueira et al., 2020). The T5 (Raffel et al., 2020) model was fine-tuned in this fashion on the MS Marco passage retrieval dataset (Nguyen et al., 2016) as monoT5 by Pradeep et al. (2021). This model is then used in a second stage to rerank BM25 rankings and achieves great results, even as a pre-trained model on other datasets and domains (Pradeep et al., 2021).

The T5 model supports 512 sub-word tokens, and the LongEval dataset consists of documents with an average length of around 800 tokens. To avoid arbitrary truncation, the document retrieval task is formulated as a passage retrieval task and the top 1000 BM25 results are split into (still arbitrary but shorter) passages with an overlap half the size of the passage. By that, the whole document texts are reranked by monoT5. Further, the maximum relevance score of all passages from one document is used as the relevance score of the document for the final ranking.

For comparison and to avoid arbitrary sequences, the full documents are used instead as well. This approach seems reasonable since not too much text is cut off from the average document, and the title and introductions with high-level terms, similar to the query terms, are often located at the beginning of a document and are therefore captured by the model.

[16] https://github.com/terrierteam/pyterrier_colbert

### 5.3.5 Doc2Query

Instead of applying a language model at the reranking stage, Doc2Query (Cheriton, 2019) uses the T5 model to generate likely queries that a document could answer. These additional queries are then indexed along the document itself. By that, natural language queries can result in exact matches using traditional ranking functions and alleged relevant terms are boosted. This results in an advanced index that can be efficiently searched independent of methods.

The effectiveness is highly dependent on the number of queries that are added to the documents during indexing since this determines how much content is added. For this experiment, we used three and ten queries. While Nogueira and Lin (2019) used up to 80 queries, a maximum of ten queries were chosen to match the available resources.

### 5.3.6 E5

Recently L. Wang et al. (2022) achieved superior performance with the E5 model family. It is the first model that outperforms BM25 in a zero-shot retrieval setting on the BEIR (Thakur et al., 2021) benchmark. The performance is attributed to the large and high-quality dataset, the contrastive pre-training and the advanced fine-tuning process. The new paired dataset CCPairs (Thakur et al., 2021) of query passage pairs was used for training. It contains 1.3 billion query document pairs from Reddit, Wikipedia, Semantic-Scoolar, CommonCrawl and Stack Exchange and news websites.

The models $E5_{small}$ and $E5_{base}$ are used in a zero-shot fashion to create embeddings for all queries and documents. The documents are truncated at 512 sub-word tokens to fit in the model and not split into passages for efficiency. A Faiss[17] flat index was created from all embeddings, and L2 was used to score the query document similarity.

## 5.4 Evaluation

In the following, results for the initial experiments on the train slice of the WT subcollection are reported, and the submitted systems are analyzed. Then, the runs and results on the full dataset are described.

### 5.4.1 System Selection

Table 4 gives an extensive overview of the initial experiments. BM25 appeared to be a strong baseline, outperformed only by some systems and most often not statistically significant on all measures. The best runs of the different types were chosen for submission, also with the goal in mind to provide a diverse set of runs for the planned pooled gold annotation (Deveaud et al., 2023).

For the official ranking, we submitted to both sub-tasks the five systems:

1. RRF(BM25+Bo1-XSqrA_M-PL2) as `IRC_RRF(BM25+Bo1-XSqrA_M-PL2)`
2. BM25+colBERT as `IRC_BM25+colBERT`

[17] https://faiss.ai/

| System | MAP | Bpref | RR | P@20 | nDCG | nDCG@20 |
|---|---|---|---|---|---|---|
| BM25 | 0.1452 | 0.3245 | 0.2604 | 0.0654 | 0.2884 | 0.2087 |
| PL2 | 0.1408 | **0.3352** | 0.2572 | 0.0650 | 0.2884 | 0.2064 |
| TF-IDF | **0.1467** | 0.3259 | **0.2637** | 0.0660 | **0.2907** | **0.2109** |
| XSqrA_M | 0.1428 | 0.3265 | 0.2629 | **0.0633** | 0.2871 | 0.2042 |
| BM25+Bo1 | **0.1470** | **0.3341** | **0.2534** | **0.0661** | **0.2922** | **0.2075** |
| BM25+RM3 | 0.1426 | 0.3295 | 0.2408 | 0.0658 | 0.2867 | 0.2035 |
| RRF(BM25, XSqrA_M, PL2) | 0.1462 | 0.3380* | 0.2646 | 0.0656 | 0.2967* | 0.2101 |
| RRF(BM25+Bo1, XSqrA_M, PL2) | **0.1511** | 0.3466* | **0.2686** | 0.0673 | **0.3040*** | **0.2156** |
| RRF(BM25+RM3, XSqrA_M, PL2) | 0.1472 | **0.3472*** | 0.2589 | **0.0676** | 0.3008* | 0.2125 |
| BM25+passages+monoT5 | 0.1540 | 0.3369 | 0.2743 | 0.0708* | 0.2969 | 0.2196 |
| BM25+monoT5 | **0.1809*** | **0.3494*** | **0.3216*** | **0.0768*** | **0.3208*** | **0.249*** |
| d2q(3)>BM25 | 0.1578 | **0.3411** | 0.2630 | **0.0752*** | 0.2940 | 0.2284* |
| d2q(10)>BM25 | **0.1638*** | 0.3382 | **0.2862*** | 0.0707* | **0.3070*** | **0.2287*** |
| colBERT | 0.1652 | 0.3435 | 0.3045* | 0.0689 | 0.2989 | 0.2290 |
| BM25+colBERT | **0.1682*** | **0.3447** | **0.3046*** | **0.0692** | **0.3082*** | **0.231*** |
| E5_small | 0.1437 | 0.3265 | 0.2705 | 0.0619 | 0.2762 | 0.2039 |
| E5_base | **0.1545** | **0.3483** | **0.2826** | **0.0634** | **0.2910** | **0.2128** |

Table 4: Results on the train slice of the WT sub-collection. The best results per group are highlighted in **bold** and significant differences with Bonferroni correction to the BM25 baseline are denoted by an asterisk (∗).

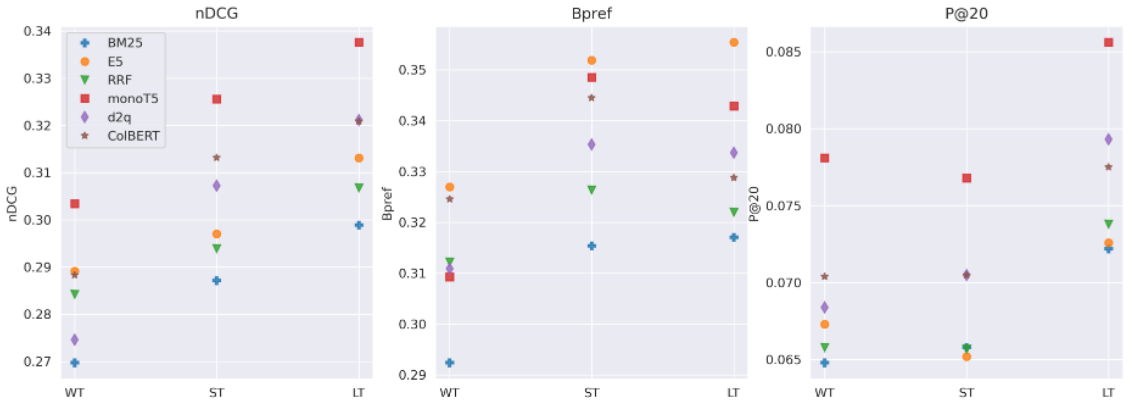


Figure 12: The ARP of nDCG (left), Bpref (center) and Recip Rank (right) from the submitted systems at WT, ST and LT.

3. BM25+monoT5 as `IRC_BM25+monoT5`
4. d2q(10)>BM25 as `IRC_d2q(10)>BM25`
5. $E5_{base}$ as `IRC_E5_base`

The BM25 baseline achieved an nDCG of 0.2884 on the WT train sub-collection slice. A MAP of 0.1452 is reported, but as initially shown in the data analysis in Section 5.2, only a few qrels per query are available; we relied on the Bpref (Buckley & Voorhees, 2004) measure instead. Here, a score of 0.3245 is achieved. Notably, compared to BM25, TF-IDF outperforms BM25 slightly but is not statistically significant. Regarding the runs with additional pseudo-relevance feedback, no significant improvements are made as well.

The RRF runs show the first significant improvements. The fusion run of the three runs BM25+Bo1, XSqrA_M and PL2 significantly outperform the BM25 baseline on MAP and nDCG by little. Larger improvements and the overall best results are achieved with BM25+monoT5. This run is significantly better on all measures and archives a 0.0324 higher nDCG. The passage retrieval version of the run performs considerably worse, similar to the baseline. The gap between the BM25 results on the two Doc2Query extended indexes is similar. While the results on the version with three additional queries per document make statistically no difference to the baseline, the results on the ten queries indexes are almost as good as the ones with BM25+monoT5 on all measures, except for P@20, which is even better. BM25+ColBERT performs slightly worse overall. Focusing on P@20, the system differs not from the baseline. Employing ColBERT as a first-stage ranker impairs the performance further. The results achieved with the E5 models as first-stage rankers are not significantly different from the baseline. Still, the base version outperforms the baseline in all measures, and the small version does on Bpref and RR.

### 5.4.2 Test Results

For the evaluation of the result, consistency is the main goal rather than high performance. The underlying assumption is that the system would continuously achieve the same performance. To evaluate this, the Result Delta ($\mathcal{R}_e\Delta$) between the averaged retrieval per-

formances at two different points in time is measured as proposed by Sáez, Mulhem, et al. (2021). The results are presented in Table 5 and visualized in Figure 12.

`IRC_RRF(BM25+Bo1-XSqrA_M-PL2):` The fused run contains at least 1000 results for all topics in the WT sub-collection. For the ST sub-collection the system could not find any documents for four queries. Namely the queries *to*, *a*, *the* and *the*[18] resulted in empty rankings. These queries consist only of stopwords, which leave an empty query string after query processing. These queries are most likely bad translations from the terms *verseau, argentique, nanterre* and *falloir*, mostly containing named entities. For the two LT sub-collection topics *cadreemploi* and *a*[19], no BM25 first stage ranking could be created. While *a* is again just a stopword, for the term *cadreemploi* no results were found, which might be due to a spelling error where actually the French job exchange website *cadremploi* might be mend. The topic *cadreemploi* is present exactly the same in the French queries.

The Average Retrieval Performance (ARP) — defined by the mean retrieval performance over multiple topics — improves slightly over time. In general, the deltas measured for the sub-collections are really small. The Δ nDCG between WT and ST is only -0.0097 and between WT and LT -0.0226.

`IRC_BM25+colBERT:` Based on the WT sub-collection for the topic *ducielalaterre*[20] no documents were found, and for all other topics, at least 1000 documents could be retrieved. Since ColBERT was employed as a reranker on top of BM25, the four topics *to*, *a*, *the* and *the*[21] still remain empty. For 28 other topics, only less than 1000 documents, ranging between three and 663, could be found. Like before, the LT sub-collection topics *cadreemploi* and the topic *a*[22] remain empty. For further 22 topics, less than 1000 results were found. For example, the fewest results were found for the topic *the audeau*.[23]

The ARP is increasing over time, as already observed for the RRF system. However, the differences are larger for this system. Between WT and ST the $\Delta nDCGs$ is -0.0249, and between WT and LT -0.0326.

`IRC_BM25+monoT5:` The composition of the runs stayed mostly the same for these runs. Since they also use BM25 as the first-stage ranking, the bottleneck for empty or short topic results remains.

As already observed on the train slice of the WT sub-collection, the ARP is the highest achieved on all measures and sub-collections compared to the other submitted systems, with small exceptions. One strong exception is the Bpref of only 0.3093 on the WT sub-collection, the smallest score achieved overall. However, the results are inconsistent, the deltas are higher, especially for Bpref.

`IRC_d2q(10)>BM25:` Through the document expansion with Doc2Query, at least 37 documents were found for the previously empty WT sub-collection topic *ducielalaterre*.[24] However, for the other sub-collections, the results stayed similar. Doc2Query performed weaker than initially on the train slice before, especially in comparison to monoT5. The

[18] LongEval ST qid: q072214697, q072222604, q072224942, q072212314
[19] LongEval LT qid: q0922511 and q092219105
[20] LongEval WT held out qid: q062216851
[21] LongEval ST qid: q072214697, q072222604, q072224942 and q072212314
[22] LongEval LT qid: q0922511, q092219105
[23] LongEval LT qid: q092220802
[24] LongEval WT held out qid: q062216851

| | | ARP | | | $\mathcal{R}_e\Delta$ | |
|---|---|---|---|---|---|---|
| | | WT | ST | LT | WT, ST | WT, LT |
| Bpref | BM25 | 0.2924 | 0.3154 | 0.3171 | -0.0230 | -0.0247 |
| | RRF | 0.3122 | 0.3264* | 0.3220 | **-0.0142** | -0.0098 |
| | ColBERT | 0.3246 | 0.3445* | 0.3288 | -0.0392 | -0.0336 |
| | monoT5 | 0.3093 | 0.3485* | 0.3429* | -0.0244 | -0.0228 |
| | d2q | 0.3109 | 0.3353* | 0.3337* | -0.0199 | **-0.0042** |
| | E5 | **0.3270** | **0.3519*** | **0.3554*** | -0.0249 | -0.0284 |
| P@20 | BM25 | 0.0648 | 0.0658 | 0.0722 | -0.0010 | -0.0074 |
| | RRF | 0.0658 | 0.0657 | 0.0738 | **0.0001** | -0.0080 |
| | ColBERT | 0.0704 | 0.0705* | 0.0775* | 0.0013 | -0.0075 |
| | monoT5 | **0.0781*** | **0.0768*** | **0.0856*** | -0.0021 | -0.0109 |
| | d2q | 0.0684 | 0.0705* | 0.0793* | **-0.0001** | -0.0071 |
| | E5 | 0.0673 | 0.0652 | 0.0726 | 0.0021 | **-0.0053** |
| nDCG | BM25 | 0.2697 | 0.2871 | 0.2989 | -0.0174 | -0.0292 |
| | RRF | 0.2842* | 0.2939* | 0.3068* | -0.0097 | **-0.0226** |
| | ColBERT | 0.2883 | 0.3132* | 0.3209* | -0.0222 | -0.0342 |
| | monoT5 | **0.3034** | **0.3256*** | **0.3376*** | -0.0326 | -0.0465 |
| | d2q | 0.2746 | 0.3072* | 0.3211* | -0.0249 | -0.0326 |
| | E5 | 0.2891 | 0.2970 | 0.3131 | **-0.0079** | -0.0240 |

Table 5: Results on the three (test) sub-collections as well as the deltas between them. The best system per measure and group is highlighted in **bold** and significant differences from the BM25 baseline are denoted with an asterisk*.

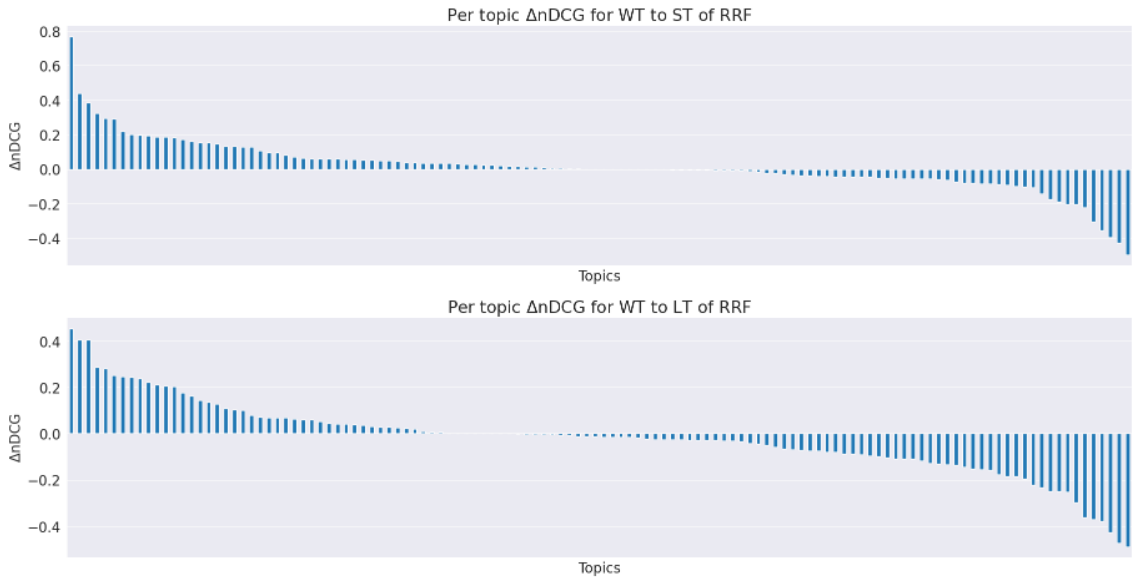


Figure 13: RRF $\Delta nDCG$ results per topic for WT to ST (top) and WT to LT (bottom). The topics are ordered according to the delta.

result deltas between WT and ST and WT and LT are among the highest for nDCG and P@20.

`IRC_E5_base`: Since the E5 model is based on k-NN and no stopwords were removed, for every topic, 1000 results were found. Compared to the train slice of the WT sub-collection, the system performed better. It achieved the highest Bpref on all three sub-collections and a high overall nDCG. The results are especially consistent between sub-collections with a $\Delta nDCG$ of 0.0079 between WT and ST and -0.0240 between WT and LT.

## 5.5 Temporal Persistence as Replicability

Building upon the result delta evaluation as introduced by Sáez, Mulhem, et al. (2021), we propose to use replicability measures to further investigate the environment effect on the systems. As described and implemented by Breuer et al. (2020), Breuer et al. (2021), the ARP may hide differences between the topic score distributions. For example, the RRF system achieved a high nDCG (0.28) at WT and is relatively stable considering the $\mathcal{R}_e\Delta(WT, ST)$ of 0.001. However, the per-topic results fluctuate between -0.4 and 0.8, as shown in Figure 13. For some topics, the retrieval performance improves, while the changes of the EE harms retrieval performance for other topics. We note that these circumstances require a more in-depth evaluation.

For a more detailed analysis of how the topic score distributions change, we cast the temporal comparison into a replication task, i.e., we evaluate the same set of systems on different data. Naturally, a direct comparison based on different sub-collections is difficult since it remains unclear if the observed effects should be attributed to the system or the changing EE. To overcome this problem, a pivot system similar as described by Sáez, Mulhem, et al. (2021) is used, and likewise, the experimental system is kept fixed in both EE. Effects are measured in comparison to this pivot system on one sub-collection and then compared to the same setup on a later sub-collection. To align the terminology, the

pivot system is a baseline run, BM25 for simplicity in this example, and the advanced run is the experimental system investigated.

In addition to the $\mathcal{R}_e\Delta$, as reported earlier in Table 5 we report the Effect Ratio (ER) and the Delta Relative Improvement ($\Delta$ RI). The ER (Breuer et al., 2020) is originally defined by the ratio between relative improvements of an advanced run over a baseline run. The relative improvements are based on the per-topic improvements, which are adapted for changing EEs as follows:

$$\Delta M_j^{EE_1} = M_j^{EE_1}(S) - M_j^{EE_1}(P), \Delta' M_j^{EE_2} = M_j^{EE_2}(S) - M_j^{EE_2}(P) \tag{15}$$

where $\Delta M_j^{EE_1}$ denotes the difference in terms of a measure $M$ between the pivot system $P$ and the experimental system $S$ for the $j$-th topic of the evaluation environment $EE_1$. Correspondingly, $\Delta' M_j^{EE_2}$ denotes the topic-wise improvement in the evaluation environment $EE_2$. The ER is then defined as:

$$\mathrm{ER}\big(\Delta' M^{EE_2}, \Delta M^{EE_1}\big) = \frac{\overline{\Delta' M^{EE_2}}}{\overline{\Delta M^{EE_1}}} = \frac{\frac{1}{n_{EE_2}} \sum_{j=1}^{n_{EE_2}} \Delta' M_j^{EE_2}}{\frac{1}{n_{EE_1}} \sum_{j=1}^{n_{EE_1}} \Delta M_j^{EE_1}}. \tag{16}$$

More specifically, the mean improvement per topic between the pivot and experimental system on one sub-collection (of $EE_1$) in comparison to the effect on the other sub-collection (of $EE_2$) is measured. Thereby, the ER is sensitive to the effect size. If the effect size is completely replicated in the second sub-collection, the ER is 1, i.e., the retrieval system is robust. If the ER is between 0 and 1, the effect is smaller, indicating a less robust system with performance drops. If the ER is larger than 1, the effect is larger, indicating performance gains caused by the change of the EE. Additionally, we include the $\Delta$ RI (Breuer et al., 2020), based on the relative improvements (RI) that are adapted to the LongEval definitions as follows:

$$\mathrm{RI} = \frac{\overline{M^{EE_1}(S) - M^{EE_1}(P)}}{\overline{M^{EE_1}(P)}}, \qquad \mathrm{RI}' = \frac{\overline{M^{EE_2}(S) - M^{EE_2}(P)}}{\overline{M^{EE_2}(P)}} \tag{17}$$

where $M^{EE}$ denotes the score of a measure $M$ determined with $EE$, and $S$ and $P$ denote the experimental and pivot system, respectively. The $\Delta$ RI is then defined as:

$$\Delta\mathrm{RI} = \mathrm{RI} - \mathrm{RI}'. \tag{18}$$

Therefore, a comparison between different sub-collections is straightforward. The ideal $\Delta$ RI of 0 is achieved if the RI is the same between both sub-collections, indicating a robust system. The more $\Delta$ RI deviates from 0, the less robust is the system, whereas negative scores indicate a more effective experimental system $S$ in the evaluation environment $EE_2$, and higher scores correspond to a less effective experimental systems than in the evaluation environment $EE_1$. All of the replicability measures were implemented with the help of `repro_eval` (Breuer et al., 2021), which is a dedicated reproducibility and replicability evaluation toolkit.

Even though the replicability measures do not necessarily require the same topics for each sub-collection, we harmonized the topics. Therefore, we only rely on the core queries that are shared between the sub-collections in this analysis. Given this methodology, the extended results are presented in Table 6. For all systems, the ARP decreases slightly at first (WT to ST) but increases in the long run (WT to LT) — a circumstance that is also reflected by the lower $\mathcal{R}_e\Delta$ scores for WT to ST compared to WT to LT.

The ER and $\Delta$ RI complement $\mathcal{R}_e\Delta$. For instance, monoT5 achieved similar P@20 scores on WT and ST, resulting in a $\mathcal{R}_e\Delta$ score of 0, which indicates perfect robustness in terms of $\mathcal{R}_e\Delta$. However, when comparing ER and also $\Delta$ RI, a more granular analysis is possible. In this case, the scores are close to but different from the perfect scores of 1 and 0, respectively, which would indicate perfect robustness. In general, the $\mathcal{R}_e\Delta$ scores do not always agree on the most robust system with ER and $\Delta$ RI. By these findings, we conclude that the replicability measures provide another perspective of the robustness, and we emphasize once again that it is also important to consider the topical variance over time.

Furthermore, we see that it is not enough to consider the differences of a single retrieval measure like nDCG. Depending on the evaluation measure, different systems perform best in terms of robustness. For instance, $\mathcal{R}_e\Delta$ of nDCG is lower for ColBERT and d2q than that of monoT5, while $\mathcal{R}_e\Delta$ of P@20 is lower for monoT5. Similarly, the replicability measures should be instantiated with different retrieval measures to get a more comprehensive understanding of robustness. While our RRF-based submissions achieve the best $\text{ER}_{\text{nDCG}}$ on both tasks, monoT5 is the most robust system in terms of $\text{ER}_{\text{P@20}}$. Likewise, ER and $\Delta$ RI identify different systems as the most robust for the same measures and tasks, which shows that it is insightful to evaluate both replicability measures.

In addition, we also included the p-values of unpaired tests based on the topic score distributions from different EE that were determined with the same experimental system as proposed in (Breuer et al., 2020). The general idea of these evaluations proposes to determine the quality of replicability (in our case, robustness) by the p-values and follows the assumption that lower p-values give a higher probability of failed replications or systems that are not robust. As can be seen, the highest p-values are achieved for the monoT5, ColBERT, or d2q, which generally agrees with our earlier observations.

The full potential of the ER and $\Delta$ RI can be seen if plotted against each other as in Figure 14. The closer the systems are located to the point (1, 0), the more persistent they are, with the preferable regions bottom right and top left. For the comparision WT to ST, the monoT5 system performs well on all three measures. However, the effect and the absolute scores are larger. The E5 system completely fails to replicate the absolute P@20 score and shows a generally larger difference. The RRF system, like most others, shows smaller absolute scores according to the $\Delta$ RI and a slightly decreased effect ratio. The plot regarding WT to LT shows more outliers with larger effect sizes for P@20 for the E5 system and Bpref for the d2q system. The systems are shifted to the top right of the plot, a trend similar to the increased $\mathcal{R}_e\Delta$ for WT to LT.

| | | ARP | | | $\mathcal{R}_e\Delta$ | | ER | | Δ RI | | p-val | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | System | WT | ST | LT | WT, ST | WT, LT | WT, ST | WT, LT | WT, ST | WT, LT | WT, ST | WT, LT |
| P@20 | BM25 | 0.070 | 0.067 | 0.085 | 0.002 | -0.015 | 1.000 | 1.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| | RRF | 0.075 | 0.069 | 0.088 | 0.006 | **-0.013** | 0.311 | 0.544 | 0.051 | 0.041 | 0.591 | 0.269 |
| | colBERT | 0.072 | 0.071 | 0.087 | 0.002 | -0.015 | 1.244 | 0.933 | **-0.011** | **0.009** | 0.875 | 0.190 |
| | monoT5 | **0.081** | **0.081** | **0.096** | **0.000** | -0.014 | **1.191** | **0.953** | -0.039 | 0.037 | **0.998** | 0.229 |
| | d2q | 0.079 | 0.072 | 0.091 | 0.007 | **-0.013** | 0.499 | 0.726 | 0.062 | 0.051 | 0.547 | **0.303** |
| | E5 | 0.071 | 0.066 | 0.088 | 0.005 | -0.017 | -1.452 | 2.903 | 0.040 | -0.022 | 0.616 | 0.125 |
| nDCG | BM25 | 0.269 | 0.272 | 0.306 | -0.003 | -0.037 | 1.000 | 1.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| | RRF | 0.285 | 0.282 | 0.314 | 0.003 | -0.030 | **0.925** | **0.786** | **0.003** | 0.013 | 0.945 | 0.227 |
| | colBERT | 0.276 | 0.275 | 0.297 | **0.001** | -0.021 | 0.441 | -1.198 | 0.015 | 0.053 | **0.967** | 0.412 |
| | monoT5 | **0.295** | **0.302** | 0.311 | -0.007 | **-0.015** | 1.146 | 0.187 | -0.013 | 0.083 | 0.817 | **0.580** |
| | d2q | 0.285 | 0.287 | **0.327** | **-0.001** | -0.042 | 0.916 | 1.317 | 0.006 | **-0.010** | 0.960 | 0.150 |
| | E5 | 0.290 | 0.300 | 0.313 | -0.010 | -0.023 | 1.333 | 0.362 | -0.025 | 0.054 | 0.720 | 0.382 |
| Bpref | BM25 | 0.314 | 0.314 | 0.324 | **-0.000** | -0.010 | 1.000 | 1.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| | RRF | 0.346 | 0.328 | 0.347 | 0.019 | -0.001 | 0.574 | **1.007** | 0.032 | **0.002** | 0.784 | 0.756 |
| | colBERT | 0.324 | 0.317 | 0.338 | 0.007 | -0.013 | 0.286 | 1.278 | 0.024 | -0.008 | 0.826 | 0.668 |
| | monoT5 | 0.337 | 0.344 | 0.337 | -0.007 | **0.000** | 1.261 | 0.553 | -0.019 | 0.034 | 0.850 | **0.997** |
| | d2q | 0.335 | 0.331 | 0.368 | 0.004 | -0.033 | **0.779** | 2.034 | **0.015** | -0.067 | **0.894** | 0.300 |
| | E5 | **0.368** | **0.354** | **0.371** | 0.014 | -0.003 | 0.738 | 0.863 | 0.045 | 0.028 | 0.692 | 0.931 |

Table 6: Extended results on the core queries, including the replicability measures.

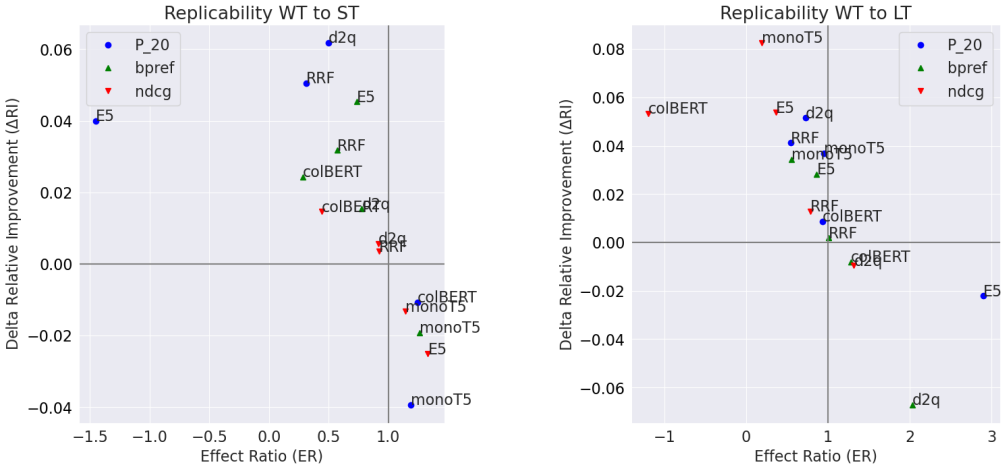


Figure 14: The ER plotted against the Δ RI for the replication WT to ST (left) and WT to LT (right).

## 5.6 Temporal Persistence as Reproducibility

For an even more detailed analysis of the topic score distributions over time, the temporal comparison is cast into a reproducibility task. The data is further harmonized so that the same documents are considered and the same systems are evaluated on this subset. This harmonization allows to directly compare runs of different points in time on up to document level persistence. Through this setup, the effect of updated documents on the systems is isolated.

For harmonization, the core queries are considered as before. Further, the documents are limited to around one million core documents, present in every sub-collection. The IDs of the documents and queries are unified in all runs and qrels, however, the actual content and the relevance labels remain untouched, so the change is still present in the harmonized collection. To avoid recomputing the runs for this comparison, all non-core documents, queries and qrels are removed. This leaves rankings with at least 281 documents, which still appears to be a reasonable length for the employed methods.

The Root Mean Square Error (RMSE) is used to evaluate the error between the effectiveness scores per topic of a system at one point in time compared to another point in time. In this context, we define the RMSE as:

$$RMSE(M^{EE_1}(S), M^{EE_2}(S)) = \sqrt{\frac{1}{n_{EE}} \sum_{j=1}^{n_{EE}} (M_j^{EE_1}(S) - M_j^{EE_2}(S))^2}. \tag{19}$$

The effectiveness, measured by a measure $M$, of a system $S$ is determined based on the two $EE$s. The difference between the two measures is averaged over the number of topics. Thereby, the RMSE requires that the topics of both EEs are the same but are not dependent on the documents or the length of the rankings. While with this measure the retrieval results of two different points in time are directly compared, the RMSE puts more weight on larger differences through squaring them (Breuer et al., 2020). The higher the RMSE is, the larger the error between the two runs. Therefore, a low RMSE denotes good persistence. Theoretically, the RMSE converges with increasing ranking length. In practice, stabilization can be observed earlier. Three modes of comparison are possible through variating the qrels for the relevance assessment. Either the qrels of the first point in time are used for evaluating both runs, or the qrels of the last period in time are used or both runs are evaluated on their corresponding qrels. The first and second mode assumes that the relevance remains, despite the document changes. Through this setup, the effect of the changing documents on the system is isolated. Therefore, a larger RMSE denotes that the system is influenced stronger by the updated documents. The last mode, where the runs are evaluated on the corresponding qrels, additionally considers changes in relevance. This again results in more affect interactions and is difficult to interpret.

On the highest level of persistence, the actual ranking position of the documents in two runs of two EEs can be considered through Kendall's $\tau$ (Kendall, 1949). It describes the agreement between the two rankings of documents and is sensitive to their ordering. As a requirement, as before for the RMSE, runs for the same topics are needed and also both

runs need to have the same length. The Kendall's $\tau$ is defined as:

$$\begin{aligned} \tau_j(S^{EE_1}, S^{EE_2}) &= \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}} \\ \overline{\tau}(S^{EE_1}, S^{EE_2}) &= \frac{1}{n_{EE}} \sum_{j=1}^{n_{EE}} \tau_j(S^{EE_1}, S^{EE_2}). \end{aligned} \tag{20}$$

Kendall's $\tau$ for a topic $j$ is calculated between the rankings of the system $S$ on two EEs, $EE_1$ and the proceeded $EE_2$. $P$ is the number of concordant pairs, i.e. document pairs that are ranked in the same order in both runs and $Q$ is the number of disconcordant pairs, ranked in opposite order. $U$ and $V$ are the ties in the rankings based on $EE_1$ and $EE_2$. Like Breuer et al. (2020) we relied on Kendall's $\tau$ Union, since result rankings are not necessarily permutations of the same list, meaning that one ranking might contain documents not present in the other. Kendall's $\tau$ Union denotes a perfect, document-wise, persistency at a value of 1. This is reached if all documents in both rankings are ordered the same. A Kendall's $\tau$ of -1 denotes a completely reversed ranking. Assuming rankings of the same system to be at least similar, with increased ranking length, the probability of rankings being similar decreases. It can be assumed that Kendall's $\tau$ Union decreases with it. Similar to the RMSE, through the proposed harmonization setup, the effect of the document updates on the rankings is observed directly.

To better account for runs with different documents and additionally weigh the ranking positions, Breuer et al. (2020) propose to use the Rank Biased Overlap (RBO) (Webber et al., 2010). Like Kendall's $\tau$, the RBO compares persistency on the document level. It is defined as:

$$\begin{aligned} RBO_j(S^{EE_1}, S^{EE_2}) &= (1 - \phi) \sum_{i=1}^{\infty} \phi^{i-1} \cdot A_i \\ \overline{RBO}(S^{EE_1}, S^{EE_2}) &= \frac{1}{n_{EE}} \sum_{j=1}^{n_{EE}} RBO_j(S^{EE_1}, S^{EE_2}) \end{aligned} \tag{21}$$

where the RBO is calculated by topic and then averaged over all topics. The parameter $\phi$ is bound between 0 and 1 and adjusts the weighting of the rank. The smaller $\phi$ is chosen, the higher the top ranks are weighted. $A_i$ is the overlap between the two rankings up to rank $i$, which can be formulized as $|S^{EE_1}_{:i} \cap S^{EE_2}_{:i}|$. The higher the RBO is, the more similar the two rankings are. Likewise, the $\overline{RBO}$ is the average of all topic vise RBOs and summarizes the document level similarity.

In table 7 and 8, the reproducibility results are reported additionally to the ARP and replicability measures. All results are again implemented through `repro_eval` (Breuer et al., 2021). The results generally do not agree with the ones achieved previously in the other evaluations due to the additional harmonization steps. The ranking of systems differs between sub-collections and measures. Considering the completely harmonized results, a generally high consistency is measured between WT and ST and only a little less between WT and LT. This highlights the impact of the newly added documents compared to the replicability evaluation. These documents, not present in the reproducibility evaluation anymore, have a strong influence on the result delta $\mathcal{R}_e\Delta$. While previously the results

| | | WT | ST | $\mathcal{R}_e\Delta$ | ER | ΔRI | p-val | RMSE | Kendall's $\tau$ | RBO |
|---|---|---|---|---|---|---|---|---|---|---|
| Bpref | BM25 | 0.308 | 0.304 | **0.004** | 1.000 | 0.000 | 1.000 | 0.032 | **0.268** | **0.952** |
| | RRF | 0.347 | 0.319 | 0.027 | **1.053** | **-0.006** | 0.489 | 0.096 | 0.153 | 0.874 |
| | colBERT | 0.337 | 0.313 | 0.024 | 0.670 | 0.032 | 0.502 | 0.042 | 0.117 | 0.905 |
| | monoT5 | 0.350 | 0.333 | 0.017 | 0.910 | 0.013 | 0.659 | **0.030** | 0.260 | 0.949 |
| | d2q | 0.332 | 0.322 | 0.010 | 1.550 | -0.043 | **0.768** | 0.093 | 0.120 | 0.817 |
| | E5 | **0.357** | **0.345** | 0.012 | 1.061 | -0.009 | 0.749 | 0.045 | 0.207 | 0.930 |
| P@20 | BM25 | 0.074 | 0.064 | 0.010 | 1.000 | 0.000 | 1.000 | **0.006** | **0.268** | **0.952** |
| | RRF | 0.077 | 0.068 | 0.009 | **1.000** | **0.000** | 0.420 | 0.009 | 0.153 | 0.874 |
| | colBERT | 0.071 | 0.067 | **0.004** | 0.500 | -0.017 | **0.656** | 0.008 | 0.117 | 0.905 |
| | monoT5 | **0.082** | **0.075** | 0.007 | 1.050 | -0.006 | 0.552 | 0.010 | 0.260 | 0.949 |
| | d2q | 0.080 | 0.071 | 0.009 | 1.188 | -0.017 | 0.416 | 0.012 | 0.120 | 0.817 |
| | E5 | 0.073 | 0.065 | 0.008 | **1.000** | **-0.000** | 0.413 | 0.011 | 0.207 | 0.930 |
| nDCG | BM25 | 0.282 | 0.268 | 0.014 | 1.000 | 0.000 | 1.000 | **0.013** | **0.268** | **0.952** |
| | RRF | 0.305 | 0.287 | 0.018 | **1.074** | **-0.006** | 0.597 | 0.062 | 0.153 | 0.874 |
| | colBERT | 0.287 | 0.276 | 0.011 | 0.399 | 0.012 | 0.708 | 0.024 | 0.117 | 0.905 |
| | monoT5 | **0.315** | **0.306** | 0.009 | 0.895 | 0.013 | 0.788 | 0.020 | 0.260 | 0.949 |
| | d2q | 0.290 | 0.281 | 0.009 | 1.496 | -0.015 | 0.780 | 0.056 | 0.120 | 0.817 |
| | E5 | 0.302 | 0.299 | **0.002** | 1.082 | **-0.006** | 0.934 | 0.021 | 0.207 | 0.930 |

Table 7: Reproducibility results WT to ST.

| | | WT | LT | $\mathcal{R}_e\Delta$ | ER | ΔRI | p-val | RMSE | Kendall's $\tau$ | RBO |
|---|---|---|---|---|---|---|---|---|---|---|
| Bpref | BM25 | 0.308 | 0.326 | 0.018 | 1.000 | 0.000 | 1.000 | 0.100 | **0.186** | **0.931** |
| | RRF | 0.347 | 0.349 | 0.002 | 0.744 | 0.034 | 0.918 | **0.096** | 0.153 | 0.874 |
| | colBERT | 0.337 | 0.334 | 0.003 | **0.950** | **0.008** | 0.929 | 0.230 | 0.025 | 0.168 |
| | monoT5 | 0.350 | 0.330 | 0.020 | 0.661 | 0.050 | 0.588 | 0.212 | 0.040 | 0.649 |
| | d2q | 0.332 | **0.375** | 0.043 | 1.144 | **-0.008** | 0.213 | 0.202 | 0.044 | 0.454 |
| | E5 | **0.357** | 0.357 | **0.000** | 0.562 | 0.074 | **0.995** | 0.302 | 0.027 | 0.162 |
| P@20 | BM25 | 0.074 | 0.077 | 0.003 | 1.000 | 0.000 | 1.000 | 0.015 | **0.186** | **0.931** |
| | RRF | 0.077 | 0.082 | 0.005 | 1.625 | -0.030 | 0.684 | **0.009** | 0.153 | 0.874 |
| | colBERT | 0.071 | 0.075 | 0.003 | -2.167 | -0.108 | 0.749 | 0.052 | 0.025 | 0.168 |
| | monoT5 | **0.082** | **0.087** | 0.005 | 0.650 | 0.037 | 0.662 | 0.046 | 0.040 | 0.649 |
| | d2q | 0.080 | 0.081 | **0.001** | **0.813** | **0.015** | **0.926** | 0.034 | 0.044 | 0.454 |
| | E5 | 0.073 | 0.079 | 0.006 | -6.500 | -0.085 | 0.602 | 0.050 | 0.027 | 0.162 |
| nDCG | BM25 | 0.282 | 0.310 | 0.028 | 1.000 | 0.000 | 1.000 | **0.041** | **0.186** | **0.931** |
| | RRF | 0.305 | 0.321 | 0.016 | **0.688** | 0.026 | 0.585 | 0.062 | 0.153 | 0.874 |
| | colBERT | 0.287 | 0.295 | 0.008 | 2.623 | -0.032 | 0.782 | 0.182 | 0.025 | 0.168 |
| | monoT5 | **0.315** | 0.311 | **0.004** | 0.456 | 0.067 | **0.903** | 0.166 | 0.040 | 0.649 |
| | d2q | 0.290 | **0.334** | 0.044 | 1.789 | -0.022 | 0.173 | 0.108 | 0.044 | 0.454 |
| | E5 | 0.302 | 0.313 | 0.011 | 0.769 | **0.018** | 0.692 | 0.183 | 0.027 | 0.162 |

Table 8: Reproducibility results WT to LT.

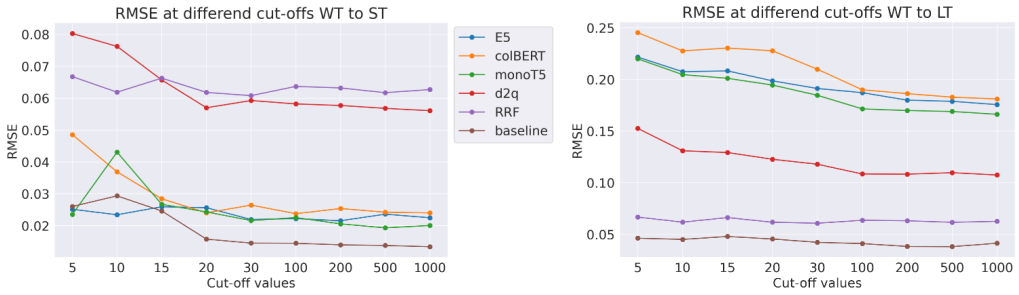


Figure 15: RMSE of nDCG between WT-ST (left) and WT-LT (right).

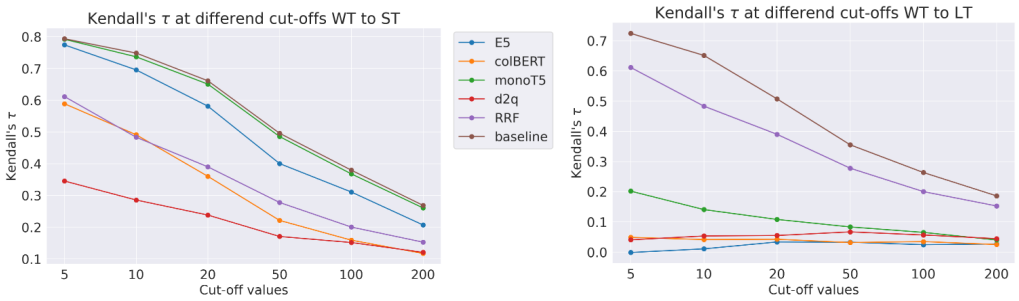


Figure 16: Kendall's $\tau$ between WT-ST (left) and WT-LT (right).

were improving over time, they are now staggering, as seen by the mainly positive $\mathcal{R}_e\Delta$ values.

On the fully harmonized dataset, the monoT5 system achieved the second lowest $\mathcal{R}_e\Delta$ considering P@20 on the WT to ST comparison. The ER and $\Delta$RI scores for P@20 are again close to perfect and thereby agree with the $\mathcal{R}_e\Delta$. However, the RMSE only partially agrees with the replicability measures, as BM25, ColBERT and RRF achieve a lower error. Considering the nDCG-based measures, monoT5 and d2q achieve the same, second lowest $\mathcal{R}_e\Delta$. However, the ER is really different and shows a divergence in different directions. This is also reflected in the RMSE measure, which is the second lowest for monoT5 and the second highest for d2q and is probably caused by the RMSE that penalizes larger errors stronger. The RMSE based on the effectiveness as measured by the nDCG is visualized over different ranking lengths in Figure 15. Comparing the plots from both points in time, the RMSE appears to be generally higher for the LT point in time. This is in line with the observation that the effects increase over time. While at ST errors up to 0.08 are measured, the errors on the LT sub-collection reach almost 0.25. The ranking of the systems is changing, except for the BM25 baseline, which retains the lowest RMSE, at least after a minimum ranking length of 15. This strengthens the choice of the system as baseline. The two systems RRF and d2q gain less RMSE from ST to LT and therefore appear to be more resilient in this experiment. In contrast, the systems colBERT, monoT5 and E5 show a lower RMSE at ST but gain strongly at LT. Most runs are relatively unstable at shorter rankings but stabilize with increasing length. At ST, a plateau is mostly met at a length of 15 documents. The effect is less harsh at LT where a slight plateau appears at around 100 documents.

By directly comparing the rankings through Kendall's $\tau$ Union, the differences are observed independently from the effectiveness. Figure 16 shows the correlation over different ranking lengths for both sub-collections. Similarly, as observed through the RMSE, effects increase from ST to LT, as shown by a generally reduced correlation. RRF and BM25 best retain the correlation between ST and LT. The measures do not necessarily agree with the replicability measures. For example, monoT5 has the second highest Kendall's $\tau$ but ranks lower considering the Bpref based measures.

This shows that even if the effectiveness is not changing much, a different ranking is presented to the user if searched with the same query at a later point in time. This behavior bears different implications for different types of queries.

## 5.7 Conclusion and Outlook

In this work, we described our participation in the LongEval Lab at CLEF 2023. As core contribution, we applied five advanced retrieval systems to the LongEval dataset and submitted the runs to both sub-tasks. As this is a new challenge, the interpretation of the results is difficult. Overall, the results for the different systems are very similar. The measured differences are statistically significant but appear small as compared to the same methods on different datasets as listed on the IR experiment platform[25] (Fröbe et al., 2023). Interestingly, an increasing ARP over time was observed for most systems and measures. Still, the performance difference, measured by $\mathcal{R}_e\Delta$, is smaller for WT to ST compared to WT to LT, which complies with the natural assumption that persistence deteriorates over time.

Further, we reported preliminary results applying replicability and reproducibility measures to quantify temporal persistence, an extension to common practices of these measures and their interpretation (Maistro et al., 2023). Through the harmonization of distinct EE components, the effect on them could be narrowed down. It was shown that the results based on different measures and likewise for different topics do not necessarily agree with each other. Therefore, we see great potential in using replicability and reproducibility measures to gain further insights into robustness. As a first validation, we saw similarities between these measures and the result deltas. All in all, a strong environment effect on the systems was shown and could be analyzed.

Future work will be including the selection of the pivot system and qualitative core queries. Also, further harmonizing the dataset by unifying the document IDs would allow us to cast the problem as a reproducibility task and investigate persistence on an even more specific level with reproducibility measures.

[25] https://www.tira.io/task/ir-benchmarks

# 6 Toward a Continuous Evaluation Framework

As initially stated, the main current evaluation methodologies are not temporally consistent by default. User studies tend not to be repeatable (Balog & Zhai, 2023; Tan et al., 2017), a necessity for temporal persistence. Test collections minimize temporal changes by abstraction but are sometimes outdated shortly after creation (Soboroff, 2006). As a reaction, the duration of an A/B test is intentionally kept as short as possible to minimize temporal effects (Kohavi, 2015). Additional efforts are needed to evaluate an IR system temporally reliably. But since evaluations are generally expensive, complete re-evaluations are rarely feasible, even if technically possible.

The components in an EE vary strongly and therefore require different evaluation methods. No method is capable of taking all components of an EE into account. Likewise, conventional evaluation methods have different strengths and weaknesses and take different assumptions on the EE. Test collections focus on the performance of the systems but thereby abstract the user. Therefore, it is observed that improved effectiveness does not necessarily correlate with actual benefits for the user (Turpin & Scholer, 2006). In contrast, user studies, which can evaluate the users' utility are hardly repeatable and very expensive (Kelly, 2009).

This shows that a single evaluation method has limited explanation power. Instead, a combination of methods should be employed to gain a good understanding of an IR system. Hofmann et al. (2016) propose to combine online and offline evaluations in an end-to-end evaluation. This is essentially what a continuous evaluation framework should achieve. While it may appear as an ambitious goal, it becomes feasible through synergies between the methods.

Schaer, Castro, et al. (2021) propose to explicitly formulate continuous evaluation in an end-to-end framework as visualized in Figure 17. Besides the IR problem, a continuous evaluation module is added, comprising test collections, user interactions and simulations. Through active users that use the system, interaction data can be logged. Further, expert users, who are domain experts that interact with the same system with a slightly modified user interface, can make annotations that are logged as relevance assessments.

Through the integration of the three evaluation methods, the spectrum of evaluations is made more continuous. This improves the coverage of the components in the EE that are recognized during evaluation. Test collections are suitable to assess the corpus components documents and the content topics. Further, some conclusions about the information need, which is explicitly defined in test collections, can be made. User interactions reflect on components closer to the user like the queries and topics. Latent relevance indicators can be estimated based on such data. Likewise, simulations can assess the user-centered components, but in a repeatable manner.

Considering all of this, it is especially important to be precise about which questions to ask and which methods to employ for answering them. Voorhees (2019) precisely formulated this as:

> "It obviously does no good to abstract an evaluation task to the point where test results do not reflect performance on the real task of interest; it is equally

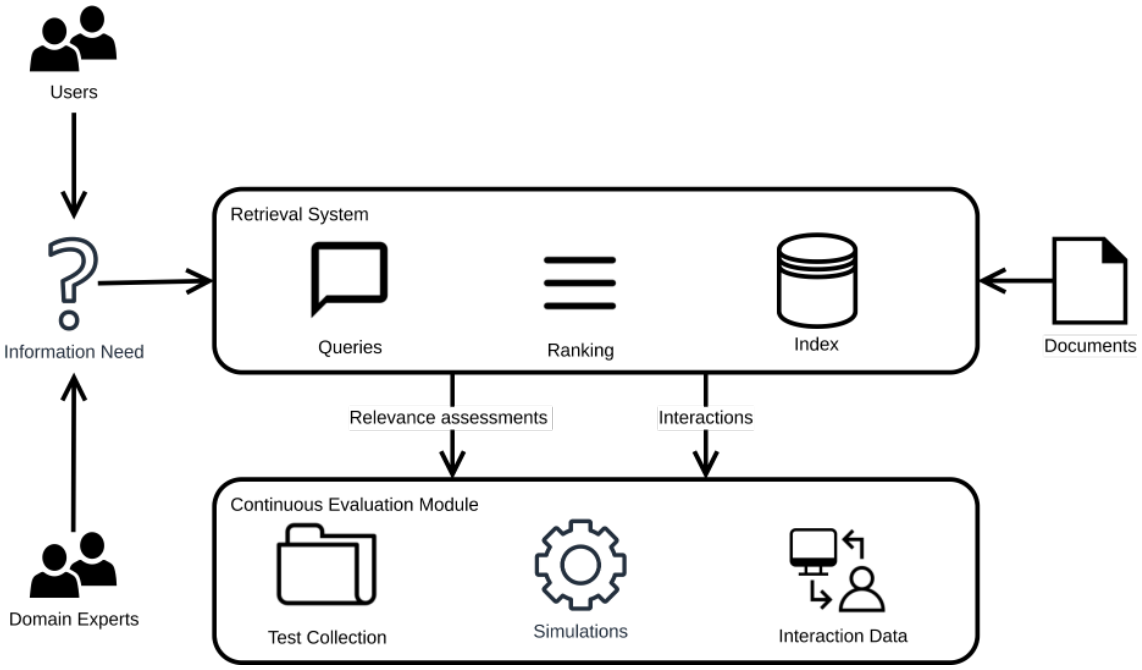


Figure 17: Schematic visualization of the continuous evaluation framework, reproduced and adapted from (Schaer, Castro, et al., 2021).

> as unhelpful to include any operational variable that might possibly influence outcomes since generalization then becomes impossible and nothing is learned."

Each evaluation methodology makes different abstractions to improve generalizability. A continuous evaluation framework needs to bear this in mind, and purposefully use methods to compensate abstractions, to achieve a holistic evaluation of a system. Further focus is drawn to the different evaluation methods in the continuous evaluation module and some synergies they can profit from.

## 6.1 Test Collections

Creating a test collection is an expensive endeavor and with growing size, demanded by deep learning approaches, increasingly difficult (Hashemi et al., 2016). Efficient methods are needed to retain feasibility. Traditional pools demand many systems to avoid a system bias (Soboroff, 2006). Beyond that, manual runs are especially valuable for the pools (Jayasinghe et al., 2014). While different systems may not necessarily be available in the continuous evaluation context, manual runs can be created through the expert users, either actively through their curation or passively based on search logs. The sampling strategy can be further optimized by sampling documents for pools not based on retrieval systems but through statistical sampling. Different strategies exist that create pools more efficiently and with less bias. For example, dynamic sampling employs active learning strategies to sample documents for further evaluation (Cormack et al., 2019). Other approaches are active sampling (D. Li & Kanoulas, 2017) or sampling with pseudo relevance feedback (Otero et al., 2023).

By iteratively and actively choosing which documents should be selected for the pool, the effort can be distributed over time, much like proposed by Tonon et al. (2015) and user

interactions, probably from expert users, could be integrated into the process. Further, this would allow to design the test collection specifically for the systems to be evaluated.

After selecting documents for the pool, these need to be annotated. Sakai et al. (2023) found that the order in which the documents are annotated is not affecting the efficiency, although it affects the ranking bias. Further, it seems to be more important that the annotators are well-trained and motivated than that they create the topics by themselves as gold annotators. The findings about the annotation efficiency imply that for continuous evaluations also bronze annotators could be employed instead of domain experts if they are well motivated for the task. These might be more accessible and available on demand through crowd-working platforms.

The goal is to achieve a stable evaluation with as little annotation effort as possible. However, the number of topics depends on its capability to evaluate the system (Roitero et al., 2018). Soboroff (2006) stated that it is better to have more topics with few qrels than to have fewer topics with many qrels.

Still, many domains evolve so quickly that a test collection becomes outdated fast, as we saw in Section 3. In these cases, Soboroff (2006) recommends creating dynamic test collections which need to be actively maintained. Doing so involves measuring how well the test collection is capable of evaluating systems at the moment, actively maintaining it and using appropriate evaluation measures like bpref. To assess the reusability of a test collection, Soboroff (2006) recommends different measures similar to the ones proposed in Section 4. Further, Hashemi et al. (2016) perform leave one out analysis to estimate the reusability. Runs, teams or topics are excluded from a test collection and the results are compared. If the results do not change significantly, the test collection is stable. If deficits are found in the collection, for example, relevant documents are not available or have changed substantially, or the ranking of it appears to be not stable anymore, the test collection needs to be adjusted.

Buckley and Voorhees (2004) propose to directly include a test collection in a larger collection and use bpref to evaluate systems that retrieve from the combined collection. Such an embedded collection is directly applicable to the continuous evaluation framework. Either, by building the embedded collection as described before, or indexing a publicly available test collection that fits the domain and purpose of the system.

## 6.2 Interaction Data

Interaction data can mainly be gathered through two experiment types. Either through artificially created experimental environments in offline Interactive Information Retrieval (IIR) or by observing users in production environments in online experiments. IIR studies are employed to evaluate specific features of a system or interface that should support the search process. While the effectiveness of the system may play a role in the evaluation, it is often only secondary. The search behavior of the user is paramount for these experiments (Kelly, 2009). These studies are often expensive because test users need to be employed and specific environments that can record the user interactions and reactions are required. Further, through learning and memorization effects, the employed subjects may need to be replaced after tests (Kelly, 2009).

In contrast, online evaluations rely on users of a production system for evaluation. They need a fully functional system and use the implicit user interactions as relevance feedback. Since the evaluation takes place in a real environment and users are not necessarily aware of the evaluation, realistic results can be achivieved (Hofmann et al., 2016). If a sufficient number of users is available, their behavior can be logged and analyzed. The system can be evaluated directly as an absolute evaluation or in relation to another system. In such A/B tests, different systems or different versions of a system are presented to separate user groups (Kohavi, 2015). The challenge is to interpret the user interactions as they are only latent relevance indicators. While clicks or dwell time are more noisy, a purchase for example is a more direct indicator of relevance (Hofmann et al., 2016).

While online experiments are powerful evaluations to measure the utility for the users, a production environment is needed with sufficient traffic. Thereby, this experiment type is the defacto standard in the industry but rarely affordable in academia (Hofmann et al., 2016). Living labs try to bridge this gap by providing access to live systems, often in shared tasks. Researchers are invited to provide systems that are deployed in production environments and evaluated on real users (Hopfgartner et al., 2019; Schaer, Breuer, et al., 2021). To exploit the traffic more efficiently, often rankings are merged through interleaving. By that, two systems are compared directly and receive interactions from one user in parallel. Here as well relevance indicators are rarely explicit and the interleaving may require further correction (Kürsten, 2012).

Evaluations based on interaction data focus on user satisfaction, which is the underlying main goal in many retrieval systems. Test collection based evaluations clearly come to an extent since they abstract the user. Online evaluations however seem to be the best way to complement the system focussed evaluations through test collections with a user perspective in a continuous evaluation framework. Having interaction data at hand, even if it is limited in quantity, can greatly improve test collection creation and ease maintenance, as described before. In return, test collections can be used to pre-test experimental systems before they are submitted to an online evaluation to avoid exposing real users to malfunctioning systems (Keller & Munz, 2022). Further, the logs of living labs are not necessarily directly re-usable. However, they may be sufficient to contribute to the creation of reusable components (Tan et al., 2017).

## 6.3 Simulations

User simulations try to simulate user behavior instead of relying on real users. The great advantage is that these simulations are repeatable, which is not the case with the interaction-based methods described earlier. Further, at least in the long run, they are more affordable. By simulating user interactions that can be re-applied, simulations connect interaction evaluations with test collection ones. Simulations are differentiated into model-based and data-driven simulations. Model-based simulations assume a user model which is based on rules or probability. The parameters are set heuristically or are estimated. In contrast, data-driven models are supervised learned from data like logs of interactions. Based on such simulated user models, different user-related components of the EE can be simulated.

For example, the queries a user would create given an information need results in, clicks on ranked results or illustrate the general browsing beahaviour (Balog & Zhai, 2023).

A user model represents a hypothesis of a user. This hypothesis can then be falsified based on interaction logs. By that, user models are validated and improved, which ultimately leads to improved evaluation measures. To create such models, information is needed that makes them distinct, e.g. the system used, the users that interact with the system or the task the users try to fullfill (Balog & Zhai, 2023). For directed evaluations, these could be derived from real interaction data in the continuous evaluation module with further differentiation between expert and normal users. Likewise, the data needed to actually fit the user models. The data demand thereby depends on the type of simulation. Breuer (2023) investigates the demand for click models in low resource settings and finds that on average 20 logged sessions per topic are required for a stable estimation. While this might be sparse for A/B tests, simulations might provide a good layer of abstraction to efficiently exploit interaction data, for example from the head queries logged.

After creating user models they can be used to conduct user-centered Evaluations. In these evaluations often the relative order of systems is assessed (Balog & Zhai, 2023). Since the user models can be re-applied, they can be used on different EEs at different points in time, almost like a user study on demand. Great power comes through parameterizing the user models. By that, it can be explored how nuanced user variations influence different components in the EE and the IR system (Balog & Zhai, 2023). Further, temporal parameterization might be possible by curating the foundational sessions.

Concerning the continuous evaluation module, simulations can function as the connecting link between test collections and interaction data. Simulations can augment static test collections by replacing the qrels with user simulations and by that reduce the abstraction of the user (Balog & Zhai, 2023). They can be used to pre-assess the utility of a system for the users before conducting online experiments and do so continuously over time. Vice versa, interaction data from online experiments may be sustainably used in simulations.

## 6.4 Summary

As initial steps toward a continuous evaluation framework, the interplay between different components is theoretically outlined. Special attention is drawn to synergies between evaluation methods, to reduce cost and increase the efficiency of evaluations. This is especially important since monitoring temporal consistency requires continuous evaluation. While this analysis is not exhaustive, the potential of continuous evaluations is estimated as an outlook for future work.

We found that through combining evaluation methods, their strengths and weaknesses could be used beneficially to gain a better understanding of the capabilities of IR systems. Test collections can be used to evaluate the performance of a system and online evaluations of the utility for the user. Simulations function as the link between both and help to make evaluations repeatable. Through employing multiple methods, the evaluation spectrum becomes more continuous – across methods and also across time. The concrete design of a framework like the one outlined is left for the future, requiring further investigation.

# 7 Conclusion

In this work, the evolving environment that surrounds and influences an IR system during evaluation is investigated. Initially, information retrieval and evaluations in the field are described as these are the theoretical foundation for this work using the example of test collections. The evolving evaluation environment is investigated in a systematic literature review. Eleven components, together with how they change over time and how these changes can be measured are, identified. Based on this evidence, the EE is further specified and arranged so that it can be used to precisely differentiate and locate effects. The findings are applied in practice to the LongEval shared task by adapting reproducibility and replicability experiments to precisely measure the effect of changes in the documents and qrels. Additionally, five retrieval systems were submitted for longitudinal evaluation. In a discussion, the continuity spectrum was extended to the evaluation methods. Synergies between methods were shown that make a continuous evaluation framework feasible.

## 7.1 Contributions

Through the extensive analysis of the EE we draw the following conclusions that let us answer the research questions initially raised. The first question **RQ1** *How is the environment of an IR system evolving?* was answered through the literature review and the specification of the EE. The eleven components *language, corpus, documents, content type, topic, information need, queries, user, relevance, results* and *system* are identified as changing. The query, corpus and results are identified as main general components in regard to the ones in the IR problem. The remaining components can be allocated to these components except the language, which influences all of them. The changes are grouped by time frame and aggregation. Many of them are the CRUD operations on the different components, others represent fluctuating quantities. Especially the components that are expressed in texts can often be captured by similar measures, based on deltas or similarity. Besides that, user and relevance measures are gathered. The quantity of the changing components presents a concern. To be able to identify the components and changes is an important prerequisite to factor these changes during evaluation.

The second research question **RQ2** *How can the evolving environment of an IR system be considered during effectiveness evaluation?* asks to factor these changes. By measuring how the components change, the first conclusions can be taken on how this might influence the effectiveness of the system. Since the changes often overlap each other, it is difficult to isolate them precisely. Based on the LongEval dataset, the temporal consistency of retrieval results is measured through result deltas. To gain a better understanding, replicability and reproducibility measures are employed. It is shown how these measures can point to distinct influences of components from the EE. The general correlation of these measures with the result deltas and the average effectiveness can be interpreted as a first validation. Finally, the theoretical outline of a continuous evaluation framework shows how multiple evaluation methods should be employed, to get a holistic evaluation of an IR system.

The literature reviewed shows that not much research in IR is focussing on the temporal influence on evaluations. Still, a general concern appears to be present. Longitudinal

evaluations ask questions on how reliable the evaluation methods are, which are considered to be the standard in academia and industry. This makes these investigations all the more important.

## 7.2 Future Work

After creating a first methodology to continuously evaluate IR systems, future work will need to substantiate it. The obvious next step would be to apply the EE to different datasets for validation. The gathered measures should be further specified and tested in practice on more collections. This would provide a better overview of the landscape of datasets. The method for measuring the impact of a changing EE on IR systems needs to be validated on a larger scale. Unfortunately, datasets that specifically focus on longitudinal evaluations are rare. However, it may be possible to employ other datasets that provide metadata for some components, which can be split into temporal sub-collections. Additional methods are also needed to investigate the impact of further components on the effectiveness of the systems. This will help to better isolate effects and thereby estimate their importance. Combining all of that will help to design a continuous evaluation framework which can be seen as the long-term goal. While temporal persistence appears to be a general concern, measuring the seriosity of it should guide future work.

# References

Adar, E., Teevan, J., Dumais, S. T., & Elsas, J. L. (2009). The web changes everything: Understanding the dynamics of web content. In R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, & B. B. Cambazoglu (Eds.), *Proceedings of the second international conference on web search and web data mining, WSDM 2009, barcelona, spain, february 9-11, 2009* (pp. 282–291). ACM. https://doi.org/10.1145/1498759.1498837

Aggarwal, K., Theocharous, G., & Rao, A. B. (2020). Dynamic clustering with discrete time event prediction. In J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, & Y. Liu (Eds.), *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020* (pp. 1501–1504). ACM. https://doi.org/10.1145/3397271.3401182

Aliannejadi, M., Chakraborty, M., Ríssola, E. A., & Crestani, F. (2020). Harnessing evolution of multi-turn conversations for effective answer retrieval. In H. L. O'Brien, L. Freund, I. Arapakis, O. Hoeber, & I. Lopatovska (Eds.), *CHIIR '20: Conference on human information interaction and retrieval, vancouver, BC, canada, march 14-18, 2020* (pp. 33–42). ACM. https://doi.org/10.1145/3343413.3377968

Alkhalifa, R., Bilal, I., Borkakoty, H., Camacho-Collados, J., Deveaud, R., El-Ebshihy, A., Espinosa-Anke, L., Gonzalez-Saez, G., Galuscakova, P., Goeuriot, L., Kochkina, E., Liakata, M., Loureiro, D., Madabushi, H. T., Mulhem, P., Piroi, F., Popel, M., Servan, C., & Zubiaga, A. (2023–September 23). Overview of the CLEF-2023 LongEval lab on longitudinal evaluation of model performance. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023).*

Alonso, O. (2013). *Stuff happens continuously: Exploring web contents with temporal information*. https://doi.org/10.1145/2487788.2488122

Altingövde, I. S., Ozcan, R., & Ulusoy, Ö. (2011). Evolution of web search results within years. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yates, T.-S. Chua, & W. B. Croft (Eds.), *Proceeding of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2011, beijing, china, july 25-29, 2011* (pp. 1237–1238). ACM. https://doi.org/10.1145/2009916.2010137

Amati, G. (2003). *Probability models for information retrieval based on divergence from randomness* (Doctoral dissertation). University of Glasgow, UK. http://theses.gla.ac.uk/1570/

Amati, G. (2006). Frequentist and bayesian approach to information retrieval. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in information retrieval, 28th european conference on IR research, ECIR 2006, london, UK, april 10-12, 2006, proceedings* (pp. 13–24). Springer. https://doi.org/10.1007/11735106_3

Baeza-Yates, R., & Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England. http://www.mir2ed.org/

Balog, K., & Zhai, C. (2023). User simulation for evaluating information access systems. *CoRR*, *abs/2306.08550*. https://doi.org/10.48550/arXiv.2306.08550

Bar-Ilan, J. (2002). Criteria for evaluating information retrieval systems in highly dynamic environments. In M. Levene & A. Poulovassilis (Eds.), *Proceedings of the second international workshop on web dynamics, WebDyn@WWW 2002, honululu, HW, USA, may 7, 2002* (pp. 70–77). CEUR-WS.org. https://ceur-ws.org/Vol-702/paper7.pdf

Bassani, E., & Romelli, L. (2022). Ranx.fuse: A python library for metasearch. In M. A. Hasan & L. Xiong (Eds.), *Proceedings of the 31st ACM international conference on information & knowledge management, atlanta, GA, USA, october 17-21, 2022* (pp. 4808–4812). ACM. https://doi.org/10.1145/3511808.3557207

Breuer, T. (2023, June 9). *Reproducible Information Retrieval Research: From Principled System-Oriented Evaluations Towards User-Oriented Experimentation*. DuEPublico: Duisburg-Essen Publications online, University of Duisburg-Essen, Germany. https://doi.org/10.17185/DUEPUBLICO/78449

Breuer, T., Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Schaer, P., & Soboroff, I. (2020). How to measure the reproducibility of system-oriented IR experiments. In J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, & Y. Liu (Eds.), *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020* (pp. 349–358). ACM. https://doi.org/10.1145/3397271.3401036

Breuer, T., Ferro, N., Maistro, M., & Schaer, P. (2021). Repro_eval: A python interface to reproducibility measures of system-oriented IR experiments. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds.), *Advances in information retrieval - 43rd european conference on IR research, ECIR 2021, virtual event, march 28 - april 1, 2021, proceedings, part II* (pp. 481–486). Springer. https://doi.org/10.1007/978-3-030-72240-1_51

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2010). *Information retrieval - implementing and evaluating search engines*. MIT Press. http://mitpress.mit.edu/books/information-retrieval

Carterette, B., Bah, A., & Zengin, M. (2015). Dynamic test collections for retrieval evaluation. In J. Allan, W. B. Croft, A. P. de Vries, & C. Zhai (Eds.), *Proceedings of the 2015 international conference on the theory of information retrieval, ICTIR 2015, northampton, massachusetts, USA, september 27-30, 2015* (pp. 91–100). ACM. https://doi.org/10.1145/2808194.2809470

Chapelle, O., & Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. In J. Quemada, G. León, Y. S. Maarek, & W. Nejdl (Eds.), *Proceedings of the 18th international conference on world wide web, WWW 2009, madrid, spain, april 20-24, 2009* (pp. 1–10). ACM. https://doi.org/10.1145/1526709.1526711

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *J. Informetrics*, *11*(4), 1175–1189. https://doi.org/10.1016/j.joi.2017.10.003

Cheng, S., Arvanitis, A., & Hristidis, V. (2013). How fresh do you want your search results? In Q. He, A. Iyengar, W. Nejdl, J. Pei, & R. Rastogi (Eds.), *22nd ACM international conference on information and knowledge management, CIKM'13, san francisco, CA, USA, october 27 - november 1, 2013* (pp. 1271–1280). ACM. https://doi.org/10.1145/2505515.2505696

Cheriton, D. R. (2019). From doc2query to docTTTTTquery.

Cho, H.-G., Tak, H., Kim, H.-H., Kim, Y., Shin, Y.-J., Lim, C., & Choi, K.-N. (2017). Evaluation of full-text retrieval system using collection of serially evolved documents. *Proceedings of the 3rd International Conference on Industrial and Business Engineering, ICIBE 2017, Sapporo, Japan, August 17-19, 2017*, 40–45. https://doi.org/10.1145/3133811.3133817

Clarke, C. L. A., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Cleverdon, C. (1997). The cranfield tests on index language devices. In *Readings in information retrieval* (pp. 47–59). Morgan Kaufmann Publishers Inc.

Cohen, D. (2021). Allowing for the grounded use of temporal difference learning in large ranking models via substate updates. In F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, & T. Sakai (Eds.), *SIGIR '21: The 44th international ACM SIGIR conference on research and development in information retrieval, virtual event, canada, july 11-15, 2021* (pp. 438–448). ACM. https://doi.org/10.1145/3404835.3462952

Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), *Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2009, boston, MA, USA, july 19-23, 2009* (pp. 758–759). ACM. https://doi.org/10.1145/1571941.1572114

Cormack, G. V., Zhang, H., Ghelani, N., Abualsaud, M., Smucker, M. D., Grossman, M. R., Rahbariasl, S., & Ghenai, A. (2019). Dynamic sampling meets pooling. In B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J.-Y. Nie, & F. Scholer (Eds.), *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2019, paris, france, july 21-25, 2019* (pp. 1217–1220). ACM. https://doi.org/10.1145/3331184.3331354

Craswell, N., Zoeter, O., Taylor, M. J., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In M. Najork, A. Z. Broder, & S. Chakrabarti (Eds.), *Proceedings of the international conference on web search and web data mining, WSDM 2008, palo alto, california, USA, february 11-12, 2008* (pp. 87–94). ACM. https://doi.org/10.1145/1341531.1341545

Croft, W. B., Metzler, D., & Strohman, T. (2009). *Search engines - information retrieval in practice*. Pearson Education. http://www.search-engines-book.com/

Dai, N., & Davison, B. D. (2010). Freshness matters: In flowers, food, and web authority. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, & J. Savoy (Eds.), *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2010, geneva, switzerland, july 19-23, 2010* (pp. 114–121). ACM. https://doi.org/10.1145/1835449.1835471

Dai, N., Shokouhi, M., & Davison, B. D. (2011). Learning to rank for freshness and relevance. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yates, T.-S. Chua, & W. B. Croft (Eds.), *Proceeding of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2011, beijing, china, july 25-29, 2011* (pp. 95–104). ACM. https://doi.org/10.1145/2009916.2009933

Delasalles, E., Lamprier, S., & Denoyer, L. (2019). Learning dynamic author representations with temporal language models. *2019 IEEE International Conference on Data Mining (ICDM)*, 120–129. https://doi.org/10.1109/ICDM.2019.00022

Delasalles, E., Lamprier, S., & Denoyer, L. (2021). Deep dynamic neural networks for temporal language modeling in author communities. *Knowledge and Information Systems*, *63*(3), 733–757. https://doi.org/10.1007/s10115-020-01539-z

Deveaud, P. G. R., Gonzalez-Saez, G., Mulhem, P., Goeuriot, L., Piroi, F., & Popel, M. (2023, April 27). *LongEval-Retrieval: French-English Dynamic Test Collection for Continuous Web Search Evaluation*. arXiv: 2303.03229 [cs]. Retrieved May 6, 2023, from http://arxiv.org/abs/2303.03229

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, MN, USA, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

Diaz, F., & Jones, R. (2004). Using temporal profiles of queries for precision prediction. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza (Eds.), *SIGIR 2004: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, sheffield, UK, july 25-29, 2004* (pp. 18–24). ACM. https://doi.org/10.1145/1008992.1008998

Duan, Y., Jatowt, A., & Yoshikawa, M. (2021). Structured representation of temporal document collections by diachronic linguistic periodization.

Dumais, S. T. (2010). Temporal dynamics and information retrieval. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 7–8. https://doi.org/10.1145/1871437.1871442

Dumais, S. T. (2014). *Putting searchers into search*. https://doi.org/10.1145/2600428.2617557

Efron, M. (2013). Query representation for cross-temporal information retrieval. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, & T. Sakai (Eds.), *The 36th international*

*ACM SIGIR conference on research and development in information retrieval, SIGIR '13, dublin, ireland - july 28 - august 01, 2013* (pp. 383–392). ACM. https://doi.org/10.1145/2484028.2484054

Forman, G. (2006). Tackling concept drift by temporal inductive transfer. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, & K. Järvelin (Eds.), *SIGIR 2006: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, seattle, washington, USA, august 6-11, 2006* (pp. 252–259). ACM. https://doi.org/10.1145/1148170.1148216

Frieder, O., & Jensen, E. C. (2006). Repeatable evaluation of information retrieval effectiveness in dynamic environments.

Fröbe, M., Reimer, J. H., MacAvaney, S., Deckers, N., Reich, S., Bevendorff, J., Stein, B., Hagen, M., & Potthast, M. (2023). The information retrieval experiment platform. *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Commun. ACM*, *30*(11), 964–971. https://doi.org/10.1145/32206.32212

Golovchinsky, G., Diriye, A., & Dunnigan, T. (2012). The future is in the past: Designing for exploratory search. In J. Kamps, W. Kraaij, & N. Fuhr (Eds.), *Information interaction in context: 2012, IIix'12, nijmegen, the netherlands, august 21-24, 2012* (pp. 52–61). ACM. https://doi.org/10.1145/2362724.2362738

Hashemi, S. H., Clarke, C. L. A., Dean-Hall, A., Kamps, J., & Kiseleva, J. (2016). An easter egg hunting approach to test collection building in dynamic domains. In E. Yilmaz & C. L. A. Clarke (Eds.), *Proceedings of the seventh international workshop on evaluating information access, EVIA 2016, a satellite workshop of the NTCIR-12 conference, national center of sciences, tokyo, japan, june 7, 2016*. National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/evia/01-EVIA2016-HashemiS.pdf

Hashemi, S. H., & Kamps, J. (2017). On the reusability of personalized test collections. In M. Bieliková, E. Herder, F. Cena, & M. C. Desmarais (Eds.), *Adjunct publication of the 25th conference on user modeling, adaptation and personalization, UMAP 2017, bratislava, slovakia, july 09 - 12, 2017* (pp. 185–189). ACM. https://doi.org/10.1145/3099023.3099044

Hofmann, K., Li, L., & Radlinski, F. (2016). Online evaluation for information retrieval. *Found. Trends Inf. Retr.*, *10*, 1–117.

Holubová, I., Klettke, M., & Störl, U. (2019). Evolution management of multi-model data - (position paper). In V. Gadepally, T. G. Mattson, M. Stonebraker, F. Wang, G. Luo, Y. Laing, & A. Dubovitskaya (Eds.), *Heterogeneous data management, polystores, and analytics for healthcare - VLDB 2019 workshops, poly and DMAH, los angeles, CA, USA, august 30, 2019, revised selected papers* (pp. 139–153). Springer. https://doi.org/10.1007/978-3-030-33752-0_10

Holzmann, H., & Risse, T. (2014a). Extraction of evolution descriptions from the web. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, 413–414.

Holzmann, H., & Risse, T. (2014b). Named entity evolution analysis on wikipedia. *Proceedings of the 2014 ACM Conference on Web Science*, 241–242. https://doi.org/10.1145/2615569.2615639

Hopfgartner, F., Balog, K., Lommatzsch, A., Kelly, L., Kille, B., Schuth, A., & Larson, M. A. (2019). Continuous evaluation of large-scale information access systems: A case for living labs. In N. Ferro & C. Peters (Eds.), *Information retrieval evaluation in a changing world - lessons learned from 20 years of CLEF* (pp. 511–543). Springer. https://doi.org/10.1007/978-3-030-22948-1_21

Ibrahim, O. A. S., & Landa-Silva, D. (2014). A new weighting scheme and discriminative approach for information retrieval in static and dynamic document collections. *14th UK Workshop on Computational Intelligence, UKCI 2014, Bradford, UK, September 8-10, 2014*, 1–8. https://doi.org/10.1109/UKCI.2014.6930160

Irfan, R., Khan, S., Rajpoot, K., & Qamar, A. M. (2018). TIE algorithm: A layer over clustering-based taxonomy generation for handling evolving data. *Frontiers Inf. Technol. Electron. Eng.*, *19*(6), 763–782. https://doi.org/10.1631/FITEE.1700517

Jaleel, N. A., Allan, J., Croft, W. B., Diaz, F., Larkey, L. S., Li, X., Smucker, M. D., & Wade, C. (2004). UMass at TREC 2004: Novelty and HARD. In E. M. Voorhees & L. P. Buckland (Eds.), *Proceedings of the thirteenth text REtrieval conference, TREC 2004, gaithersburg, maryland, USA, november 16-19, 2004*. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf

Jatowt, A., & Tanaka, K. (2012). Large scale analysis of changes in english vocabulary over recent time. In X.-w. Chen, G. Lebanon, H. Wang, & M. J. Zaki (Eds.), *21st ACM international conference on information and knowledge management, CIKM'12, maui, HI, USA, october 29 - november 02, 2012* (pp. 2523–2526). ACM. https://doi.org/10.1145/2396761.2398682

Jayasinghe, G. K., Webber, W., Sanderson, M., & Culpepper, J. S. (2014). Extending test collection pools without manual runs. In S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, & K. Järvelin (Eds.), *The 37th international ACM SIGIR conference on research and development in information retrieval, SIGIR '14, gold coast , QLD, australia - july 06 - 11, 2014* (pp. 915–918). ACM. https://doi.org/10.1145/2600428.2609473

Joho, H., Jatowt, A., Blanco, R., Naka, H., & Yamamoto, S. (2014). Overview of NTCIR-11 temporal information access (temporalia) task. In N. Kando, H. Joho, & K. Kishida (Eds.), *Proceedings of the 11th NTCIR conference on evaluation of information access technologies, NTCIR-11, national center of sciences, tokyo, japan, december 9-12, 2014*. National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/OVERVIEW/01-NTCIR11-OV-TEMPORALIA-JohoH.pdf

Kaluarachchi, A. C., Roychoudhury, D., Varde, A. S., & Weikum, G. (2011). SITAC: Discovering *semantically identical temporally altering concepts* in text archives. In A. Ailamaki, S. Amer-Yahia, J. M. Patel, T. Risch, P. Senellart, & J. Stoyanovich (Eds.), *EDBT 2011, 14th international conference on extending database technology, uppsala, sweden, march 21-24, 2011, proceedings* (pp. 566–569). ACM. https://doi.org/10.1145/1951365.1951442

Kaluarachchi, A. C., Varde, A. S., Bedathur, S. J., Weikum, G., Peng, J., & Feldman, A. (2010). Incorporating terminology evolution for query translation in text retrieval with association rules. In J. X. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, & A. An (Eds.), *Proceedings of the 19th ACM conference on information and knowledge management, CIKM 2010, toronto, ontario, canada, october 26-30, 2010* (pp. 1789–1792). ACM. https://doi.org/10.1145/1871437.1871730

Kanhabua, N. (2009). Exploiting temporal information in retrieval of archived documents. In J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, & J. Zobel (Eds.), *Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2009, boston, MA, USA, july 19-23, 2009* (p. 848). ACM. https://doi.org/10.1145/1571941.1572169

Kanhabua, N., & Anand, A. (2016). Temporal information retrieval. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1235–1238. https://doi.org/10.1145/2911451.2914805

Keller, J. (2023, July 10). *Continuous Evaluation in Information Retrieval* (Version 1). https://doi.org/10.5281/ZENODO.8126865

Keller, J., & Munz, L. P. M. (2022). Evaluating research dataset recommendations in a living lab. In A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, & N. Ferro (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction - 13th international conference of the CLEF association, CLEF 2022, bologna, italy, september 5-8, 2022, proceedings* (pp. 135–148). Springer. https://doi.org/10.1007/978-3-031-13643-6_11

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, *3*(1-2), 1–224. https://doi.org/10.1561/1500000012

Kendall, M. G. (1949). Rank correlation methods.

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, & Y. Liu (Eds.), *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020* (pp. 39–48). ACM. https://doi.org/10.1145/3397271.3401075

Kim, H. D., Nikitin, D., Zhai, C., Castellanos, M., & Hsu, M. (2013). Information retrieval with time series query. In O. Kurland, D. Metzler, C. Lioma, B. Larsen, & P. Ingwersen (Eds.), *International conference on the theory of information retrieval, ICTIR '13, copenhagen, denmark, september 29 - october 02, 2013* (p. 14). ACM. https://doi.org/10.1145/2499178.2499195

Kitchenham, B. A., & Charters, S. (2007, July 9). *Guidelines for performing systematic literature reviews in software engineering* (EBSE 2007-001). Keele University and Durham University Joint Report / Keele University. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

Kleinberg, J. M. (2016). Temporal dynamics of on-line information streams. In M. N. Garofalakis, J. Gehrke, & R. Rastogi (Eds.), *Data stream management - processing high-speed data streams* (pp. 221–238). Springer. https://doi.org/10.1007/978-3-540-28608-0_11

Kohavi, R. (2015). Online controlled experiments: Lessons from running A/B/n tests for 12 years. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, & G. Williams (Eds.), *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, sydney, NSW, australia, august 10-13, 2015* (p. 1). ACM. https://doi.org/10.1145/2783258.2785464

Kuang, N. L., & Clement H.C., L. (2019). Analysis of evolutionary behavior in self-learning media search engines. *2019 IEEE International Conference on Big Data (Big Data)*, 643–650. https://doi.org/10.1109/BigData47090.2019.9006191

Kürsten, J. (2012). *A generic approach to component-level evaluation in information retrieval* (Doctoral dissertation). Chemnitz University of Technology. https://d-nb.info/1028505493

Leibscher, R. (2004). Temporal context: Applications and implications for computational linguistics. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, July 21-26, 2004 - Student Research Workshop*. https://aclanthology.org/P04-2004/

Li, B., Li, W., & Lu, Q. (2006). Enhancing topic tracking with temporal information. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, & K. Järvelin (Eds.), *SIGIR 2006: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, seattle, washington, USA, august 6-11, 2006* (pp. 667–668). ACM. https://doi.org/10.1145/1148170.1148308

Li, D., & Kanoulas, E. (2017). Active Sampling for Large-scale Information Retrieval Evaluation. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 49–58. https://doi.org/10.1145/3132847.3133015

Ma, C., Ren, Y., Castells, P., & Sanderson, M. (2022). Evaluation of herd behavior caused by population-scale concept drift in collaborative filtering. In E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, & G. Kazai (Eds.), *SIGIR '22: The 45th international ACM SIGIR conference on research and development in information retrieval, madrid, spain, july 11 - 15, 2022* (pp. 1984–1989). ACM. https://doi.org/10.1145/3477495.3531792

Macdonald, C., & Tonellotto, N. (2020). Declarative experimentation in information retrieval using PyTerrier. In K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, & K. Berberich (Eds.), *ICTIR '20: The 2020 ACM SIGIR international conference on the theory of information retrieval, virtual event, norway, september 14-17, 2020* (pp. 161–168). ACM. https://doi.org/10.1145/3409256.3409829

Maistro, M., Breuer, T., Schaer, P., & Ferro, N. (2023). An in-depth investigation on the behavior of measures to quantify reproducibility. *Inf. Process. Manag.*, *60*(3), 103332. https://doi.org/10.1016/j.ipm.2023.103332

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Mansouri, B., Zahedi, M. S., Rahgozar, M., & Campos, R. (2017). Detecting seasonal queries using time series and content features. In J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, & E. Yilmaz (Eds.), *Proceedings of the ACM SIGIR international conference on theory of information retrieval, ICTIR 2017, amsterdam, the netherlands, october 1-4, 2017* (pp. 297–300). ACM. https://doi.org/10.1145/3121050.3121100

Mohammed, S., Crane, M., & Lin, J. (2017). Quantization in append-only collections. In J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, & E. Yilmaz (Eds.), *Proceedings of the ACM SIGIR international conference on theory of information retrieval, ICTIR 2017, amsterdam, the netherlands, october 1-4, 2017* (pp. 265–268). ACM. https://doi.org/10.1145/3121050.3121092

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated MAchine reading COmprehension dataset. In T. R. Besold, A. Bordes, A. S. d'Avila Garcez, & G. Wayne (Eds.), *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*. CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

Nishida, K., Hoshide, T., & Fujimura, K. (2012). Improving tweet stream classification by detecting changes in word probability. In W. R. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), *The 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR '12, portland, OR, USA, august 12-16, 2012* (pp. 971–980). ACM. https://doi.org/10.1145/2348283.2348412

Nogueira, R., Jiang, Z., Pradeep, R., & Lin, J. (2020). Document Ranking with a Pretrained Sequence-to-Sequence Model, 708–718. https://doi.org/10.18653/v1/2020.findings-emnlp.63

Nogueira, R., & Lin, J. (2019). From doc2query to docTTTTTquery.

Nunes, S., Ribeiro, C., & David, G. (2010). Term frequency dynamics in collaborative articles. In A. Antonacopoulos, M. J. Gormish, & R. Ingold (Eds.), *Proceedings of the 2010 ACM symposium on document engineering, manchester, united kingdom, september 21-24, 2010* (pp. 267–270). ACM. https://doi.org/10.1145/1860559.1860620

Otero, D., Parapar, J., & Barreiro, Á. (2023). Relevance feedback for building pooled test collections. *Journal of Information Science*, *0*(0), 01655515231171085. https://doi.org/10.1177/01655515231171085

Pass, G., Chowdhury, A., & Torgeson, C. (2006). A picture of search. *Proceedings of the 1st International Conference on Scalable Information Systems - InfoScale '06*, 1–es. https://doi.org/10.1145/1146847.1146848

Perkiö, J., Buntine, W. L., & Tirri, H. (2005). A temporally adaptive content-based relevance ranking algorithm. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait (Eds.), *SIGIR 2005: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, salvador, brazil, august 15-19, 2005* (pp. 647–648). ACM. https://doi.org/10.1145/1076034.1076171

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell Pub
OCLC: ocm60360309.

Pradeep, R., Nogueira, R., & Lin, J. (2021, January 14). *The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models*. arXiv: 2101.05667 [cs]. Retrieved May 1, 2023, from http://arxiv.org/abs/2101.05667

Qian, X., Lin, J., & Roegiest, A. (2016). Interleaved evaluation for retrospective summarization and prospective notification on document streams. 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* (pp. 175–184). ACM. https://doi.org/10.1145/2911451.2911494

Radinsky, K., Diaz, F., Dumais, S. T., Shokouhi, M., Dong, A., & Chang, Y. (2013). Temporal web dynamics and its application to information retrieval. In S. Leonardi, A. Panconesi, P. Ferragina, & A. Gionis (Eds.), *Sixth ACM international conference on web search and data mining, WSDM 2013, rome, italy, february 4-8, 2013* (pp. 781–782). ACM. https://doi.org/10.1145/2433396.2433500

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

Ren, P., Chen, Z., Ma, J., Zhang, Z., Si, L., & Wang, S. (2017). Detecting temporal patterns of user queries. *J. Assoc. Inf. Sci. Technol.*, *68*(1), 113–128. https://doi.org/10.1002/asi.23578

Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E. M., Wang, L. L., & Hersh, W. R. (2020). TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19. *J. Am. Medical Informatics Assoc.*, *27*(9), 1431–1436. https://doi.org/10.1093/jamia/ocaa091

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994, January 1). *Okapi at TREC-3.*

Roitero, K., Soprano, M., Brunello, A., & Mizzaro, S. (2018). Reproduce and improve: An evolutionary approach to select a few good topics for information retrieval evaluation. *ACM J. Data Inf. Qual.*, *10*(3), 12:1–12:21. https://doi.org/10.1145/3239573

Ryu, C.-K., Kim, H.-J., & Cho, H.-G. (2008). Reconstructing evolution process of documents in spatio-temporal analysis. *2008 Third International Conference on Con-*

*vergence and Hybrid Information Technology*, *1*, 136–142. https://doi.org/10.1109/ICCIT.2008.233

Sáez, G. N. G., Goeuriot, L., & Mulhem, P. (2021). Addressing different evaluation environments for information retrieval through pivot systems. In A. Doucet & A.-G. Chifu (Eds.), *COnférence en recherche d'Informations et applications - CORIA 2021, french information retrieval conference, grenoble, france, april 15, 2021*. ARIA. https://doi.org/10.24348/coria.2021.long_6

Sáez, G. N. G., Mulhem, P., & Goeuriot, L. (2021). Towards the evaluation of information retrieval systems on evolving datasets with pivot systems. In K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, & N. Ferro (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction - 12th international conference of the CLEF association, CLEF 2021, virtual event, september 21-24, 2021, proceedings* (pp. 91–102). Springer. https://doi.org/10.1007/978-3-030-85251-1_8

Sahraoui, A. K., & Faiz, R. (2017). Time sensitivity for personalized search. *14th IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2017, Hammamet, Tunisia, October 30 - Nov. 3, 2017*, 585–592. https://doi.org/10.1109/AICCSA.2017.77

Sakai, T., Tao, S., Chen, N., Li, Y., Maistro, M., Chu, Z., & Ferro, N. (2023). On the ordering of pooled web pages, gold assessments, and bronze assessments. *ACM Trans. Inf. Syst.* https://doi.org/10.1145/3600227

Salles, T., Rocha, L., Pappa, G. L., Mourão, F., Jr., W. M., & Gonçalves, M. A. (2010). Temporally-aware algorithms for document classification. In F. Crestani, S. Marchand-Maillet, H.-H. Chen, E. N. Efthimiadis, & J. Savoy (Eds.), *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2010, geneva, switzerland, july 19-23, 2010* (pp. 307–314). ACM. https://doi.org/10.1145/1835449.1835502

Sánchez, P., & Bellogín, A. (2018). Time-aware novelty metrics for recommender systems. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval - 40th european conference on IR research, ECIR 2018, grenoble, france, march 26-29, 2018, proceedings* (pp. 357–370). Springer. https://doi.org/10.1007/978-3-319-76941-7_27

Sánchez, P., Mesas, R. M., & Bellogín, A. (2018). New approaches for evaluation: Correctness and freshness: Extended Abstract. In J. Tramullas, R. T. Lado, & J. Nogueras-Iso (Eds.), *Proceedings of the 5th spanish conference on information retrieval, CERI 2018, zaragoza, spain, june 26-27, 2018* (14:1–14:2). ACM. https://doi.org/10.1145/3230599.3230614

Sanderson, M. (2010). BEST PRACTICES FOR TEST COLLECTION CREATION AND INFORMATION RETRIEVAL SYSTEM EVALUATION.

Schaer, P., Breuer, T., Castro, L. J., Wolff, B., Schaible, J., & Tavakolpoursaleh, N. (2021). Overview of LiLAS 2021 - living labs for academic search (extended overview). In G. Faggioli, N. Ferro, A. Joly, M. Maistro, & F. Piroi (Eds.), *Working notes of CLEF 2021 - conference and labs of the evaluation forum*.

Schaer, P., Castro, L. J., & Hienert, D. (2021). Infrastructures for Living Labs - Project Phase II.

Sidana, S., Mishra, S., Amer-Yahia, S., Clausel, M., & Amini, M.-R. (2016). Health monitoring on social media over time. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 849–852. https://doi.org/10.1145/2911451.2914697

Silva, R. L. S., & Neiva, F. W. (2016). Systematic Literature Review in Computer Science - A Practical Guide. https://doi.org/10.13140/RG.2.2.35453.87524

Sisman, B., & Kak, A. C. (2012). Incorporating version histories in Information Retrieval based bug localization. In M. Lanza, M. D. Penta, & T. Xie (Eds.), *9th IEEE working conference of mining software repositories, MSR 2012, june 2-3, 2012, zurich, switzerland* (pp. 50–59). IEEE Computer Society. https://doi.org/10.1109/MSR.2012.6224299

Soboroff, I. (2006). Dynamic test collections: Measuring search effectiveness on the live web. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, & K. Järvelin (Eds.), *SIGIR 2006: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, seattle, washington, USA, august 6-11, 2006* (pp. 276–283). ACM. https://doi.org/10.1145/1148170.1148220

Stowe, K., & Gurevych, I. (2021). Combating temporal drift in crisis with adapted embeddings. *CoRR*, *abs/2104.08535*. https://arxiv.org/abs/2104.08535

Strötgen, J., Alonso, O., & Gertz, M. (2012). Retro: Time-based exploration of product reviews. In R. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, & F. Silvestri (Eds.), *Advances in information retrieval - 34th european conference on IR research, ECIR 2012, barcelona, spain, april 1-5, 2012. Proceedings* (pp. 581–582). Springer. https://doi.org/10.1007/978-3-642-28997-2_71

Sun, P., Wu, L., & Wang, M. (2018). Attentive recurrent social recommendation. In K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, & E. Yilmaz (Eds.), *The 41st international ACM SIGIR conference on research & development in information retrieval, SIGIR 2018, ann arbor, MI, USA, july 08-12, 2018* (pp. 185–194). ACM. https://doi.org/10.1145/3209978.3210023

Svore, K. M., Teevan, J., Dumais, S. T., & Kulkarni, A. (2012). Creating temporally dynamic web search snippets. In W. R. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), *The 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR '12, portland, OR, USA, august 12-16, 2012* (pp. 1045–1046). ACM. https://doi.org/10.1145/2348283.2348461

Takeda, N., Seki, Y., Morishita, M., & Inagaki, Y. (2017). Evolution of information needs based on life event experiences with topic transition. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, & R. W. White (Eds.), *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, shinjuku, tokyo, japan, august 7-11, 2017* (pp. 1009–1012). ACM. https://doi.org/10.1145/3077136.3080703

Tan, L., Baruah, G., & Lin, J. (2017). On the reusability of "Living Labs" test collections: : A case study of real-time summarization. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, & R. W. White (Eds.), *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, shinjuku, tokyo, japan, august 7-11, 2017* (pp. 793–796). ACM. https://doi.org/10.1145/3077136.3080644

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In J. Vanschoren & S.-K. Yeung (Eds.), *Proceedings of the neural information processing systems track on datasets and benchmarks 1, NeurIPS datasets and benchmarks 2021, december 2021, virtual*. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html

Tikhonov, A., Bogatyy, I., Burangulov, P., Ostroumova, L., Koshelev, V., & Gusev, G. (2013). Studying page life patterns in dynamical web. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, & T. Sakai (Eds.), *The 36th international ACM SIGIR conference on research and development in information retrieval, SIGIR '13, dublin, ireland - july 28 - august 01, 2013* (pp. 905–908). ACM. https://doi.org/10.1145/2484028.2484185

Tonon, A., Demartini, G., & Cudré-Mauroux, P. (2015). Pooling-based continuous evaluation of information retrieval systems. *Inf. Retr. J.*, *18*(5), 445–472. https://doi.org/10.1007/s10791-015-9266-y

Tsevas, S., & Iakovidis, D. K. (2011). Fusion of multimodal temporal clinical data for the retrieval of similar patient cases. *2011 10th International Workshop on Biomedical Engineering*, 1–4.

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, & K. Järvelin (Eds.), *SIGIR 2006: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, seattle, washington, USA, august 6-11, 2006* (pp. 11–18). ACM. https://doi.org/10.1145/1148170.1148176

Uehara, M., Sato, N., & Sakai, Y. (2005). Adaptive calculation of scores for fresh information retrieval. *11th International Conference on Parallel and Distributed Systems, ICPADS 2005, Fuduoka, Japan, July 20-22, 2005*, 750–755. https://doi.org/10.1109/ICPADS.2005.65

van Dam, M., & Hauff, C. (2014). Large-scale author verification: Temporal and topical influences. In S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, & K. Järvelin (Eds.), *The 37th international ACM SIGIR conference on research and development in information retrieval, SIGIR '14, gold coast , QLD, australia - july 06 - 11, 2014* (pp. 1039–1042). ACM. https://doi.org/10.1145/2600428.2609504

Voorhees, E. M. (2019). The evolution of cranfield. In N. Ferro & C. Peters (Eds.), *Information retrieval evaluation in a changing world - lessons learned from 20 years of CLEF* (pp. 45–69). Springer. https://doi.org/10.1007/978-3-030-22948-1_2

Vouzoukidou, D. (2015). *Continuous top-k queries over real-time web streams. (Evaluation de requêtes top-k continues à large-échelle)* (Doctoral dissertation). Pierre and Marie Curie University, Paris, France. https://tel.archives-ouvertes.fr/tel-01366673

Wang, B., Buccio, E. D., & Melucci, M. (2021). Sequential modeling in vector space. In V. W. Anelli, T. D. Noia, N. Ferro, & F. Narducci (Eds.), *Proceedings of the 11th italian information retrieval workshop 2021, bari, italy, september 13-15, 2021*. CEUR-WS.org. https://ceur-ws.org/Vol-2947/paper12.pdf

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, *abs/2212.03533*. https://doi.org/10.48550/arXiv.2212.03533

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D. A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., ... Kohlmeier, S. (2020). CORD-19: The covid-19 open research dataset. *CoRR*, *abs/2004.10706*. https://arxiv.org/abs/2004.10706

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, *28*(4), 20:1–20:38. https://doi.org/10.1145/1852102.1852106

Wei, C.-P., & Chang, Y.-H. (2007). Discovering event evolution patterns from document sequences. *IEEE Trans. Syst. Man Cybern. Part A*, *37*(2), 273–283. https://doi.org/10.1109/TSMCA.2006.886377

Wei, C.-P., & Dong, Y.-X. (2001). A mining-based category evolution approach to managing online document categories. *34th Annual Hawaii International Conference on System Sciences (HICSS-34), January 3-6, 2001, Maui, Hawaii, USA*. https://doi.org/10.1109/HICSS.2001.927093

Whiting, S., Klampanos, I. A., & Jose, J. M. (2012). Temporal pseudo-relevance feedback in microblog retrieval. In R. Baeza-Yates, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, & F. Silvestri (Eds.), *Advances in information retrieval - 34th european conference on IR research, ECIR 2012, barcelona, spain, april 1-5, 2012. Proceedings* (pp. 522–526). Springer. https://doi.org/10.1007/978-3-642-28997-2_55

Whiting, S., Moshfeghi, Y., & Jose, J. M. (2011). Exploring term temporality for pseudo-relevance feedback. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yates, T.-S. Chua, & W. B. Croft (Eds.), *Proceeding of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2011, beijing, china, july 25-29, 2011* (pp. 1245–1246). ACM. https://doi.org/10.1145/2009916.2010141

Wu, Z., Li, C., Zhao, Z., Wu, F., & Mei, Q. (2018). Identify shifts of word semantics through bayesian surprise. In K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, & E. Yilmaz (Eds.), *The 41st international ACM SIGIR conference on research & development in information retrieval, SIGIR 2018, ann arbor, MI, USA, july 08-12, 2018* (pp. 825–834). ACM. https://doi.org/10.1145/3209978.3210040

Yadav, M. M., Sudhir, S., Saketh, M., Avinash, S., & Alekya, M. (2021). Analyzing evolving trends on social networks by using temporal bipartite networks. *2021 Third Inter-*

*national Conference on Inventive Research in Computing Applications (ICIRCA)*, 380–386. https://doi.org/10.1109/ICIRCA51532.2021.9544529

Yang, C. C., Shi, X., & Wei, C.-P. (2009). Discovering event evolution graphs from news corpora. *IEEE Trans. Syst. Man Cybern. Part A*, *39*(4), 850–863. https://doi.org/10.1109/TSMCA.2009.2015885

Yang, Y., & Kisiel, B. (2003). Margin-based local regression for adaptive filtering. *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003*, 191–198. https://doi.org/10.1145/956863.956902

Yeniterzi, R., & Callan, J. (2014). Analyzing bias in CQA-based expert finding test sets. In S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, & K. Järvelin (Eds.), *The 37th international ACM SIGIR conference on research and development in information retrieval, SIGIR '14, gold coast , QLD, australia - july 06 - 11, 2014* (pp. 967–970). ACM. https://doi.org/10.1145/2600428.2609486

Yoon, T., Myaeng, S.-H., Woo, H.-W., Lee, S.-W., & Kim, S.-B. (2018). On temporally sensitive word embeddings for news information retrieval. In D. Albakour, D. P. A. Corney, J. Gonzalo, M. Martinez-Alvarez, B. Poblete, & A. Valochas (Eds.), *Proceedings of the second international workshop on recent trends in news information retrieval co-located with 40th european conference on information retrieval (ECIR 2018), grenoble, france, march 26, 2018* (pp. 51–56). CEUR-WS.org. https://ceur-ws.org/Vol-2079/paper11.pdf

Zein, D. E. (2021). User knowledge and search goals in information retrieval: A benchmark and study on the evolution of users' knowledge gain. In O. Alonso, S. Marchesin, M. Najork, & G. Silvello (Eds.), *Proceedings of the second international conference on design of experimental search & information REtrieval systems, padova, italy, september 15-18, 2021* (pp. 189–190). CEUR-WS.org. https://ceur-ws.org/Vol-2950/paper-22.pdf

Zhang, L., Ai, W., Yuan, C., Zhang, Y., & Ye, J. (2018). Taxi or hitchhiking: Predicting passenger's preferred service on ride sharing platforms. In K. Collins-Thompson, Q. Mei, B. D. Davison, Y. Liu, & E. Yilmaz (Eds.), *The 41st international ACM SIGIR conference on research & development in information retrieval, SIGIR 2018, ann arbor, MI, USA, july 08-12, 2018* (pp. 1041–1044). ACM. https://doi.org/10.1145/3209978.3210153

Zhang, L., Joho, H., & Yu, H. (2022). Semantic modelling of document focus-time for temporal information retrieval. In F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, & L. Médini (Eds.), *Companion of the web conference 2022, virtual event / lyon, france, april 25 - 29, 2022* (pp. 896–902). ACM. https://doi.org/10.1145/3487553.3524668

Zhang, X., Li, X., Jiang, S.-y., Li, X., & Xie, B. (2019). Evolution Analysis of Information Retrieval based on co-word network. *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, 1837–1840. https://doi.org/10.1109/EITCE47263.2019.9094904

Zheng, L., Guo, N., Chen, W., Yu, J., & Jiang, D. (2020). Sentiment-guided sequential recommendation. In J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, & Y. Liu (Eds.), *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020* (pp. 1957–1960). ACM. https://doi.org/10.1145/3397271.3401330

Zhou, A. J., Luo, J., & Yang, H. (2015). DUMPLING: A novel dynamic search engine. In R. Baeza-Yates, M. Lalmas, A. Moffat, & B. A. Ribeiro-Neto (Eds.), *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, santiago, chile, august 9-13, 2015* (pp. 1049–1050). ACM. https://doi.org/10.1145/2766462.2767873

Zhou, H., Yu, H., & Hu, R. (2017). Topic evolution based on the probabilistic topic model: A review. *Frontiers Comput. Sci.*, *11*(5), 786–802. https://doi.org/10.1007/s11704-016-5442-5

# Appendix

## Temporal Review Queries

### Scopus

```
TITLE-ABS-KEY("continuous evaluation" OR "temporal shift"
OR "longitudinal evaluation" OR "dynamic test collection" OR "evolving test
collection" OR "evolving dataset" OR "evolv" OR "temporal" OR "temporal
generalisability" OR "monitoring" OR "temporal decay" OR "delta" OR
"evaluation environment" OR "time-evolving") AND TITLE-ABS("information
retrieval") AND ( LIMIT-TO ( SUBJAREA,"COMP" ) ) AND ( LIMIT-TO ( LANGUAGE,
"English" ) OR LIMIT-TO ( LANGUAGE, "German" ) )
```

### ACM Digital Library:

```
(Title: "continuous evaluation" OR Title: "temporal shift" OR Title:
"longitudinal evaluation" OR Title: "dynamic test collection" OR
Title: "evolving test collection" OR Title: "evolving dataset" OR
Title: "evolv" OR Title: "temporal" OR Title: "temporal persistance"
OR Title: "temporal generalisability" OR Title: "temporal decay" OR
Title: "temporal evolution" OR Title: "evolution" OR Title: "delta" OR
Title: "evaluation environment" OR Title: "time-evolving" OR Abstract:
"continuous evaluation" OR Abstract: "temporal shift" OR Abstract:
"longitudinal evaluation" OR Abstract: "dynamic test collection" OR
Abstract: "evolving test collection" OR Abstract: "evolving dataset"
OR Abstract: "evolv" OR Abstract: "temporal" OR Abstract: "temporal
persistance" OR Abstract: "temporal generalisability" OR Abstract:
"temporal decay" OR Abstract: "temporal evolution" OR Abstract:
"evolution" OR Abstract: "delta" OR Abstract: "evaluation environment"
OR Abstract: "time-evolving" OR Keyword: "continuous evaluation"
OR Keyword: "temporal shift" OR Keyword: "longitudinal evaluation"
OR Keyword: "dynamic test collection" OR Keyword: "evolving test
collection" OR Keyword: "evolving dataset" OR Keyword: "evolv" OR
Keyword: "temporal" OR Keyword: "temporal persistance" OR Keyword:
"temporal generalisability" OR Keyword: "temporal decay" OR Keyword:
"temporal evolution" OR Keyword: "evolution" OR Keyword: "delta" OR
Keyword: "evaluation environment" OR Keyword: "time-evolving") AND
(Title: "information retrieval" OR Abstract: "information retrieval"
OR Keyword: "information retrieval")
```

### IEEE Xplore:

```
"Document Title":"continuous evaluation" OR "Document Title":"temporal
shift" OR "Document Title":"longitudinal evaluation" OR "Document
Title":"dynamic test collection" OR "Document Title":"evolving test
```

```
collection" OR "Document Title":"evolving dataset" OR "Document
Title":"evolv" OR "Document Title":"temporal" OR "Document Title":"temporal
persistance" OR "Document Title":"temporal generalisability" OR
"Document Title":"temporal decay" OR "Document Title":"temporal
evolution" OR "Document Title":"evolution" OR "Document Title":"delta"
OR "Document Title":"evaluation environment" OR "Document
Title":"time-evolving" OR "Abstract":"continuous evaluation" OR
"Abstract":"temporal shift" OR "Abstract":"longitudinal evaluation"
OR "Abstract":"dynamic test collection" OR "Abstract":"evolving test
collection" OR "Abstract":"evolving dataset" OR "Abstract":"evolv"
OR "Abstract":"temporal" OR "Abstract":"temporal persistance" OR
"Abstract":"temporal generalisability" OR "Abstract":"temporal
decay" OR "Abstract":"temporal evolution" OR "Abstract":"evolution"
OR "Abstract":"delta" OR "Abstract":"evaluation environment" OR
"Abstract":"time-evolving" OR "Author Keywords":"continuous evaluation"
OR "Author Keywords":"temporal shift" OR "Author Keywords":"longitudinal
evaluation" OR "Author Keywords":"dynamic test collection" OR "Author
Keywords":"evolving test collection" OR "Author Keywords":"evolving
dataset" OR "Author Keywords":"evolv" OR "Author Keywords":"temporal" OR
"Author Keywords":"temporal persistance" OR "Author Keywords":"temporal
generalisability" OR "Author Keywords":"temporal decay" OR "Author
Keywords":"temporal evolution" OR "Author Keywords":"evolution" OR "Author
Keywords":"delta" OR "Author Keywords":"evaluation environment" OR "Author
Keywords":"time-evolving") AND ("Document Title":"information retrieval"
OR "Abstract":"information retrieval" OR "Author Keywords":"information
retrieval"
```