Evaluation of IR Systems

NICOLA FERRO, University of Padua, Italy
MARIA MAISTRO, University of Copenhagen, Denmark

ACM Reference Format:
Nicola Ferro and Maria Maistro. 2024. Evaluation of IR Systems. 1, 1 (September 2024), 75 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Information Retrieval (IR) system performance can be evaluated from two different standpoints, *efficiency* and *effectiveness*. Efficiency is concerned with the algorithmic costs of IR systems, i.e., how fast they are in processing the needed information and how demanding they are in terms of the computational resources required, namely CPU, memory, and storage. Effectiveness, instead, is concerned with the ability of IR systems to retrieve and properly rank relevant documents while at the same time suppressing the retrieval of non relevant ones. The ultimate goal is to satisfy the user's information needs.

While efficiency could also be assessed formally, e.g., by proving the computational complexity of the adopted algorithms, effectiveness can be assessed only experimentally and this is why IR is a discipline strongly rooted in experimentation since its inception, as Spärck Jones [372] and Harman [189] have deeply discussed. Over the years, experimental evaluation has thus represented a main driver of progress and innovation in the IR field, providing the means to assess, understand, and improve the performance of IR systems from the viewpoint of effectiveness. Experimental evaluation has not only propelled research in the field but, as pointed out by Rowe et al. [327], it also had a continued and remarkable economic impact in terms of return on investment for both academia and industry.

Experimental evaluation traditionally covers a very wide spectrum of cases, ranging from *system-oriented evaluation*, accurately described by Sanderson [346], to *user-oriented evaluation*, carefully summarized by Kelly [229]. This way of "categorizing" evaluation is still appropriate and current. However, IR systems have greatly evolved over the decades and, nowadays, they embrace a paradigm which de Rijke [113] called *retrieval as interaction*[1] where their development is best thought of as a two-stage process: offline development followed by continued online adaptation based on interactions with users. Both these stages are driven by evaluation: in the offline phase to tune parameters and learn from annotated datasets or log data; in the online phase to learn and adapt to live interaction from users.

[1]This vision emerged through a series of technical papers over the last decade, stemming from early works by Hofmann et al. [202, 203] to recent ones by Oosterhuis and de Rijke [298].

Authors' addresses: Nicola Ferro, nicola.ferro@unipd.it, University of Padua, Padua, Italy; Maria Maistro, mm@di.ku.dk, University of Copenhagen, Copenhagen, Denmark.

XXXX-XXXX/2024/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Therefore, in this chapter, we present evaluation as ranging from *offline evaluation* to *online evaluation*, also addressing the possibility of mixed approaches where you perform offline evaluation but exploiting online data. Note that this vision does not fully cover the spectrum from system-oriented to user-oriented evaluation, especially for what concerns evaluation of interactive IR. For a detailed discussion on interactive IR and its evaluation please refer to, e.g., Ingwersen and Järvelin [209], Ruthven and Kelly [332], and White [422].

In particular, Section 2 discusses the basis of offline evaluation, namely the *Cranfield paradigm*, and how it is implemented, detailing how corpora and ground truth are created and how the paradigm is embodied by evaluation campaigns. Section 3 presents some of the most widely used performance measures, used in offline evaluation. Section 4 introduces the basic notions about statistical significance testing and, then, it focuses on *ANalysis Of VAriance (ANOVA)* and how to properly compare multiple systems. Section 5 starts to bridge from offline towards online evaluation, explaining how to calibrate offline evaluation measures with online data. Section 6 describes the basic principles of online evaluation and then presents two widely used alternatives for it, namely A/B testing and interleaving, as well as performance measures for online evaluation. Section 7 deals with an issue transversal to offline and online evaluation, namely the foundational aspects of measurement and their implications for IR evaluation. Finally, Section 8 wraps up the chapter and presents some future trends in the field.

2 OFFLINE EVALUATION

This section introduces the building blocks of experimentation and evaluation of IR systems. Section 2.1 starts with the Cranfield Paradigm, which is at the basis of IR evaluation. It was proposed in the sixties but it still represents the core paradigm to build experimental collections and run evaluation campaigns. The rest of the section covers different aspects involved in the creation and usage of test collections. Section 2.2 describes how test collections can be created as a joint effort during evaluation campaigns. Section 2.3 lists some popular test collections created in the context of evaluation campaigns and widely used within the research community. Then the focus shifts on how to collect annotations for test collections, which is the most expensive and demanding activity when creating a test collection. Section 2.4 describes pooling, i.e., how to select the candidate documents to be annotated. Section 2.5 explains how to use crowdsourcing to assess such documents. Finally, Section 2.6 describes multi-armed bandits, as an alternative to create pools of documents.

2.1 The Cranfield Paradigm

The Cranfield paradigm is at the core of offline evaluation and it was proposed by Cleverdon [92, 93] back in the sixties of the past century. Cyril Cleverdon was a librarian at the College of Aeronautics at Cranfield in the UK and he had the goal of comparing alternative strategies for the (manual) indexing of books and papers in the library, in order to understand which one was more effective for searching and retrieving the library holdings.

The Cranfield paradigm is based on *experimental collections* $C = (D, T, A)$ where: a corpus of documents D represents the domain of interest; a set of topics T represents (a surrogate of) the user information needs; and, human-made relevance assessments A are the "correct" answers, or ground-truth, determining, for each topic, the relevant documents[2]. Voorhees [408] provides a detailed discussion of the evolution of this fundamental paradigm over the years.

[2]Relevance assessments are often also called *relevance judgements* or *qrels*; all these terms are used interchangeably in the field.

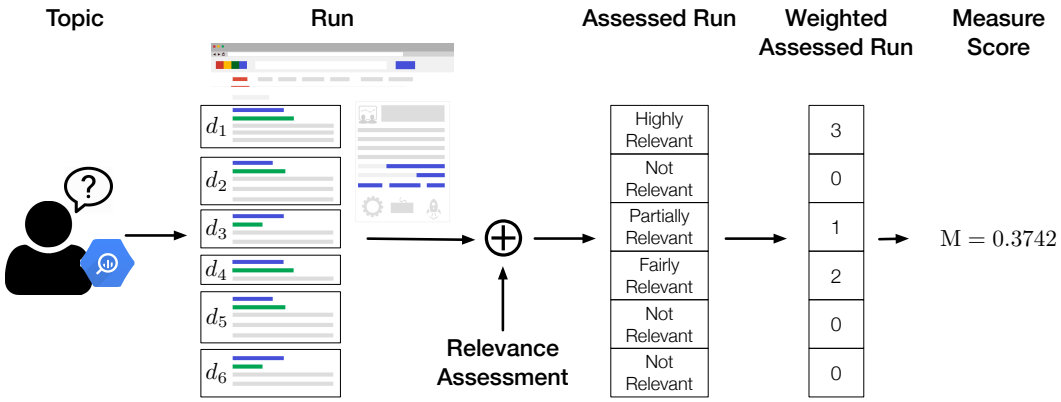


Fig. 1. Offline evaluation according to the Cranfield paradigm.

Figure 1 shows how offline evaluation works according to the Cranfield paradigm. Each topic $t \in T$ is used as input to the examined IR system which searches the corpus of documents D against that topic and produces a ranked result list of documents $d_1, d_2, d_3, \ldots$ where the higher the document in the ranking, i.e., d_1, the higher is the estimate of its relevance by the system. This list, i.e., the output of the IR system, is called the *run*. The relevance assessments A are then used to judge each of the retrieved documents and to produce the *assessed run*. Relevance assessments are typically expressed as either *binary relevance*, i.e., relevant or not relevant, or as *graded relevance*, e.g., not relevant, partially relevant, highly relevant, as proposed by Kekäläinen and Järvelin [228]. Assessments in the assessed run are then mapped to weights and this originates the *weighted assessed run*. Typically, 0 and 1 are used in the case of binary relevance and integer numbers in the case of graded relevance; however, recent studies, e.g., Maddalena et al. [266], used more articulated mappings. The weighted assessed run is finally scored using one of the many evaluation measures M available in IR. Since an experimental collection consists of many topics, the performance of an IR system is characterized by a set of scores, one for each topic. Therefore, when comparing two systems, we can aggregate such scores, e.g., by comparing the mean performance of the two systems, or even better, we can use such scores to conduct a *statistical significance test* between the two systems.

The main goal of this experimental setup is to be able to compare the performance of different IR systems in a robust and repeatable way, as they are all scored with respect to the same experimental collection. From an experimental setup point of view, as observed by Fuhr [167] and shown in Figure 2, experimental collections and evaluation measures are *controlled variables*, since they are kept fixed during experimentation; IR systems are *independent variables*, since they are the object of experimentation, compared one against the other; and, performance scores are the *dependent variables*, since their observed value changes as IR systems change.

This paradigm seems straightforward and intuitive but implementing it in a proper way is not trivial. It is necessary to guarantee that the experimental results are scientifically *valid*, first of all in terms of *internal validity*, i.e., "the ability to draw conclusions about causal relationships from the results of a study" [100, p. 157]. Some examples of the issues you have to face are: how do you sample documents and topics to avoid biases? How do you create relevance assessments? How "big" should experimental collections be to guarantee statistically reliable inferences? How many documents? How many topics? How many relevance assessments?

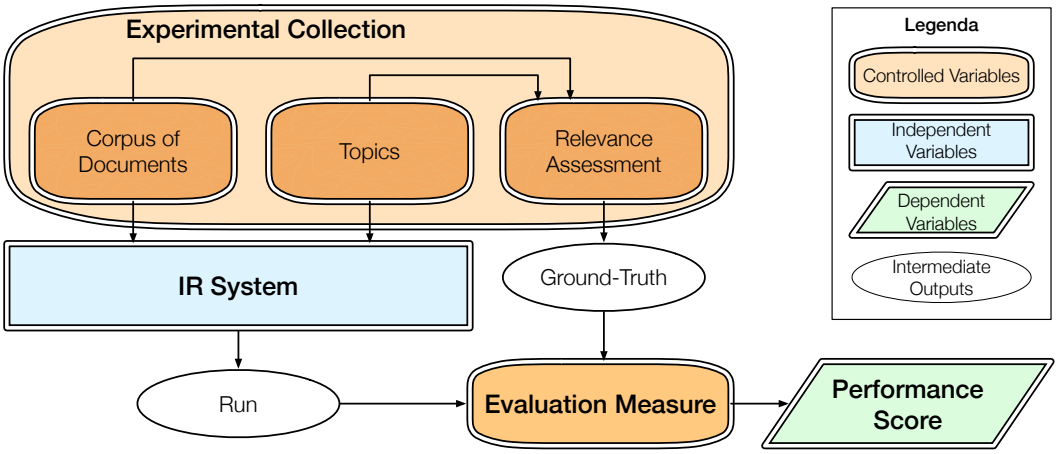


Fig. 2. The Cranfield paradigm from an experimental setup point of view.

A rigorous methodology and validity are not the only challenges. Another major concern of experimental evaluation is to be *realistic*. IR applications cover a wide range of domains: Web search, medical and eHealth search, product search and e-commerce, intellectual property search, digital libraries, just to name a few. Evaluating an IR system in a specific domain requires to develop an experimental collection which reasonably represents that domain. Therefore, you need to gather real documents and topics from a domain and you cannot rely on synthetic data, as it is common in other areas of computer science. For example, in the case of Web search, you need to crawl a reasonable amount of Web pages – which means in the order of millions and more, considering the actual size of the Web – and sample topics from the query log of a real search engine.

Finally, experimental evaluation according to the Cranfield paradigm does not only allow us to compare two (or more) IR systems in order to determine which one performs best, but it also perfectly fits in the above vision of IR system development as a two-stage process, constituted by an offline and an online development phases. Indeed, this evaluation paradigm is the same adopted by *Learning to Rank (LtR)* [257] in the offline phase, where experimental collections, split into a training, a validation, and a test set, become the annotated data used for the learning process and evaluation measures are used to define the cost function of the learning algorithms.

2.2 Evaluation Campaigns

Carrying out experimental evaluation according to the Cranfield paradigm is very demanding in terms of both the time and the effort required to prepare the experimental collection. Therefore, it is usually carried out in publicly open and large-scale evaluation campaigns, at international level, to share the effort, compare state-of-the-art systems and algorithms on a common and reproducible ground, and maximize the impact.

Spärck Jones and van Rijsbergen [373] wrote, in the early seventies of the last century, a seminal paper on how to implement the Cranfield paradigm in the context of a large-scale and shared evaluation campaign, but it took until 1992, when the *Text REtrieval Conference (TREC)*[3] [190] was initiated in the United States. The TREC conference series has constituted the blueprint for the organization of evaluation campaigns, providing guidelines and paving the way for others to follow.

[3]http://trec.nist.gov/

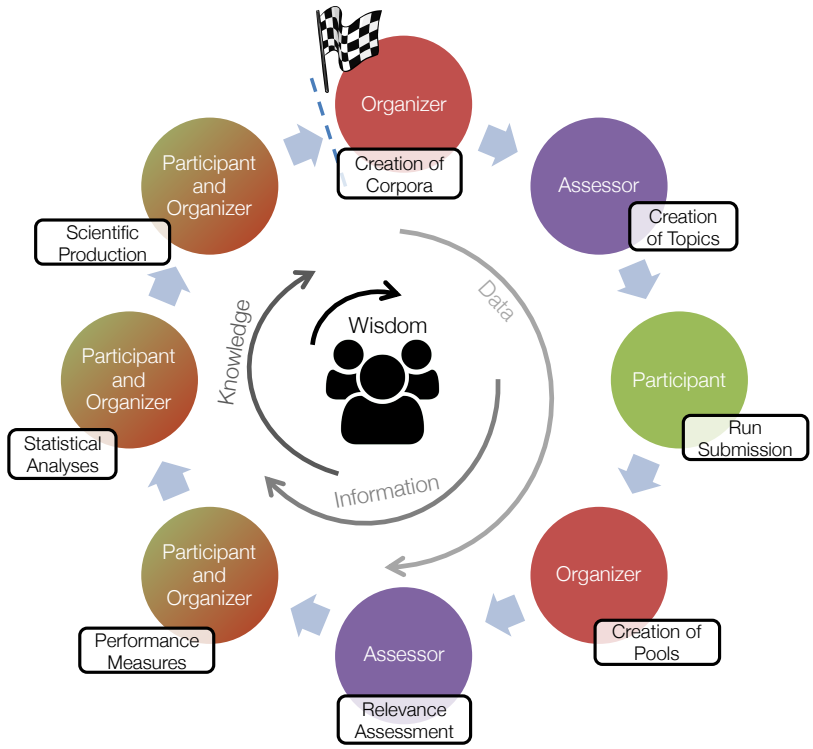


Fig. 3. Typical cycle of an evaluation campaign.

In 1999 the *NII Testbeds and Community for Information access Research (NTCIR)*[4] [342] was launched in Japan and Asia while, in 2000, the *Conference and Labs of the Evaluation Forum (CLEF)*[5] [154] was initiated in Europe. More recently, in 2008, the *Forum for Information Retrieval Evaluation (FIRE)*[6] began in India. Each of these initiatives has been studied to meet the perceived needs of a specific community, reflecting linguistic, cultural and resource differences, while being designed within a common theoretical framework. This common background has facilitated discussions and exchange of ideas among different groups, sometimes resulting in tasks run across evaluation campaigns. The aim is to avoid the duplication of effort and to provide complementary challenges, thus achieving a synergy of ideas and activities.

Figure 3 shows the typical cycle according to which an evaluation campaign proceeds. *Organizers* and *assessors* prepare document corpora and topics; then, *participants*, i.e. researchers and developers, run their systems on the provided document corpora and topics and produce their runs. These result lists are then sampled according to some criteria in order to create a *pool* of documents, which are then judged by assessors in order to produce the relevance judgments. At this point, performance measures, descriptive statistics, and statistical analyses are computed to evaluate the performance of each system and compare the proposed solutions. All of this information is then used for feeding the scientific production and the design and development of next generation systems. Sakai [340] provides a detailed discussion on how to setup and run an evaluation task according to

[4]http://research.nii.ac.jp/ntcir/
[5]https://www.clef-initiative.eu/
[6]http://fire.irsi.res.in/

this cycle. Dussin and Ferro [121] observed how this cycle can be framed within the well-know *Data, Information, Knowledge, Wisdom (DIKW)* hierarchy [163, 328]: during an evaluation campaign, we start from raw data, i.e., documents and topics; next, we obtain information by relating these raw data via runs, pools, and relevance assessment; then, we extract knowledge in terms of performance measures and statistical analyses; eventually, we distill wisdom by means of papers and insights on how to design the next generation of systems.

Shared evaluation campaigns have always played a central role in IR research. They have produced huge improvements in the state-of-the-art and helped strengthen a common systematic methodology, achieving not only scholarly impact [24, 245, 387, 392, 393] but also economic results [327], estimated in a return-on-investment about 3-5 times the funding provided. The experimental results accumulated by evaluation campaigns have also been used to conduct longitudinal studies [26, 57, 158, 232] to quantitatively track the progress and improvement of IR technologies over the years.

Finally, during their life-span, these large-scale campaigns also produce a huge amount of extremely valuable experimental data and require the support of proper software infrastructures to be operated. Over the years, this led to the development and adoption of various platforms, such as *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* [114, 121, 361], EvaluatIR [25], TIRA [309], TIREx [166], EvALL [13] as well as the Evaluation-as-a-Service approach [205]. Moreover, recent tools such as ir_datasets [264] and ir_metadata [55] facilitate the access to shared experimental collections and the description of system runs, streamlining the implementation and adoption of the Cranfield paradigm.

2.3 Document Corpora and Topics

As previously discussed, over the years, evaluation campaigns produced or relied on many document corpora. Below, you can find a not-exhaustive list of a few of them, just to get an understanding of their kind of documents, their size, and the purposes for which are used.

- **TIPSTER**[7]: 528,155 documents (news articles, US government reports, etc.), Disks 4 and 5 excluding Congressional Record subcollection. It was used in several TREC tracks, among which the Adhoc track [416, 417] and Robust track [402, 404].
- **WT10g**[8]: 1,692,096 Web pages crawled in 2001. It was used in the TREC Web track [192, 193];
- **GOV2**[9]: 25,205,179 Web pages crawled from .gov sites in early 2004. It was used in several TREC tracks, among which the TREC Web track [104, 105] and the TREC Terabyte track [85, 90].
- **CLEF Multilingual Corpus**[10]: 4,883,227 multilingual news articles corpus in 13 languages (Bulgarian, Dutch, English, Farsi, Finnish, French, German, Hungarian, Italian, Portuguese, Spanish) gathered in 1994, 1995 and 2002. Topics in 28 different languages (Bengali, Bulgarian, Chinese, Czech, Dutch, English, Farsi, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Marathi, Norwegian, Oromo Polish, Portuguese, Russian, Slovenian, Spanish, Swedish, Tamil, Telugu, Thai). It was used in several CLEF Adhoc tracks [5, 48–52, 115–117, 153]
- **ClueWeb 2009**[11]: 1,040,809,705 Web pages in 10 languages crawled between January and February 2009. It was used in several TREC tracks, among which the TREC Web track [86, 87]

[7] https://catalog.ldc.upenn.edu/LDC93T3A
[8] http://ir.dcs.gla.ac.uk/test_collections/wt10g.html
[9] http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
[10] http://catalog.elra.info/en-us/repository/browse/ELRA-E0008/ and http://catalog.elra.info/en-us/repository/browse/ELRA-E0036/
[11] https://lemurproject.org/clueweb09/

- **ClueWeb 2012**[12]: 733,019,372 English Web pages crawled between February 10, 2012 and May 10, 2012. It was used in several TREC tracks, among which the TREC Web track [96, 97].
- **ClueWeb 2022**[13]: 10 billion web pages, complemented with enriched information, such as the visual representation of pages rendered by a web browser, making it suitable for various tasks, including ranking, neural model pre-training, and language generation [303]. It was used in the CLEF 2023 Touché lab [45] and the TREC 2023 IKAT track[14].
- **Chuweb21D**[15]: 82.5 million English document collection for web search tasks [78]. It was used in the NTCIR-16 We Want Web with CENTRE (WWW-4) task [344] and the NTCIR-17 Fair Web (FairWeb-1) task [383].
- **The New York Times Annotated Corpus**[16]: 1,855,658 news articles from January 1987 through December 2007 from New York Times. It was used in the TREC Common Core track [8].
- **TREC Washington Post Corpus**[17]: 595,037 news articles and blog posts from January 2012 through August 2017 from Washington Post. It was used in several TREC tracks, among which the TREC Common Core track [9] and the TREC News track [369, 370].
- **MS MARCO**[18]: 3.2 million English documents, 8.8 million passages, 1 million questions. It was used in the TREC 2019 and 2020 Deep Learning tracks [106, 107].
- **MS MARCO V2**[19]: 11.9 million English documents, 138.3 million passages. It was used in the TREC 2021 to 2023 Deep Learning tracks [102, 103].

Figure 4 shows a topic taken from the CLEF multilingual collection. A topic typically consists of: *title*, a brief statement expressing the information need and resembling the typical search engine query; *description*, a more detailed formulation of the information need; and, *narrative*, instructions for assessors on when to consider a document relevant. In the case of CLEF, you can note how the content of these fields has been translated to several languages.

Topics are used to create the actual queries used by IR systems to search the document corpus. This may happen in two ways, either *automatic* or *manual*. In the automatic way, IR systems directly take topics as input and generate actual queries to be searched, typically using the title and/or description fields; the simplest way is to use the topic fields as they are, but other more sophisticated strategies are possible, such as boosting the weight of the terms of a field (typically the title field) or expanding and enriching the query. In the manual way, a person reads the topic and generates one or more queries corresponding to that topic which are then fed to the IR system.

Topics are thought to be (uniformly) sampled from an hypothetical distribution of all the possible topics and, thus, to be representative of this distribution. But, as said before, how many topics do we need to ensure that we are conducting statistically sound inferences? Academic experimental collections usually comprise at least 50 topics but, most often, a few hundreds of them, accumulated across several cycles of a campaign when using the same document corpus. In an industrial context, especially in Web search, it is common to use thousands of topics.

Spärck Jones and van Rijsbergen [373] suggested that under 75 topics there was no real value while 250 topics was a more acceptable size and more than 1,000 topics might be needed for some purposes. Buckley and Voorhees showed that the reliability of a single comparison of two IR

[12]https://lemurproject.org/clueweb12/
[13]https://lemurproject.org/clueweb22.php/
[14]https://www.trecikat.com/
[15]https://github.com/chuzhumin98/Chuweb21D
[16]https://catalog.ldc.upenn.edu/LDC2008T19
[17]https://trec.nist.gov/data/wapost/
[18]https://microsoft.github.io/msmarco/
[19]https://microsoft.github.io/msmarco/TREC-Deep-Learning.html

```
<?xml version="1.0" encoding="UTF-8"?>
<topic>
    <identifier>41</identifier>

    <title lang="en">Pesticides in Baby Food</title>
    <title lang="fr">Des pesticides dans la nourriture pour bébés</title>
    <title lang="it">Pesticidi negli alimenti per bambini</title>
    <title lang="ru">Пестициды в детском питании</title>
    <title lang="zh">嬰兒食品中含有殺蟲劑</title>
    <title lang="ja">ベビーフード中の病虫害防除剤</title>
    <title lang="th">ยาฆ่าแมลง ใน อาหาร เด็กอ่อน</title>
    <title lang="so">Sunta cayayaanka ee Cuntada Ilmaha</title>
    <title lang="sw">Dawa za kuulia wadudu katika Chakula cha Mtoto</title>

    <description lang="en">Find reports on pesticides in baby food.</description>
    <description lang="fr">
        Rechercher des documents sur les pesticides dans la nourriture pour bébés.
    </description>
    <description lang="it">
        Trova documenti che parlano dei pesticidi negli alimenti per bambini.
    </description>
    <description lang="ru">Найти статьи о пестицидах в детском питании</description>
    <description lang="zh">查詢有關嬰兒食品中含有殺蟲劑的報導。</description>
    <description lang="ja">ベビーフード中の病虫害防除剤に関する記事を探したい。</description>
    <description lang="th">หา รายงาน ที่ เกี่ยวของ กับ ยาฆ่าแมลง ใน อาหาร เด็กอ่อน</description>
    <description lang="so">Hel wargelinada sunta cayayaanka ee cuntada ilmaha.</description>
    <description lang="sw">
        Pata ripoti kuhusu dawa za kuulia wadudu katika chakula cha mtoto.
    </description>

    <narrative lang="en">
        Relevant documents give information on the discovery of pesticides in baby food.
        They report on different brands, supermarkets, and companies selling baby food
        which contains pesticides. They also discuss measures against the contamination
        of baby food by pesticides.
    </narrative>
    <narrative lang="fr">
        Les documents pertinents informent sur la découverte de pesticides dans la
        nourriture pour bébés. Ils contiennent des informations sur les différentes
        marques, les supermarchés et les firmes ayant mis en vente de la nourriture pour
        bébés renfermant des pesticides. Ils relatent également les mesures prises contre
        la contamination de la nourriture pour bébés par les pesticides.
    </narrative>
    <narrative lang="it">
        I documenti rilevanti forniscono informazioni sulla scoperta di pesticidi nei
        cibi per bambini. Riportano i diversi marchi, i supermercati e le ditte che hanno
        venduto alimenti per bambini con i pesticidi. Sono anche rilevanti i documenti
        che discutono le misure contro la contaminazione degli alimenti per bambini con
        i pesticidi.
    </narrative>
</topic>
```

Fig. 4. Example of topic from the CLEF multilingual collection.

systems depends on the used evaluation measure, the magnitude of the difference in the scores, and the number of topics, being 25 topics a bare minimum [58, 412]. Sanderson and Zobel [348] then showed that 25 topics are definitely not enough and, later on, Voorhees [406] reported that even 50 topics may be not enough to draw strong conclusions. Finally, Sakai [338] has proposed a statistically-grounded method to determine how many topics are needed to meet a set of targeted statistical requirements.

2.4 Pooling

When creating experimental collections a question immediately arises: how to perform relevance assessment? Even if one of the "holy grails" of the field is to perform relevance assessment in an automatic way [325], it substantially remains a human task where an assessor needs to inspect each document and to decide whether or not and/or to what extent that document is relevant for the topic at hand. Given the size of the document corpora, it is clearly infeasible to assess each document in the corpus with respect to a given topic; even more, when you consider that the

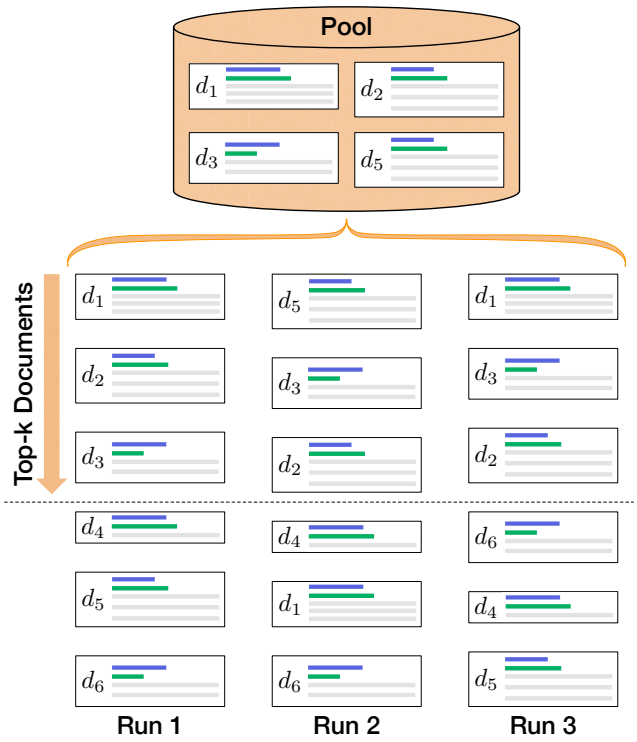


Fig. 5. Top-k pooling.

process should be repeated for each topic in the collection. Therefore, some sampling techniques are needed to select a manageable subset of documents to be actually inspected by an assessor. Unfortunately, uniform sampling would not work because there usually are very few relevant documents for a topic with respect to the corpus size and you would need to sample way too many documents to obtain a reasonable fraction of the relevant ones.

Figure 5 shows the traditional *top-k pooling*, which is built as follows: (1) the threshold k is fixed ahead (typically $k = 100$); (2) the top k documents from each run are selected; (3) duplicated documents are removed. This constitutes the pool of documents which will be actually judged by the human assessors for a given topic. The same procedure is repeated for each topic in order to create the overall pool. It is assumed that documents which are not pooled, i.e., those below the threshold k, are not relevant, even if they are not actually judged by an assessor. This assumption ensures that documents which are retrieved by one run, but not actually judged, have a relevance assessment anyway, making it possible to compute the performance score for the run.

Pooling represents a further motivation for conducting experimental evaluation in large-scale evaluation campaigns. To obtain a good quality pool, you need a good number of diverse runs, where diverse means that there is a small overlap between runs and they retrieve different documents. In this way, the pool will contain different documents and this will increase the probability of sampling a good fraction of all the existing relevant documents for that topic. In turn, having (almost) all the relevant documents for a topic in the pool will allow an accurate and reliable scoring of system performance and comparison among IR systems.

But what does a good quality pool mean? Experimental collections are demanding resources to be created and you wish to re-use them also beyond the specific campaign cycle where they have been created. This means that you wish to use them to evaluate IR systems which have not participated in that campaign cycle and which have not taken part in the pooling process, i.e., whose documents are (potentially) not part of the pool. Therefore, a good quality pool should not be biased towards systems whose documents were included in the pool and it should allow a robust comparison of IR systems even if they did not participate in the pooling process, reliably determining which one is better in both cases. Zobel [435] has shown that top-k pooling is a robust approach for this purpose.

On the other hand, Rashidi et al. [316] have shown that IR evaluation measures (see Section 3), especially those based on the recall base, could be fragile in case of judgment errors, leading to changes in how systems are ordered by their performance.

2.5 Crowdsourcing

One of the main drawbacks of the traditional pooling approach is to be extremely demanding in terms of resources needed to perform it. Reading and judging documents is a lengthy process and you need to hire assessors who are often professionals – in the early days of TREC they were retired CIA analysts for their expertise in seeking for specific information through huge amounts of documents – thus making relevance assessment a quite costly effort.

Crowdsourcing [11, 12, 233, 247, 270] has emerged as a viable option for relevance assessment since it allows to quickly gather judgments from crowd-assessors and even to cheaply collect multiple assessments for each document. However, it raises many questions regarding the quality of the collected assessments since crowd-assessors may not have the same expertise as professional assessors and may not pay the same level of care in performing the task. Therefore, in order to obtain relevance assessments good enough to be used for evaluation purposes, the possibility of discarding low quality crowd-assessors and/or combining them with more or less sophisticated algorithms has been considered.

Research in crowdsourcing has focused on several different issues: aggregating labels from multiple assessors to improve the quality of the gathered assessments, by using unsupervised [34, 206], supervised [306, 318, 319], and hybrid [191] approaches; behavioural aspects [184, 185, 227]; proper and careful design of *Human Intelligent Tasks (HITs)* [10, 179, 210, 226], also using gamification to improve quality [124] and game theory to increase crowd-assessors engagement [290] and judgement quality [289]; routing tasks to proper crowd-assessors [220, 246], as well as the effect of relevance scales on relevance judgments made by crowd-workers [324].

The problem of merging multiple crowd-assessors has been addressed mostly from a classification point of view, i.e., choosing among the set of possible relevance assessments (labels) those best supported by the evidence provided by the crowd-assessors. In detail, traditional approaches typically determine the "best" relevance assessment, combining those produced by multiple crowd-assessors according to some criteria, use them to compute an evaluation measure, and score IR systems. We call this an *upstream* approach, because the aggregated relevance assessments are created before systems are evaluated and performance scores are computed.

The most common, and still very effective, of these family of upstream approaches is *Majority Vote (MV)* [382]: it assigns to each document the most popular assessment among those expressed by crowd-assessors; to deal with variable quality workers, several weighted versions of MV have been proposed, e.g., [388]. *Expectation Maximization (EM)* [34, 206] addresses the problem in a probabilistic way, by iteratively estimating the probability of relevance of each document and then by assigning it the most probable assessment. Ferrante et al. [138, 144] have proposed a stochastic vision to relevance assessment, where each relevance judgement is a binomial random variable

whose expectation p indicates the quantity of relevance assigned to a document. They leveraged this vision to define a *Binomial Majority Vote (BMV)* strategy, where the merged amount of relevance of each topic/document pair is estimated from the observed values of the binomial random variables associated to each crowd-assessor.

Gaussian Processes are another technique adopted to merge and denoise crowd-workers data [251, 252, 330] by assuming that the label generation process can be represented through a probabilistic graphical model which allows for learning latent true labels from the observed crowd labels.

On a different stance, Ferrante et al. [142] proposed *Assessor-driven Weighted Averages for Retrieval Evaluation (AWARE)*, a new *downstream* approach. AWARE is motivated by the observation that upstream approaches, e.g., MV, choose the "best" relevance assessment ahead, at the pool level, disregarding how this choice may affect different IR systems and evaluation measures. For example, consider two systems that retrieve the same document but at different rank positions. Then, a correct or wrong label for this document will affect the performance of these two systems in different ways. Moreover, IR evaluation measures score systems differently, depending on how the measure weights the relevance of a document, its rank position, and so on (see Section 3). For example, a mislabelled document at rank position 10, can have different effects depending on the user model underlying the evaluation measure (see Sections 3 and 5). If a measure assumes a patient user who goes deeply in the ranking, the mislabelled document will affect the measure score to a possibly great extent. On the other side, if a measure consider an impatient users, that examines documents up to rank position 3 (e.g., with a mobile screen), then the mislabelled document at position 10 will not have an impact on the measure score. As a consequence, even a small assessment error over a whole pool of documents may affect IR systems and evaluation measures in quite different ways. AWARE addresses these issues with a probabilistic downstream framework, which works as follows. First, evaluation measures are computed with respect to relevance assessments from each crowd-assessor, this happens without merging relevance assessments as in upstream approaches. Then, the measure scores, computed from each crowd-assessor separately, are merged by optimizing parameters that account for both the systems and the measures under consideration.

2.6 Multi-armed Bandits

Both traditional top-k pooling and crowdsourcing approaches assume that the pool of documents is prepared ahead, according to some sampling strategy, afterwards the assessors, being them experts or crowd-workers, judge the documents in the pool. In both cases, the depth k of the pool (or any other sampling parameter) depends on the trade-off between finding as many relevant documents as possible and the resources (time and effort) available for ground-truth creation. In general, the higher the value for k, the higher the chance of finding more relevant documents but this can also sensibly increase the total size of the pool and thus the resources required to judge its documents.

An alternative which is emerging recently is to *adaptively* select which is the next document to be judged in such a way that maximizes the total number of relevant documents found, given a fixed amount of resources for carrying out the relevant assessment. Indeed, not all the runs are equally good in contributing relevant documents to the pool and sampling the same number of documents from all the runs, as it happens with top-k pooling, may waste resources in assessing not relevant documents. On the contrary, we would like to sample more documents from those runs with an higher chance of contributing relevant documents.

Multi-armed bandits [365] are a reinforcement learning problem where you have a fixed amount of resources and you can pick from a set of alternative choices in order to maximize some gain. The name armed bandits comes from gambling: the player is in front of a row of slot machines, also called *one-armed bandits*, and she/he has to decide which one to play, i.e., which lever (arm) to pull. The goal of the gambler is to maximize the amount of money won, given her/his fixed budget.

To this end, the gambler has to decide how many times to play with a machine (exploitation) and when to try a different machine (exploration) based on the estimate of the expected payoff from each machine.

This corresponds to an algorithm with K possible actions, i.e., the *arms*, to choose from and T rounds to perform, i.e., the limited resources. In each round, the algorithm picks an action and this produces a reward, which comes from an unknown distribution depending on the chosen action. Round after round, the algorithm estimates such distribution better and better thanks to the collected rewards. However, the algorithm is challenged with an *exploitation* vs *exploration* trade-off: if it always picks the same action, how does it know if another action would be more rewarding? Therefore, in some rounds it has to pick the same action (exploitation) but in some other rounds it has to try another action (exploration), relying on the estimates of the unknown distributions and in such a way that maximizes the total gain at the end of the T rounds.

Losada et al. [259, 260] applied multi-armed bandits approaches to the construction of a pool. Each run represents a one-armed bandit and, at each round, the algorithm has to decide from which run to pick the next document to be assessed; the reward is whether the chosen document is relevant or not; the fixed amount of resources is the number of rounds, i.e., the total number of documents to be pooled.

Bandits approaches may suffer from some issues. They may lack diversity, i.e., they may keep picking documents from the same run, favoring exploitation over exploration, if that run contains many relevant document. Even if this may maximise the total number of relevant documents found, this may also introduce bias in the pool. Indeed, as discussed in Section 2.4, relevant documents coming from different runs are preferable because they make the pool more robust and fair also for runs which have not participated in the pool. Bandits methods may also penalize "slow-start" runs, i.e., those runs which do not immediately retrieve relevant documents in the top ranks positions but may be able to retrieve many of them in slightly lower ranks. In this respect, Voorhees [407] adopted multi-armed bandits approaches in the construction of the TREC CORE 2017; she concluded that the greedy approach common to most bandit methods can be unfair even to the runs participating in the collection-building process when the judgment budget is small relative to the (unknown) number of relevant documents. A study by Lipani et al. [256] compares many traditional and adaptive pooling strategies in order to determine not only which one is the most effective in maximising the total number of relevant documents but also which one is the less biased, leading to more reusable experimental collections. Otero et al. [301] studied the effectiveness of multi-armed bandits approaches from a different point of view, i.e. their impact on the statistically significant differences among systems, in order to be able to more reliably ansewer the question "is system A significantly better than system B?".

Multi-armed bandits approaches are not limited to pooling strategies but they can also be used in crowdsourcing. Indeed, as discussed in the previous section, to ensure the quality of the collected relevance judgments, the same HIT, i.e., a (topic, document) pair, is judged by more crowd-workers. How many crowd-workers for each HIT is typically decided ahead but this might be a not optimal allocation of resources. Indeed, some pairs may be easier to assess than others, e.g., the document (url) `www.facebook.com` for the query `facebook`, and they may require less crowd-workers to ensure a high quality judgement. Therefore, also in this case, we are faced with a quality/cost trade-off that can be suitably modeled as an exploration/exploitation trade-off with multi-armed bandits. In this context, Abraham et al. [2] have proposed an algorithm, based on multi-armed bandits approaches, to dynamically decide when to stop assigning crowd-workers to a HIT, i.e., when the quality of the judgement is satisfactory, showing that the proposed algorithm performs significantly better than assigning the same number of crowd-workers to each HIT.

3 EVALUATION MEASURES

As it emerges from the discussion in the previous section, offline evaluation abstracts away many details of how and why users interact with IR systems in real settings, in order to provide a very controlled environment which allows for repeatedly running experiments in a replicable way. In this context, evaluation measures not only quantify the effectiveness of IR systems, but they also bring back some notion of user by embedding the so-called *user models*, which provide an abridged template of the user behaviour when scanning and interacting with the ranked result list. Therefore, it becomes crucial how much realistic such user models are, since they shape how close to actual users we are in quantifying IR system performance.

Carterette [67] pointed out that model-based measures are actually composed from three distinct underlying models:

(1) a *browsing model* that describes how a user interacts with results;
(2) a model of *document utility*, describing how a user derives utility from individual relevant documents;
(3) a *utility accumulation* model that describes how a user accumulates utility in the course of browsing.

Carterette considers as browsing model that of a user scanning down ranked results one-by-one and stopping at some rank k and, therefore, he models the probability distribution of the *stopping rank*. As a model of document utility, he makes use of binary or graded relevance judgments. Finally, the utility accumulation model depends on the actual evaluation measure.

As an alternative, Moffat et al. [285, 286, 287] proposed the C/W/L framework, and its extensions, as a theoretically principled approach to describe and design IR evaluation measures, where W is a probability distribution of weights for each rank position, C is the conditional probability that the users continues scanning from rank i to $i + 1$, and L is the probability that the i-th document in the ranking is the last one observed by the user.

3.1 Precision and Recall

Precision and *Recall* are two fundamental measures in IR since its inception [398], where Precision is the the proportion of retrieved documents that are actually relevant, while Recall is the the proportion of relevant documents actually retrieved. Together, Precision and Recall measure retrieval effectiveness, meant as the ability of a system to retrieve relevant documents while at the same time holding back non-relevant ones, and they capture the trade-off between retrieving the more relevant documents possible (Recall) and rejecting not relevant ones (Precision). As observed by van Rijsbergen [400], maximizing Precision and Recall corresponds to optimal retrieval in the sense of the Probability Ranking Principle, i.e., ordering documents by their decreasing probability of being relevant. This connection between Precision/Recall and optimal retrieval creates a tight link between retrieval models and evaluation.

Figure 6 shows the relation between relevant and retrieved documents in the corpus of documents. The set of relevant documents A is the set of relevance assessments of the experimental collection. The set of retrieved documents B is what in Figure 1 is called a run. Eq. (1) defines Precision and Recall in terms of the sets shown in Figure 6.

$$\begin{aligned} P &= \frac{|A \cap B|}{|B|} \\ R &= \frac{|A \cap B|}{|A|} \end{aligned} \tag{1}$$

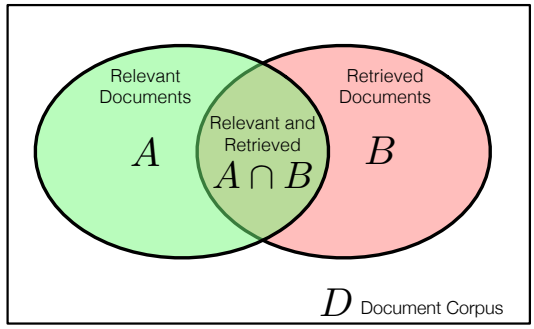


Fig. 6. Relation between relevant and retrieved documents.

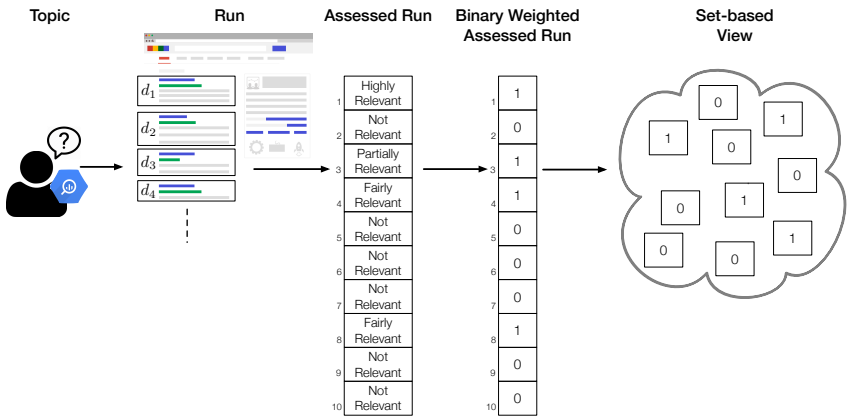


Fig. 7. Example of computation of binary evaluation measures for a single run.

Precision and Recall are set-based evaluation measures, meaning that they do not consider the ranking of documents produced by a system but just which documents are retrieved. In this respect there is no browsing model in the sense of Carterette because it is like if the user inspects the whole list in one single shot. Precision and Recall are binary measures and, thus, the document utility model is to consider 0 for not relevant documents and 1 for relevant ones (in case of graded relevance 1 for whatever above not relevant). Precision and Recall rely on the cardinality of the different sets involved and, thus, the utility accumulation model is just counting how many documents there are in the different sets.

Note that Recall requires to know $|A|$, i.e., the total number of relevant documents for a topic, and this is more a sort of assumption rather than something we can really know exactly. Indeed, if you consider how pooling happens and relevance assessment is performed, we can just have an estimate of $|A|$ and the better this estimate the better the quality of the pool we created.

Figure 7 shows an example of how to compute Precision and Recall for a single run. Assume that: the run retrieves 10 documents, i.e., $|B| = 10$; there are 8 relevant documents for this topic, i.e.,

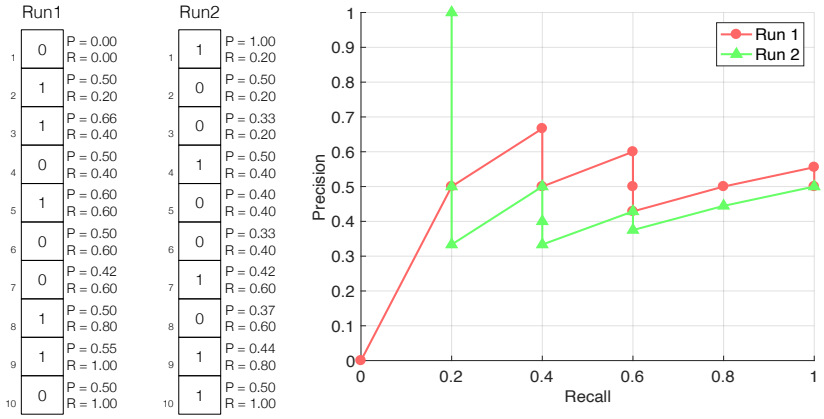


Fig. 8. Example of Precision-Recall curve.

$|A| = 8$; and, 4 documents retrieved by the run are relevant, i.e., $|A \cap B| = 4$. Therefore, we have:

$$P = \frac{4}{10} = 0.40$$
$$R = \frac{4}{8} = 0.50$$

We can adopt a rank-based view to Precision and Recall by computing them at a given document cut-off value k:

$$P(k) = \frac{1}{k} \sum_{i=1}^{k} \hat{r}_t[i]$$
$$R(k) = \frac{1}{RB} \sum_{i=1}^{k} \hat{r}_t[i] \tag{2}$$

where $\hat{r}_t[i] \in \{0, 1\}$ is the relevance degree of the i-th document in the run and $RB = |A|$ is the so-called *recall base*, i.e., the total number of relevant documents for a topic.

If we go back to the example of Figure 7, we obtain that:

$$P(5) = \frac{3}{5} = 0.600$$
$$R(5) = \frac{3}{8} = 0.375$$

We can introduce also the *Rprec* measure, which corresponds to $P(RB)$, i.e., Precision computed at a document cut-off equal to the recall base. Note that $k = RB$ is the rank position at which it is possible to achieve perfect retrieval, since the system had enough rank positions to potentially retrieve all the relevant documents, and it is also the rank position at which $P(RB) = R(RB)$. In our example we have that $Rprec = P(8) = \frac{4}{8} = 0.50$.

3.2 Precision-Recall Curve and Interpolated Precision

Let us consider the two runs shown in Figure 8 and let us assume that $RB = 5$. If we go step-by-step from the top to the bottom rank position, we can compute $P(k)$ and $R(k)$ for each rank position and plot them in the so-called *precision-recall curve*. This curve allow us to determine which level of Precision we achieved at a given level of Recall, abstracting from the specific rank positions. For

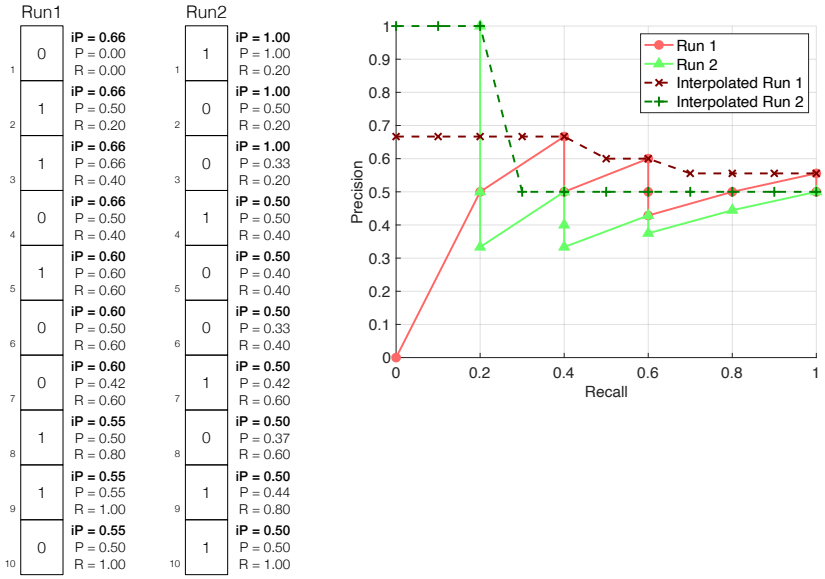


Fig. 9. Example of interpolated Precision at standard Recall values curve for the runs of Figure 8.

example, we can see that Run1 achieved 60% of Precision at the 60% of Recall. If we consider that maximizing Precision and Recall corresponds to optimal retrieval [400], we can understand the importance of the precision-recall curve and how the *Area Under the Curve (AUC)* is an indicator of the overall effectiveness of an IR system. Indeed, the greater the AUC, the better the performance of the system.

Unfortunately, the precision-recall curve has this typical saw-tooth shape where we may have multiple Precision values for the same Recall value as, for example, it happens for Run1 at 60% of Recall. Moreover, it may be difficult to compare runs because they may not have the same Recall values; for example, Run2 does not have a Precision value for 0% of Recall as Run1 does instead.

For all these reasons, Precision is interpolated as follows:

$$iP(R_j) = \max_{R \geq R_j} P(R) \tag{3}$$

where, for a given standard Recall value R_j, the interpolated precision iP is the maximum Precision obtained for any actual Recall value R greater than or equal to R_j. The typical standard Recall values correspond to 11 points: $R_j \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

Figure 9 shows how the interpolated precision at standard recall values curve looks for the runs of Figure 8. $\times$ and + markers are used to show the values of iP at the 11 standard recall values for Run1 and Run2, respectively. For example, in the case of Run1 there is no precision value for $R = 0.3$; the maximum precision value for a recall $R \geq 0.3$ is $P = 0.66$ which happens at $R = 0.4$ (rank positions 3 and 4); therefore, the interpolated precision is $iP = 0.66$ for $R = 0.3$.

Figure 10 shows the interpolated precision at standard recall values curve in the case of real runs submitted to the CLEF 2009 Adhoc Persian track [153]. The figure shows the typical inverse relationship between Precision and Recall which well captures the trade-off at the core of optimal retrieval and which has been deeply studied [56, 94, 123]. This kind of plots is one of the most widely used to compare the performance of IR systems.

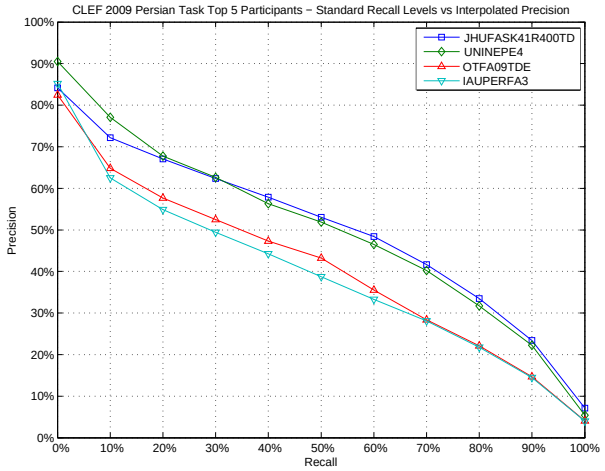


Fig. 10. Interpolated Precision at standard Recall values curve for the top runs of CLEF 2009 Adhoc Persian track, taken from [153].

3.3 Average Precision

As discussed in Section 3.1, the goal of a good IR system is to jointly optimize Precision and Recall. However, Precision and Recall are set based, thus their main limitation is that they account solely for the proportion of relevant documents retrieved and not the rank positions where relevant documents are actually retrieved. *Average Precision (AP)* aims at addressing this issue, while at the same time optimizing Precision and Recall. AP is defined as follows:

$$\begin{aligned} AP &= \frac{1}{RB}\sum_{k\in\mathcal{R}} P(k) = \\ &= \frac{1}{RB}\sum_{n=1}^{N}\left(\frac{1}{n}\sum_{m=1}^{n}\hat{r}_t[m]\right)\hat{r}_t[n] = \underbrace{\frac{rr}{RB}}_{\text{Recall}} \cdot \underbrace{\frac{1}{rr}\sum_{k\in\mathcal{R}} P(k)}_{\text{arithmetic mean of } P(k)} \end{aligned} \tag{4}$$

where $\mathcal{R}$ is the set of the rank positions of the relevant retrieved documents; $rr = |\mathcal{R}|$ is the total number of relevant retrieved documents ($|A \cap B|$ in Figure 6); and, N is the total number of retrieved documents, i.e., the length of the run.

The original definition of AP by Buckley and Voorhees [60] is to be the sum of the Precision achieved at each relevant retrieved document ($k \in \mathcal{R}$) averaged by the recall base, as shown in the first row of Eq. (4). The motivation behind AP is to provide a single-score summary for the overall effectiveness of an IR system, since it is known to also correspond to the AUC, very conveniently summarizing the whole precision-recall curve. If we look at the second row of Eq. (4), we can observe that AP is actually given by the aritmetic mean of the Precision achieved at each relevant retrieved document times the Recall achieved by the system; this let us further understand how it fully embeds the precision-recall trade-off and in which sense it can be an average.

If we go back to the example of Figure 7, AP is computed as follows

$$AP = \frac{1}{RB}\Big(P(1) + P(3) + P(4) + P(8)\Big) = \frac{1}{8}\left(1 + \frac{2}{3} + \frac{3}{4} + \frac{4}{8}\right) = \frac{35}{96} = 0.36$$

Mean Average Precision (MAP) is the arithmetic mean of AP over the set of topics. Differently from the other measures, this mean has its own name since it is the most widely used single number to summarise the whole performance of a system and to compare systems.

AP represents the "gold standard" measure in IR [427], known to be stable [58] and informative [27], with a natural top-heavy bias[20] and an underlying theoretical basis as approximation of the AUC. Nevertheless, due to its dependence on the recall base, it assumes a perfect knowledge of the relevance of each document in the collection, which is an approximation when pooling is adopted and not assessed documents are assumed to be not relevant [187], and is even more exacerbated in the case of large scale or dynamic collections [59, 427].

However, the strongest criticism to AP comes from the absence of a convincing user model for it, a feature which is deemed extremely important in order to make the interpretation of a measure meaningful and to bridge the gap between system-oriented and user-oriented studies [67, 287, 368]. In this respect, Moffat and Zobel [288] argued that the model behind AP is abstract, complex, and far from the real behavior of users interacting with an IR system, especially when it comes to its dependence on the recall base which is something actually unknown to real users. As a consequence, Robertson [322] proposed a simple but somehow plausible user model for AP, which allows for a mix of different behaviors in the population of users.

3.4 Discounted Cumulated Gain

When Järvelin and Kekäläinen [214] proposed *Discounted Cumulated Gain (DCG)*, it was an innovative measure for several reasons:

(1) it relied on graded relevance instead of binary relevance. The document utility model is able to distinguish between different types of relevant documents and to accrue different levels of utility from them;
(2) it relied on an explicit browsing model of a user. The browsing model considers a user that sequentially scans the ranked result list from the top to the bottom;
(3) it relied on an utility accumulation model, where the utility provided by a document is discounted proportionally to the rank position at which that document is retrieved. The underlying idea is that a relevant document retrieved at the top of the ranking can be more useful to the user than, the same document retrieved at the bottom of the ranking.

DCG is defined as follows:

$$\begin{aligned} DCG(k) &= \begin{cases} \sum_{i=1}^{k} \hat{r}_t[i] & \text{if } i < b \\ DCG(k-1) + \frac{\hat{r}_t[k]}{\log_b(k)} & \text{if } k \geq b \end{cases} \\ &= \sum_{i=1}^{k} \frac{\hat{r}_t[i]}{\max(1, \log_b(k))} \end{aligned} \tag{5}$$

where b is the base of the logarithm used for discounting. As you can note from Eq. (5), DCG does not apply the discount for rank positions below the logarithm base; for this reason the logarithm base is often interpreted as the *patience* of the user, where $b = 2$ models an impatient user and $b = 10$ a patient one[21]. The original formulation of DCG by Järvelin and Kekäläinen uses a recursive definition, as shown in the first row of Eq. (5). In the second row of Eq. (5) we provide an alternative and iterative definition, which is more convenient for computation.

[20] A top-heavy measure is a measure which rewards more the relevant documents retrieved in the top rank positions.
[21] Log base $b = 2$ is considered an impatient users since she/he starts discounting documents, i.e., valuing them less, as early as the second rank position.

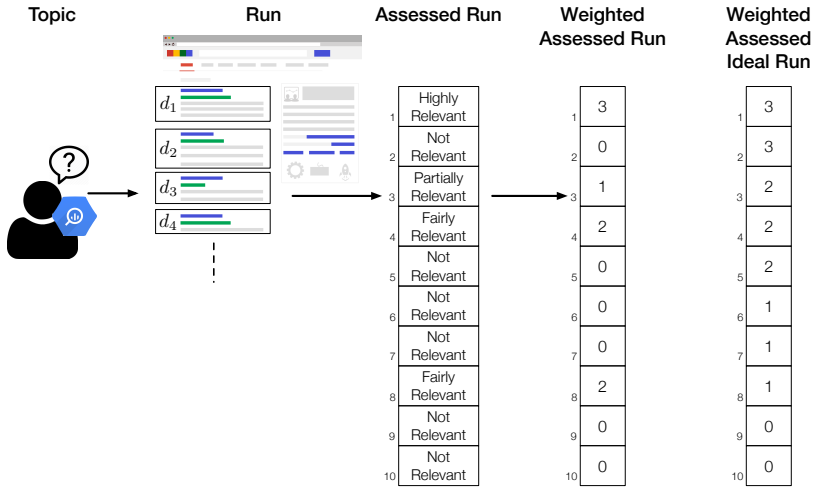


Fig. 11. Example of computation of graded evaluation measures for a single run.

Figure 11 shows an example on how to compute DCG for a single run. Assume that the run retrieves 10 documents, whose relevance degrees are indicated in the figure, and $b = 2$, i.e., an impatient user. DCG is given by:

$$DCG = 3 + \frac{1}{\log_2(3)} + \frac{2}{\log_2(4)} + \frac{2}{\log_2(8)} = 5.2976$$

While being intuitive to use, DCG suffers from the limitation of not being normalized in a specific range. This makes the measure scores hard to interpret. For example, it is not clear if DCG = 5.3 above represents a good or bad performance score. In order to normalize DCG, we need to consider the so-called *ideal run*, i.e., the run generated by ranking all the relevant documents in the pool in decreasing order of their relevance (the rightmost run of Figure 11, assuming 8 relevant documents in the pool). The *Normalized Discounted Cumulated Gain (nDCG)* is defined as

$$nDCG(k) = \frac{DCG(k)}{iDCG(k)} \tag{6}$$

where *iDCG* is the DCG score of the ideal run.

In our example, we obtain:

$$DCG = 5.2976$$
$$iDCG = 10.1996$$
$$nDCG = 0.5194$$

Note that while DCG is independent from the recall base, a somewhat desirable property for an evaluation measure, nDCG is not, since the ideal run is a materialization of the recall base.

Measures for Other Retrieval Tasks

The IR measures described in the previous sections are among the most commonly used measures and they are designed primarily for ad-hoc retrieval tasks, i.e. returning a list of possibly relevant documents in response to a user query, but they are often used also in other contexts, e.g. DCG is a measure very commonly used also for *Recommender Systems (RSs)*.

Nevertheless, there are many other measures which have been developed with specific tasks in mind, among which particularly worth of mention are *diversity* and *fairness*.

Diversity is concerned with how good are retrieval results at representing different angles of a topic and it plays a quite important role in Web search; a typical example is the query "jaguar" which may concern both the jaguar animal and the jaguar car brand. Note that there is a difference between diversity and novelty, even if they are often treated together, as outlined by Clarke et al. [88, p. 659]: novelty is the need to avoid redundancy while diversity is the need to resolve ambiguity. Widely used measures for diversity are: α-nDCG [88], which focuses more on novelty; nDCG-IA [6] and D-measure [343], which focus on the different intents behind a query. For a more comprehensive discussion on measures for diversity, please see, Wu et al. [425] and Kunaver and Požrl [241].

Fairness is generally understood as treating users, results, items which are alike in a similar way and it can be studied either for individual users/items/results or for groups of them. Widely used measures for fairness are based on: pair-wise parity [426], ensuring that the proportion of protected candidates matches a target distribution [430], exposure [118, 362], equity of attention [40], dissatisfaction [128], and quantification [127]. For a more comprehensive discussion on measures for fairness, please see, Zehlike et al. [431, 432], Rampisela et al. [315], Amigó et al. [15], and Rai and Ekstrand [314].

4 STATISTICAL SIGNIFICANCE TESTING

Statistical significance testing [159, 248, 295] plays a fundamental role in experimental evaluation since it provides us with the means to properly assess differences among compared systems and to understand when they actually matter.

In the early days of IR, the use of significance testing was not so widespread, mostly because IR experimental data do not fully match the assumptions of such tests, as pointed out by Saracevic [349] and van Rijsbergen [399]. In absence of such testing, Spärck Jones [371] proposed a famous rule-of-thumb about absolute differences among systems: differences less than 5% should be discarded, in the range of 5%−10% are to be considered *noticeable*, and above 10% they are *material*. Several years later, Hull [208] showed that departing from the assumptions behind significance test was not impairing the inferences and conclusion drawn and, thus, it should not have prevented their application to IR experimental data. Tague-Sutcliffe and Blustein [379] report one of the first systematic applications of significance testing to the analysis and comparison of the runs submitted to TREC. Since then the use and study of which significance tests to adopt in IR increased [68, 352, 366], as summarized by Sakai [335] in a brief history of significance in IR.

However, much has still do be done to achieve a widespread and systematic adoption in IR experimental practice. Sanderson and Zobel [348] surveyed 26 papers from the ACM SIGIR conference in 2003-04 and found that 14 papers (61%) did not explicitly state if a significance test was used or failed to name the test; six of the 26 (23%) reported small experimental differences that most likely needed a significance test or another type of statistical method to examine the difference further. Later, Sakai [337] surveyed over 850 papers (drawn from SIGIR and TOIS) and found that around 30% failed to report any form of significance test (or other techniques such as confidence intervals). We can only stress the importance of always applying proper statistical techniques and significance testing at the analysis of your experimental results.

Sakai [339] provides an updated account on how to apply statistical significance testing in IR experimentation. There are many significance tests that can be used, among which:

- **Sign Test** [173] is a non parametric test which looks at the signs of the differences among two paired samples x_i and y_i; the null hypothesis is that the median of the differences is zero. It can be used to compare two systems.

- **Wilcoxon Rank Sum Test** (or Mann-Whitney U Test) [173, 424] is a non parametric test which looks at the ranks of two paired samples x_i and y_i; the null hypothesis is that the two samples have the same median. It can be used to compare two systems.
- **Wilcoxon Signed Rank Test** [173, 424] is a non parametric test which looks at the signs and ranks of the differences among two paired samples x_i and y_i; the null hypothesis is that the median of the differences is zero. It can be used to compare two systems.
- **Randomization Test** [122, 160] is a non parametric test for the null hypothesis that two samples come from the same distribution and requires random resampling of the data to compute the test statistics. It can be used to compare two systems.
- **Student's t Test** [376] is a parametric test for the null hypothesis that two paired samples x_i and y_i come from a normal distribution with same mean and unknown variance. It can be used to compare two systems.
- **ANOVA** [159, 242] is a parametric test for the null hypothesis that q samples come from a normal distribution with same mean and unknown variance. It can be used to compare two or more systems.
- **Kruskal-Wallis Test** [173, 239] is a nonparametric version of the one-way ANOVA for the null hypothesis that q samples come from a distribution with same median. It is based on the ranks of the different samples and it can be considered as an extension of the Wilcoxon rank sum test to the comparison of multiple systems at the same time. It can be used to compare two or more systems.
- **Friedman Test** [164, 165, 173] is a nonparametric version of the two-way ANOVA for the null hypothesis that the effects of the q samples are the same. It is based on the ranks of the different samples. It can be used to compare two or more systems.

The analysis of Sakai [337] showed that, of those papers using a test, the t-test was found to be dominant with the Wilcoxon, sign, bootstrap, and randomisation tests accounting for the vast majority of tests used. However, Sakai found that the surveyed papers provided little justification for test selection. Again, we can only underline the need for motivating the choice of a specific significance test and this should be made before conducting the experiment and not afterwards, picking the significance test which gives us the most favorable outcomes.

In this section, we will present ANOVA, which is a very robust and powerful analysis procedure. Tague-Sutcliffe and Blustein [379] were the first to use ANOVA to analyse TREC data and also Zobel [435] reported its adoption. Recently, there is a revamped interest on how to exploit ANOVA within IR experimentation to better model effectiveness and determine significance. Voorhees et al. [418] and Ferro et al. [152, 155, 156] have jointly exploited sharding of the document corpus together with advanced ANOVA models to improve the accuracy in the estimation of significant differences among systems; Ferro and Sanderson [157] proposed a methodology to conduct a more comprehensive analysis of the behaviour of a significance test and its trade-offs, applying it to the ANOVA models previously described. Faggioli et al. [132] have proposed a new methodology for comparing *Query Performance Prediction (QPP)* algorithms [66] and accounting for query variants based on ANOVA.

4.1 Basic Intuition about Statistical Significance Testing

Before going into the details of ANOVA, we provide a basic intuition about what is the problem addressed by statistical significance tests and how they approach a solution to it.

Consider Figure 12 and assume that we are observing two set of scores with sample mean $\hat{\mu}_X$ and $\hat{\mu}_Y$ respectively. In the IR context, the set of scores are the performance scores of two systems

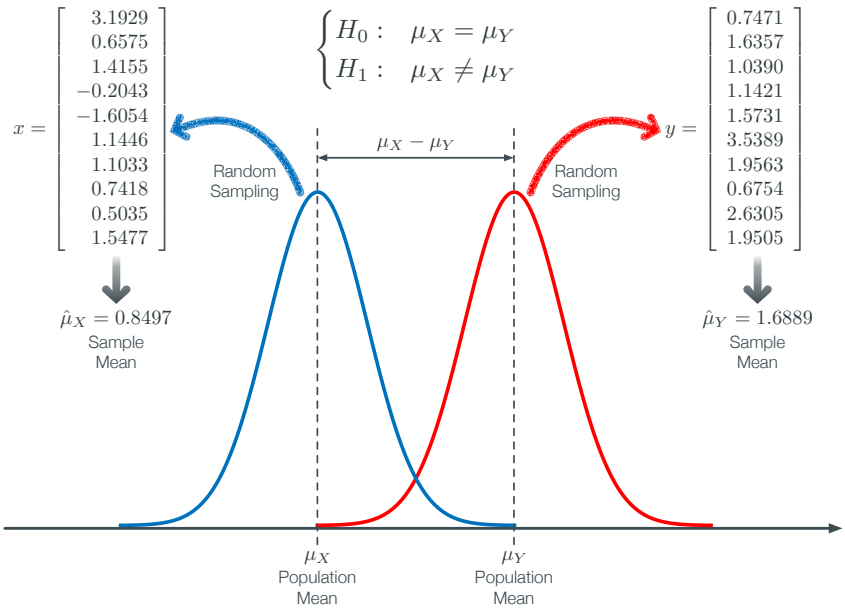


Fig. 12. The problem addressed by statistical significance testing.

X and Y across a set of topics and $\hat{\mu}_X$ and $\hat{\mu}_Y$ are the means of those scores, e.g., the MAP of the two systems.

The problem is to determine whether the observed difference between $\hat{\mu}_X$ and $\hat{\mu}_Y$ is due to the fact that they are two random samples taken from the same distribution or to the fact that they are two random samples actually coming from two different distributions. In the former case, the sample means are different just because of the sampling process but the underlying distribution is the same; in the latter case, they are different because the underlying distributions are different.

We are thus trying to understand from the data whether the population means μ_X and μ_Y are equal, i.e., we are talking about the same distribution – and this is called the *null hypothesis* $H_0 : \mu_X = \mu_Y$ – or whether they are different, i.e., we are talking about two different distributions – and this is called the *alternative hypothesis* $H_1 : \mu_X \neq \mu_Y$. We say that the observed difference is *statistically significant* if the data are unlikely to be a realisation of the null hypothesis with respect to a chosen threshold α, called *significance level*. In this case we *reject* the null hypothesis; in the opposite case, we *fail to reject* the null hypothesis.

In the case of IR, we are comparing two systems X and Y and, for each of them, we have a sample of performance scores computed according to some evaluation measures over a set of topics. The question is whether we observe a difference in the performance of the two systems just because of the (topic) sampling effect, or because the two systems have indeed different performance. If the data let us reject the null hypothesis H_0, that the performance distribution of the two systems is the same, we consider the two systems to be *significantly different*; otherwise, they are not.

Figure 13 illustrates intuitively how statistical significance tests let us decide whether to reject or not the null hypothesis H_0. Each test is built around its own *test statistic*. The distribution of the test statistic is known under the null hypothesis H_0, i.e., when we assume H_0 to be true. The significance level α allows us to determine a *critical value* t_{crit} for the test statistics. When the null hypothesis H_0 is assumed true, the critical value t_{crit} corresponds to a probability α of observing

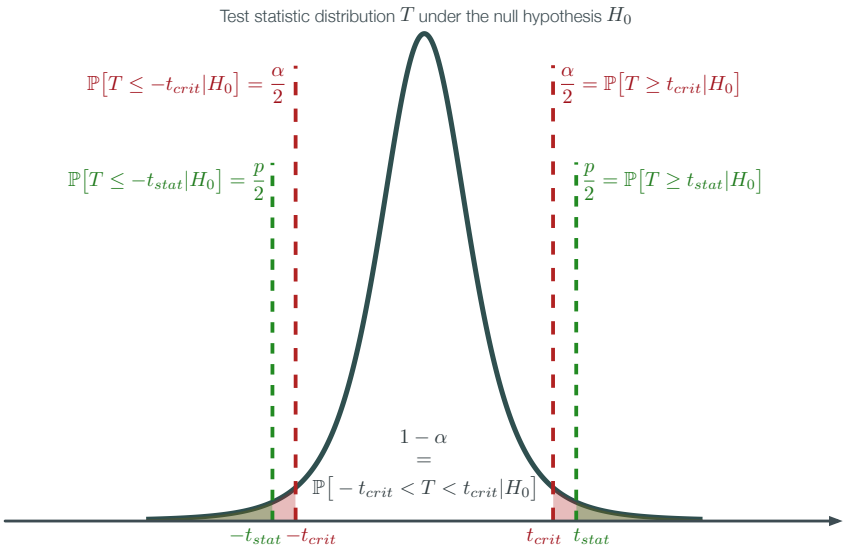


Fig. 13. Rejecting or not the null hypothesis H_0.

values more extreme than t_{crit} itself. We then compute t_{stat}, the value of the test statistic from the actual data. If t_{stat} is more extreme than t_{crit}, we reject the null hypothesis H_0.

Indeed, we consider very unlikely (at least with respect to the chosen significance level α) to observe that value t_{stat} of the test statistic when the null hypothesis H_0 is actually true, i.e., when there are no differences. This means that we "feel confident" in rejecting H_0 because we can attribute that unlikely value of t_{stat} to an actual difference. How much unlikely? This is the so-called p-value, which can be computed from t_{stat} as the probability of observing a value more extreme than t_{stat}. Summing up, we reject the null hypothesis H_0 when t_{stat} is more extreme than t_{crit} or, equivalently, when $p \leq \alpha$.

Figure 13 shows the case of a *two-tailed test*, i.e., when "more extreme" means that it can be either "greater than" or "less than" and we attribute half of the probability α to each of these cases. This is the common case in IR when you do not know a-priori which system is better than the other. The other option is a *one-tailed test*, when you have a-priori motivations to know that a system can only be better (if it happens) than another one.

Figure 14 shows the confusion matrix between the truth/falseness of the null hypothesis (rows) and outcomes of a significance test (columns). We commit a *Type I error* when we reject the null hypothesis H_0 while indeed it is true, i.e., when we consider two systems as significantly different while they are not. The probability of a Type I error is exactly the significance level α. Typical values are $\alpha = 0.05$ or $\alpha = 0.01$; for example, setting $\alpha = 0.05$ means that we accept a 5% chance of committing a Type I error. What does this 5% chance mean? If we take two random samples from the same distribution and we perform a significance test over them, we expect to not reject the null hypothesis since we know that, by construction, these two samples are drawn from the same distribution. However, if we repeat this procedure 100 times, (on average) in 5 cases we will reject the null hypothesis anyway, i.e., we will commit a Type I error. Type I errors are false positives and, in experimentation, you wish to keep their rate controlled since they may have severe consequences; for example, due to a Type I error you may mistakenly conclude that a drug is effective for treating a disease when actually it is not.

	We fail to reject H_0 [not statistically significant]	We reject H_0 [statistically significant]
H_0 is true [e.g. systems are equivalent]	Correct conclusion [true negative] Probability $1 - \alpha$	**Type I error** [false positive] Probability α
H_0 is false [e.g. systems are not equivalent]	**Type II Error** [false negative] Probability β	Correct conclusion [true positive] Probability $1 - \beta$

Fig. 14. Confusion matrix between the truth/falseness of the null hypothesis (rows) and outcomes of a significance test (columns).

On the other hand, when the null hypothesis H_0 is false but we fail to reject it, i.e., when two systems actually are different but we deem them to be not significantly different, we commit a *Type II error*. Type II errors are false negatives and they are sort of "missed opportunities": we could correctly consider two systems different but we wrongly miss this opportunity. The probability of a Type II error is β and the probability $1 - \beta$ of not committing Type II errors is called the *power* of a statistical significance test. You wish to keep the Type II error rate controlled as well but often in a less strict way, e.g., typical values are $\beta = 0.2$.

The discussion so far concerned the comparison of two IR systems but what happens if we need to compare many of them? In evaluation campaigns and everyday development, it is common to compare different systems or versions of the same system. Performing multiple comparisons increases the Type I error probability, i.e., it is easier to reject the null hypothesis when you should not, as shown in the following:

$$\mathbb{P}\big[\text{No Type I Error}\big] = (1 - \alpha)$$

$$\mathbb{P}\big[\text{No Type I Error}\mathbf{s}\big] = \prod_{i=1}^{c}(1 - \alpha) = (1 - \alpha)^c$$

$$\mathbb{P}\big[\text{At Least One Type I Error}\big] = 1 - (1 - \alpha)^c \longrightarrow 1$$

where c is the number of independent comparisons. As c increases, the probability that at least one Type I error occurs tends to 1. As underlined by both Fuhr [168] and Sakai [341], when performing multiple comparisons, you should always apply a proper correction procedure to ensure to control the Type I error rate.

4.2 ANalysis Of VAriance

A *General Linear Mixed Model (GLMM)* [272, 331] explains the variation of a dependent variable ("Data") in terms of a controlled variation of independent variables ("Model") in addition to a residual uncontrolled variation ("Error"): $Data = Model + Error$.

The most basic example of GLMM is a simple linear regression, where $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. The dependent variable Y_i represents the score of the i-th subject. Y_i is explained (predicted) as a sum of 3 terms: (1) an intercept β_0; (2) an independent variable X_i (predictor) times the regression coefficient β_1, i.e., the slope of the regression line; (3) a residual error ε_i, not explained by the model, which follows a Gaussian distribution with mean 0.

In GLMM terms, *ANalysis Of VAriance (ANOVA)* attempts to explain data (the dependent variable scores) in terms of the experimental conditions (the model) and an error component. Typically, ANOVA is used to determine under which experimental condition do dependent variable score means differ and what proportion of variation in the dependent variable can be attributed to differences between specific experimental groups or conditions, as defined by the independent variable(s). An ANOVA can be regarded as a particular type of regression analysis that employs only categorical predictors.

The GLMM regression model is expressed in ANOVA terms as $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, where i represents the subject and j the experimental condition. Y_{ij} is the score of the dependent variable computed for subject i with the experimental condition j. The parameter μ is the grand mean of the experimental condition population means that underlies all subjects' dependent variable scores. The parameter α_j is the effect of the j-th experimental condition. The random variable ε_{ij} is the error term, which reflects variation due to any uncontrolled source. The above regression model corresponds to the ANOVA version once you add as many X_{ij} predictors and as many levels as there are in the experimental condition α_j, e.g., by using dummy coding.

For a given model, the ANOVA table summarizes the outcomes of the ANOVA test indicating, for each factor, the *Sum of Squares (SS)*, the *Degrees of Freedom (DF)*, the *Mean Squares (MS)*, the F statistics, and the p-value of that factor, which allows us to determine the significance of that factor.

When it comes to independent variables they can be either *fixed effects* – i.e., they have precisely defined levels, and inferences about its effect apply only to those levels – or *random effects* – i.e., they describe a randomly and independently drawn set of levels that represent variation in a clearly defined wider population. The latter case is more sophisticated because when it estimates the variance attributed to the different factors, it accounts also for the additional randomness due to sampling of effect levels.

The experimental design determines how you compute the model and how you estimate its parameters μ and α_j. In particular, it is possible to have an *independent measures* design, where different subjects participate to different experimental conditions (factors), or a *repeated measures* design, where each subject participates to all the experimental conditions (factors).

A final distinction is between *crossed/factorial* designs, where every level of one factor is measured in combination with every level of the other factors, and *nested* designs, where levels of a factor are grouped within each level of another nesting factor.

4.2.1 Estimating the Model. Figure 15 shows the experimental layout for the ANOVA model reported in equation (7), which is also called a *two-way* ANOVA since it is constituted by two factors. This is the typical IR setting where you have a set of topics and a set of systems which are run against those topics; in ANOVA terms this is a crossed/factorial repeated measures design. Note that this is the same model used by Banks et al. [33] and Tague-Sutcliffe and Blustein [379] to analyse TREC data.

For m topics and n systems, the two-way ANOVA is defined as follows:

$$Y_{ij} = \mu_{..} + \tau_i + \alpha_j + \varepsilon_{ij} \tag{7}$$

where: $\mu_{..}$ is the grand mean; τ_i with $i = 1, \ldots, m$ represents the effect of topics; α_j with $j = 1, \ldots, n$ represents the effect of systems; and, ε_{ij} is the residual error.

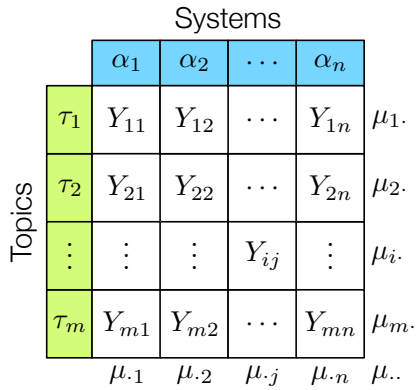


Fig. 15. Two-way ANOVA model for the topic and system effects.

The model of equation (7) has the following estimators:

- grand mean

$$\hat{\mu}_{\cdot\cdot} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} Y_{ij}$$

- topic marginal mean and topic effect

$$\hat{\mu}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$$
$$\hat{\tau}_i = \hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot}$$

- system marginal mean and system effect

$$\hat{\mu}_{\cdot j} = \frac{1}{m} \sum_{i=1}^{m} Y_{ij}$$
$$\hat{\alpha}_j = \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$$

Therefore, the score predicted by the model is

$$\hat{Y}_{ij} = \hat{\mu}_{\cdot\cdot} + \hat{\tau}_i + \hat{\alpha}_j = \hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot}$$

and the prediction error is:

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})$$

4.2.2 Assessment of the Model. We can compute the *Sum of Squares (SS)*, *Degrees of Freedom (DF)*, *Mean Squares (MS)* and F statistics as follows:

- total effects

$$\mathrm{SS}_{\mathrm{Total}} = \sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \hat{\mu}_{\cdot\cdot})^2$$
$$\mathrm{df}_{\mathrm{Total}} = mn - 1$$
$$\mathrm{MS}_{\mathrm{Total}} = \frac{\mathrm{SS}_{\mathrm{Total}}}{\mathrm{df}_{\mathrm{Total}}}$$

- topic effects

$$\begin{aligned}
\mathrm{SS}_{\mathrm{Topic}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\tau}_i^2 = n\sum_{i=1}^{m}\hat{\tau}_i^2 = n\sum_{i=1}^{m}(\hat{\mu}_{i\cdot} - \hat{\mu}_{\cdot\cdot})^2 \\
\mathrm{df}_{\mathrm{Topic}} &= m-1 \\
\mathrm{MS}_{\mathrm{Topic}} &= \frac{\mathrm{SS}_{\mathrm{Topic}}}{\mathrm{df}_{\mathrm{Topic}}} \\
\mathrm{F}_{\mathrm{Topic}} &= \frac{\mathrm{MS}_{\mathrm{Topic}}}{\mathrm{MS}_{\mathrm{Error}}}
\end{aligned}$$

- system effects

$$\begin{aligned}
\mathrm{SS}_{\mathrm{System}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\alpha}_j^2 = m\sum_{j=1}^{n}\hat{\alpha}_j^2 = m\sum_{j=1}^{n}(\hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})^2 \\
\mathrm{df}_{\mathrm{System}} &= n-1 \\
\mathrm{MS}_{\mathrm{System}} &= \frac{\mathrm{SS}_{\mathrm{System}}}{\mathrm{df}_{\mathrm{System}}} \\
\mathrm{F}_{\mathrm{System}} &= \frac{\mathrm{MS}_{\mathrm{System}}}{\mathrm{MS}_{\mathrm{Error}}}
\end{aligned}$$

- error effects

$$\begin{aligned}
\mathrm{SS}_{\mathrm{Error}} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\varepsilon}_{ij}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{ij} - (\hat{\mu}_{i\cdot} + \hat{\mu}_{\cdot j} - \hat{\mu}_{\cdot\cdot})\right)^2 \\
\mathrm{df}_{\mathrm{Error}} &= (m-1)(n-1) \\
\mathrm{MS}_{\mathrm{Error}} &= \frac{\mathrm{SS}_{\mathrm{Error}}}{\mathrm{df}_{\mathrm{Error}}}
\end{aligned}$$

Note that:

$$\mathrm{SS}_{\mathrm{Total}} = \mathrm{SS}_{\mathrm{Topic}} + \mathrm{SS}_{\mathrm{System}} + \mathrm{SS}_{\mathrm{Error}}$$

We can then compute the critical value for the F statistics of a factor, i.e., $\mathrm{F}_{crit} = \mathrm{F}_{\left(\mathrm{df}_{fact},\mathrm{df}_{err}\right)}$, and determine its significance if $\mathrm{F}_{fact} > \mathrm{F}_{crit}$; this allows us also to obtain the p-value for that factor.

4.2.3 Effect Size. As the sample size increases, the F statistics tend to increase and the p-value tends to decrease. As a consequence, we also consider the *effect size* of a factor, which accounts for the amount of variance explained by the model. To do this, we can use the following unbiased estimator [297, 334]:

$$\hat{\omega}^2_{\langle fact\rangle} = \frac{\mathrm{df}_{fact}(\mathrm{F}_{fact} - 1)}{\mathrm{df}_{fact}(\mathrm{F}_{fact} - 1) + mn} \tag{8}$$

where F_{fact} is the F-statistic; df_{fact} are the degrees of freedom for the factor; and mn is the total number of samples, i.e., total number of topics times total number of systems. In this way, we are able to assess not only if a factor is significant but also how much it matters.

The common rule of thumb [331] when classifying $\hat{\omega}^2_{\langle fact\rangle}$ effect size is: 0.14 and above is a *large size effect*, 0.06–0.14 is a *medium size effect*, and 0.01–0.06 is a *small size effect*. Note, $\hat{\omega}^2_{\langle fact\rangle}$ can be negative, in such cases it is considered as zero.

4.2.4 Multiple Comparisons. As discussed above, if one simultaneously compares multiple system pairs, the probability of committing a *Type I* error increases. This probability is called the *Family-wise Error Rate (FWER)* and is computed as $\text{FWER} = 1 - (1 - \alpha)^c$, where c is the total number of comparisons to be performed [199, pp. 7–8].

Tukey [394] proposed the *Honestly Significant Difference (HSD)* test, which creates confidence intervals for all pairwise differences between factor levels, while controlling the FWER. Two systems u and v are considered significantly different when:

$$|tk| = \frac{|\hat{\mu}_{\cdot u} - \hat{\mu}_{\cdot v}|}{\sqrt{\frac{\text{MS}_{\text{Error}}}{m}}} > Q^{\alpha}_{n,\text{df}_{\text{Error}}} \tag{9}$$

where: $\hat{\mu}_{\cdot u}$ and $\hat{\mu}_{\cdot v}$ are the marginal means of the systems u and v as estimated from the actual data; df_{Error} are the DF of the error; MS_{Error} is the MS of the error, i.e., an estimation of the variance left unexplained; and $Q^{\alpha}_{n,\text{df}_{\text{Error}}}$ is the upper $100 * (1 - \alpha)$-th percentile of the studentized range distribution [294]; m is the total number of topics and n is the total number of systems.

The test statistic $|tk|$ allows us to compute the p-value

$$p = \mathbb{P}\left[Q^{\alpha}_{n,\text{df}_{\text{Error}}} \geq |tk|\right] \tag{10}$$

of observing a more extreme value of the Studentized range distribution. We can then compare this p-value to the desired significance level α and, if it is $\leq \alpha$, the two systems u and v are significantly different, still controlling the FWER. Eqs. (9) and (10) are two equivalent ways to perform multiple comparisons controlling the FWER.

The Tukey HSD test of eq. (9) allows us to define exact confidence intervals for the system main effects, still controlling the FWER. Hochberg and Tamhane [199] suggest creating a half-width confidence interval around the marginal mean of a system u:

$$\hat{\mu}_{\cdot u} \pm \frac{1}{2} Q^{\alpha}_{n,\text{df}_{\text{Error}}} \sqrt{\frac{\text{MS}_{\text{Error}}}{m}} \tag{11}$$

Systems u and v are significantly different, according to the Tukey HSD test of eq. (9), if and only if their confidence intervals of eq. (11) do not overlap [199, p. 116].

Voorhees et al. [418] adopted a different approach to the multiple comparison problem and opted for the *False Discovery Rate (FDR)* control technique proposed by Benjamini and Hochberg [37]. While FWER controls the probability of making even one Type I error, FDR controls the proportion of Type I errors. The main benefit of FDR with respect to FWER is that it is more powerful, i.e., it allows for detecting more significant differences, than FWER, especially when the number of comparison increases, as it happens in the IR case. However, the main drawback is that this additional power comes at the price of an increased number of false positives.

The approach to be adopted for multiple comparisons is a quite debated issue in statistics but the answer on what approach to use is mostly determined by the expected use of the experimental findings and by the distinction between exploratory analysis, where all the possible cases are examined in the search for interesting patterns, as opposed to the selection of just a few cases to consider, motivated by some exogenous cause and knowledge, as highlighted by Tukey [395]. In this respect, Hochberg and Tamhane [199, p. 11] suggest to adopt FWER when exploratory analysis is conducted and high validity is required: "As a rule of thumb, one may adhere to controlling the FWER whenever the findings are presented without highlighting the selection process and/or without emphasizing the need for further confirmation" – and, this is a typical scenario in IR.

Moreover, when it comes to all-pairwise comparisons, there is agreement in recommending – see for example Hsu [207] and Maxwell and Delaney [272] – the Tukey HSD test as more appropriate

and powerful than other methods such as the Bonferroni [46] and Scheffe [354] methods. Finally, Benjamini and Hochberg [37, p. 291] observes that "any procedure that controls the FWER also controls the FDR" and, therefore, adopting a FWER approach we draw conclusions which hold also in the case of the FDR approach used by Voorhees et al. [418]. In this respect, Faggioli and Ferro [131] compared traditional ANOVA approaches to bootstrap ANOVA ones by Voorhees et al. as well as adopting FWER or FDR. Faggioli and Ferro concluded that traditional ANOVA is more stable and less computationally expensive than bootstrap ANOVA. Moreover, as expected, traditional ANOVA using FWER is stricter, i.e., it identifies less significant different pairs, than bootstrap ANOVA using FDR; however, traditional ANOVA using FDR provides similar results to bootstrap ANOVA using FDR, still being a little bit more stable. Finally, Ferro and Sanderson [157] have recently investigated the stability of several ANOVA models under different conditions, finding them to be strikingly consistent in the settings of a large-scale evaluation campaign. However, in the narrower case of some specific participant experiments, Ferro and Sanderson found that ANOVA models may behave in a quite less consistent way.

4.2.5 Assumptions. ANOVA is based on the following assumptions [242]:

- normality of the error terms;
- equal variance (homoskedasticity) of the error terms;
- independence of the error terms, i.e., they are a random sample.

ANOVA is known to be quite robust to violations of the first two assumptions. Ito [211, p. 205] observes that "the F-test is found to be remarkably insensitive to general nonnormality. In the commonly occurring case where the group sample sizes are equal, it is not very sensitive to heterogeneity of variance from group to group". Similarly, Mendenhall and Sincich [276] note that, for relatively large samples (e.g., 20 or more observations per factor), ANOVA is robust to violations of the normality assumption and that it is also robust to unequal variances in the case of balanced design. On the other hand, violation of the third assumption may severely impact the F-test and hamper the drawn conclusions as noted, for example, by Scariano and Davenport [353].

IR performance scores are known to violate the first two ANOVA assumptions [68, 379]. Tague-Sutcliffe and Blustein [379] noted that performance scores did not satisfy the homoskedasticity assumption and applied a transformation, which is typically used in the case of ratio data, consisting of taking the arcsine of the square root of the original scores. However, they noted very few differences in the analysis conducted on the transformed data and they decided to stay with the untransformed scores, which are more easily interpretable. Carterette [68] observed that both the first two assumptions are violated because of performance scores typically being bounded in $[0, 1]$; however Carterette concluded that ANOVA is robust to the kind of violations of normality due to IR performance scores and that also the violations of homoskedasticity have a fairly limited impact, in agreement with previous findings of Tague-Sutcliffe and Blustein [379].

4.2.6 Example. We now report an example of analysis of the results by using ANOVA; this analysis resembles what done by Banks et al. [33] and by Tague-Sutcliffe and Blustein [379].

We use data from the TREC 08 Adhoc track [417] which consists of: 528, 155 documents of the TIPSTER disks 4-5 corpus minus congressional record; 50 topics, each with binary relevance judgments and a pool depth of 100; 129 system runs were submitted. We use *Average Precision (AP)* as evaluation measure and we set the significance level at $\alpha = 0.05$

Figure 16 shows the *box plot* [275] of the AP scores of the runs submitted to TREC 08; systems, on the x-axis, are sorted in descending order of their MAP. A boxplot is a graphical tool to summarise a distribution of data: the box shows the first quartile (Q1), the second quartile (Q2, the median) as a line inside the box, and the third quartile (Q3); the box itself represents the Inter-Quartile Range

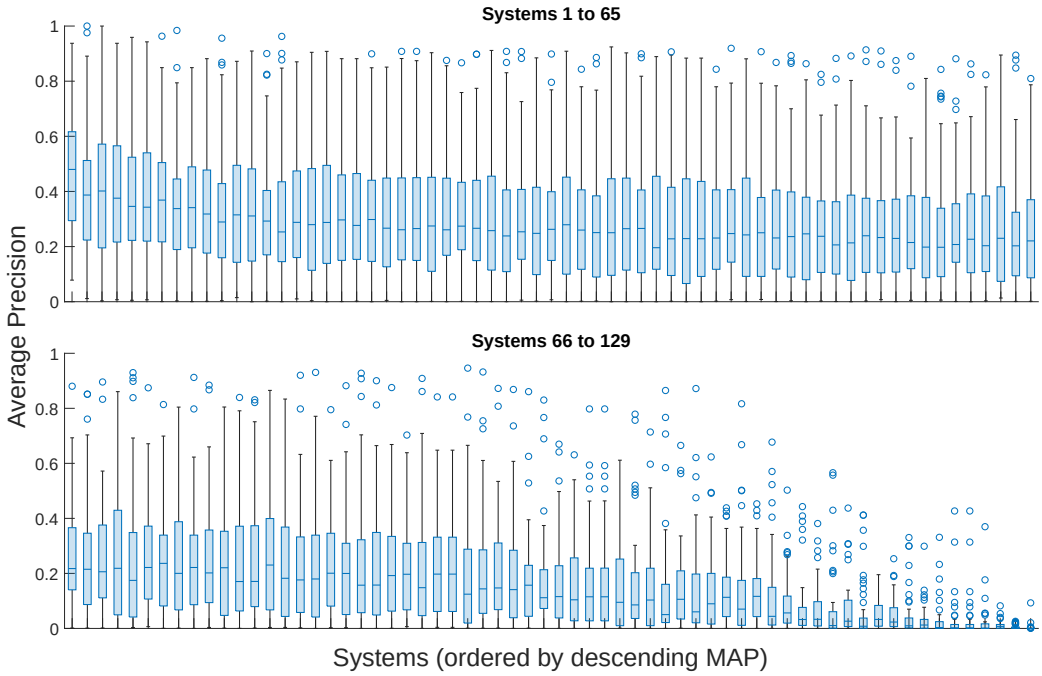


Fig. 16. Boxplot of the AP scores of the runs submitted to TREC 08. Systems on the x-axis are sorted in descending order of their MAP.

Table 1. ANOVA model of Eq. (7) applied to TREC 08 data.

Source	SS	DF	MS	F	p-value	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	167.9974	49	3.4285	251.9463	0	0.6559
System	60.0299	128	0.4690	34.4635	0	0.3991
Error	85.3502	6272	0.0136			
Total	313.3375	6449				

(IQR), i.e., the difference Q3-Q1; the extension of the whiskers (the lines outside the box) represents $1.5 \cdot \text{IQR}$ and they roughly cover 99% of the data, assuming a normal distribution; any data outside the whiskers is considered an outlier, represented with a $\circ$ symbol.

We can observe as there is quite a sizeable difference between top performing systems (top plot) and worst performing ones (bottom plot) but there is also a quite wide range of performance for each system, as represented by the extension of the boxes and of the whiskers. Therefore, it would be difficult to tell ahead which systems are significantly better than others without using a significance test.

Table 1 shows an example of a typical ANOVA table used for summarizing the results. For each factor, the table reports all the main indicators discussed in Section 4.2: the *Sum of Squares (SS)*, the *Degrees of Freedom (DF)*, the *Mean Squares (MS)*, the F statistic, the p-value, and the ω^2 *Strength of Association (SOA)*. Both the Topic and the System factors are statistically significant (the estimated p-value is 0 for both of them), i.e., there is at least a pair of topics and a pair of systems which are significantly different. We can also observe that the SS of the Topic, System, and Error factors

Table 2. Some offline measures and their parameters, which can be calibrated with online data.

Measure	g_i	$W(i)$	User model
RBP	$\hat{r}_t[i]$	p^{i-1}	positional (top → bottom)
ERR	$\prod_{k=1}^{i-1}(1-R_k)R_i$	$1/i$	cascade (top → bottom)
TBG	g_i	D(T(i))	positional with time (top → bottom)
click-based utility	$\mathbb{P}[C_i = 1]$	-	defined by the click model
click-based effort	$s_i\mathbb{P}[C_i = 1]$	$1/i$	defined by the click model
MP	$P(i)$	π_i	Markovian with/without time

do sum up to the Total SS. Finally, both the Topic and the System factors are large-size effects (estimated ω^2 above 0.14 for both of them) and the Topic factor is about 1.6 times bigger than the System one. This is a well-known phenomenon in the literature: the variability across topics is much bigger than the variability across systems.

There are $n = 129$ submitted runs and therefore there are $\frac{n(n-1)}{2} = 8,256$ system pairs to be compared; out of them, 3,423 pairs turn out to be significantly different according to the Tukey HSD test with a half-width confidence interval equal to 0.0515[22].

5 OFFLINE EVALUATION WITH ONLINE DATA

As described in Section 2, offline evaluation allows us to compare IR systems in a controlled and reproducible experimental setting. However, one might wonder: given two systems A and B, if system A is better than system B for a given offline measures M, do users prefer system A over B? Sanderson et al. [347] investigate whether offline evaluation with IR measures correlates with user preferences. They show that nDCG and ERR (presented in Section 5.1) correlate more with user preferences, while P@10 poorly represents users' preferences. Carterette [67] investigates the same problem from a different perspective: he proposes a conceptual framework to analyze discount functions of IR measures and shows that fat-tail discount functions, as the one used by nDCG, lead to more robust measures, closely aligned with user behaviour.

To align IR evaluation measures and user behaviour even more, several measures have been proposed, which can be used for offline evaluation, but with online data to calibrate their parameters. The typical form of these measures is as follows:

$$M(r_t) = \sum_{i=1}^{N} g_i \cdot W(i) \tag{12}$$

where g_i is the gain contributed by the document at rank position i (or utility model in Section 3) and $W(i)$ is the discount (or weight), associated to the rank position i. Next we will present some measures, where g_i and $W(\cdot)$ are defined upon different user models and can be calibrated with online data. Table 2 summarizes their user models and the value of their parameters.

5.1 Measures Calibrated with Online Data

Rank-Biased Precision (RBP) [288] considers a user who linearly scans the ranked list of results. The user will start from the first rank position and proceeds towards the bottom of the ranking with *persistence* or probability p: at each rank position the user will decide to visit the next document

[22]The 0.05 half-size width of the confidence interval means that a MAP distance of 0.10 between two systems is required to consider them as significantly different. It is somewhat surprising how much these figures recall the 5% − 10% absolute difference rule-of-thumb by Spärck Jones [371], even if this obviously is just a coincidence.

with probability p, or to finish the examination of the ranking with probability $1 - p$. The choice between visiting the next document or terminating the examination is taken independently of the relevance of the document and of the depth reached in the ranking. RBP is computed as follows:

$$\text{RBP}(r_t) = (1 - p) \sum_{i=1}^{N} \hat{r}_t[i] \cdot p^{i-1} \tag{13}$$

where $(1 - p)$ is a normalization factor[23] ensuring that RBP scores are in $[0, 1]$, $\hat{r}_t[i]$ is the binary relevance of the document at rank position i, and p^{i-1} is the probability that the user visits the document at rank position i. The persistence p is a value in $[0, 1]$ which represents how much patient the user is, similarly to the log base b for nDCG (see Section 3.4). Commonly used values are $p = 0.5$ for highly impatient users, $p = 0.95$ for more persistent users, and $p = 0.8$ for users in between. When real user data are available, p can be calibrated directly from the data, e.g., with clicks, mouse movements, etc.

Expected Reciprocal Rank (ERR) [70] is an extension of *Reciprocal Rank (RR)*[24] [364] based on the cascade model, opposed to the position model used by RBP. The cascade model assumes that users scan the ranked list of documents from top to bottom. However, differently from the position model, the choice of visiting the next document is based not only on the rank position, but also on the relevance of previously visited documents. ERR is computed as follows:

$$\text{ERR}(r_t) = \sum_{i=1}^{N} \frac{1}{i} \prod_{k=1}^{i-1} (1 - R_k) \, R_i \tag{14}$$

where R_i is the probability of the user being satisfied with the document at position i. The product on the right-hand side of Equation (14) is the probability that a user terminates the examination with the document at position i, i.e., the product between the probability that the user was not satisfied by the previous $i - 1$ documents and the probability that the user is satisfied by the current document at position i. For offline evaluation, the probabilities R_i can be computed from the relevance grade of each document, while for online evaluation R_i can be estimated with maximum likelihood from click logs data.

Time-Biased Gain (TBG) [367, 368] assumes a user that scans the ranking from top to bottom and in addition accounts for the temporal dimension. Indeed, a limitation of RBP and ERR is that they implicitly assume a user who examines documents at constant rate. This is a simplistic assumption, especially when we need to evaluate modern search engines, which display short snippets of Web pages that users can quickly skip if not relevant. TBG is computed as follows:

$$\text{TBG}(r_t) = \sum_{i=1}^{N} g_i D(T(i)) \tag{15}$$

where g_i is the gain associated with the document at rank i, $T(i)$ is the expected time a user would take to reach the rank position i, and $D(\cdot)$ is a decay function, which models the probability that a user continues to examine the ranking until a given time. All the parameters and functions in Equation (15) need to be estimated from click log data. Alternatively, Smucker and Clarke [368] propose to calibrate $T(i)$ as the cumulative sum of the time needed to read the snippet T_S, and the time needed to read the document T_D[25], conditioned to the probability of the user clicking on that document. The gain g_i is defined as a function of the relevance label and the decay $D(\cdot)$ is a negative exponential function.

[23]The geometric series converges to $1/(1 - p)$.

[24]RR is defined as the inverse of the rank position of the first relevant retrieved document.

[25]T_D depends on the length of the document.

Click model based measures [81] exploit the output of a click model as a parameter to estimate the user utility or effort. Click models [80] are probabilistic models of the user behaviour, which learn how to predict users clicks on a *Search Engine Result Page (SERP)*. Following the framework proposed in [67] there are two families of click based measures: *utility based* and *effort based* measures. Utility based measures account just for the utility contributed by each document, estimated from the relevance grade and the click probability. Utility based measures are computed as follows:

$$uM(r_t) = \sum_{i=1}^{N} \mathbb{P}[C_i = 1]\hat{r}_t[i] \tag{16}$$

where $\hat{r}_t[i]$ is the relevance grade of the document at rank position i and $\mathbb{P}[C_i = 1]$ is the probability that the user clicks on that document. The probability $\mathbb{P}[C_i = 1]$ is returned as output of a click model trained on online data, thus the underlying user behaviour depends on the click model being used. Effort based measures account for the user effort and the probability of a user stopping the examination at a given rank position. Effort based measures are computed as follows:

$$rrM(r_t) = \sum_{i=1}^{N} s_i \mathbb{P}[C_i = 1]\frac{1}{i} \tag{17}$$

where $s_i = \mathbb{P}[S_i = 1|C_i = 1]$ is the probability that the user is satisfied by the document at rank position i, given that he/she clicked on that document, and the reciprocal of the rank position $1/i$ discounts for the user effort. Note that to compute effort based measures the underlying click models need to be able to estimate user satisfaction (e.g., *Dynamic Bayesian Network (DBN)* [71], *Dependent Click Model (DCM)* [181][26]).

Markov Precision (MP) [140] accounts for the temporal dimension and does not assume a user that examines the ranking in a linear way. As shown by Thomas et al. [385], users might have a complex behaviour, which can not be described with sequential models. Thus, MP uses a Markov chain to model the user behaviour, which allows the user to skip documents, move backwards, and re-visit the same document multiple times. The state space of the Markov chain is the set or a subset of rank positions and the transition probabilities represent the probability that a user moves from one document to any other in the ranking. MP is computed as follows:

$$\mathrm{MP}(r_t) = \sum_{i \in \mathcal{R}} \pi_i \cdot P(i) \tag{18}$$

where $\mathcal{R}$ is the set of rank positions of relevant retrieved documents, π_i is the invariant distribution of the Markov chain restricted to relevant documents, and $P(i)$ is precision computed at cut-off i. The transition probabilities can be either predefined, for example by assuming a user that moves only among adjacent documents, or can be calibrated directly with click log data or eye-tracking. Moreover, MP and its user model can be extended to evaluate IR systems at session level [397].

6 ONLINE EVALUATION

This section goes beyond Cranfield framework and evaluation with test collections [346] presented in Section 2. Offline evaluation has several limitations: collecting relevance assessments is expensive and time consuming, which in turns pose a limit on the number of topics that can be included in test collections (usually only 50), even though real world search engines receive and process billions of queries per day[27].

[26]The effort based measure in Equation (17) computed with DCM click model corresponds to ERR
[27]https://www.internetlivestats.com/google-search-statistics/

Furthermore, IR systems are becoming increasingly complex, personalized and context dependent to attempt in correctly interpreting users' information need, which is extremely subjective. Therefore, topical relevance is not sufficient and objective judgements provided by assessors can not be used to adequately evaluate complex and highly personalized retrieval tasks [200].

Online evaluation exploits the behaviour and interactions of real users, while they engage with an IR system. This allows to evaluate IR systems from the perspective of real users, aiming at optimizing their utility.

6.1 Description of Online Evaluation

In the evaluation spectrum proposed by Kelly [229], online evaluation is placed close to the centre of the spectrum, shifting the focus from systems to humans[28]. Specifically, Hofmann et al. [200] define *online evaluation* as "the evaluation of a fully functioning system based on implicit measurement of real users' experience of the system in a natural usage environment".

The *natural usage environment* is represented by a user issuing a query to an IR system. The IR system returns a ranked list of document snippets, usually grouped on different pages with 10/20 snippets per page. Each page is called *Search Engine Result Page (SERP)*. The user interacts with the SERP by clicking on the snippets to open and read the documents. The user's session starts when the user issues the first query, but it is complicated to determine when the session ends [212, 213]. For example, within the same search episode, a user can issue multiple queries related to the same information need, every time refining the query terms based on the results in the SERP. This case should be treated as a single session and the query terms can be analyzed to understand when there is a shift in the information need. Moreover, many times users do not close the connection with the IR system even if their session is over, therefore it is convenient to set a time-out after a long period of inactivity (usually 30 minutes [183]).

Implicit measurements include any type of interaction, obtained by watching users while they naturally engage with an IR system [230, 359]. Examples of implicit feedback are clicks on search results, mouse hover, dwell time or reading time, purchase decisions, etc. Note that collecting implicit feedback is transparent to users and does not require any extra effort, opposed to explicit feedback, which requires users to actively engage in additional tasks.

Online evaluation requires access to a large amount of log data from real users. For example query or click logs are a valuable and informative source of user feedback: they can be collected by IR systems at almost no cost, they do not require any effort from users, they are available in large quantities and in real time, and they represent personalized preferences of users [216]. Nevertheless, such data might not be easily accessible, due to privacy or ethical constraints and it is inherently biased and noisy. Indeed, a document can be frequently clicked simply because it is displayed at the top of the ranked list of documents, rather than being relevant (position bias [108, 216]). In a different situation, a user may find the answer in a document snippet, so even if the document is relevant, the user does not need to click on it (good abandonment [253]). As a consequence, it is hard to interpret the actual user behaviour and to remove bias and noise from log data.

Furthermore, compared to offline evaluation, online evaluation is hardly reproducible. Indeed, the evaluation outcome depends on the users and the interpretation of their implicit signals: thus testing two systems in two different time periods might lead to different conclusions, especially if the user population is not large and diverse enough. On the other side, with offline evaluation the collection of documents, the set of topics and the relevance assessments are fixed, thus the evaluation scores of baseline systems can be used as benchmark.

[28] In [229] online evaluation corresponds to log analysis.

To conclude, we summarize advantages and disadvantages of different evaluation strategies analyzed so far:

- Offline Evaluation: allows for direct comparison of multiple IR systems, it is simple and can be easily replicated/reproduced, but does not account for the actual user experience and satisfaction.
- Offline Evaluation Calibrated with Online Data: in-between offline and online evaluation, it is better aligned with the user behaviour than offline evaluation and it is easier to reproduce than online evaluation.
- Online Evaluation: accounts for the actual user behaviour while interacting with the IR system, but it is complex and in most of the cases it is not reproducible.

6.2 Online Controlled Experiments

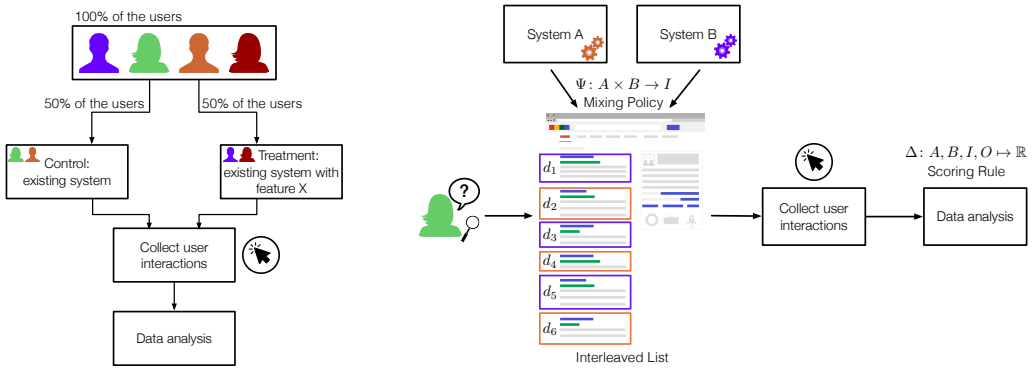
The goal of a controlled experiment is to explain the relation between the cause (independent variable) and its effect (dependent variable). Researchers formulate an hypothesis and then test it by controlling the independent variable and quantitatively measure its effect on the dependent variable. The *Overall Evaluation Criterion (OEC)* is a set of observations or indicators which are collected to quantitatively measure the effect of changes in the independent variable.

In online evaluation, controlled experiments are run to assess the quality of different IR systems. For example, assume that a system B is developed with a new feature and this system is supposed to improve over the current system A. An online controlled experiment exploits interactions of users to answer the question: is system A better than system B? Thus a large number of users is exposed to systems A and B and their interactions are collected and analysed. In this setting, examples of OEC can be the click-through rate, the conversion rate[29], and the overall time spent in engaging with the system (see Section 6.5). The tested hypothesis assumes that the effect of the new feature in system B can be measured and quantified, for example with an increased click-through rate or conversion rate.

When running controlled experiments, we need to consider *confounding variables*: these are external factors, that the researcher can not control, which can affect both the dependent and independent variables. Confounding variables can lead to wrong conclusions: for instance a researcher can find a cause-and-effect relationship which does not exist, because it is due to the confounding variables, or he/she can fail to detect a real cause-and-effect relationship, because the confounding variable masks the effects on the dependent variable. Going back to the two systems A and B, an example of confounding variable can be the presence of an extremely popular item, thus a large increment in click-through rate or purchase rate might be caused by this popular item, rather than the new developed feature of system B. Randomization is the most robust way to limit the impact of confounding variables: if the number of users involved in the experiment is large enough, the value of confounding variables is somehow averaged across the population, and they will not affect the cause-and-effect relationship. Finally, statistical tests should be run to compare the observations from the control and treatment groups and ensure that the differences are not due to chance (see Section 4).

There are two main types of controlled randomized experiments: *between-subject* and *within-subject* experiments. In between-subject experiments users are exposed just to a single condition, while in within-subject experiments users are exposed simultaneously to both conditions. In IR online evaluation, A/B testing is commonly used for between-subject controlled experiments (see Section 6.3) and interleaving is used for within-subject experiments (see Section 6.4).

[29]The conversion rate is the percentage of users who take a desired action, e.g., percentage of users who purchase an item in a e-commerce website [28].



(a) A/B Testing (b) Interleaving

Fig. 17. Online controlled experiments: difference between A/B testing and Interleaving.

6.3 A/B Testing

As illustrated in Figure 17a, A/B Testing[30] [234, 237] randomly assigns each user to the control group (the current system A) or the treatment group (the new system B). Thus the independent variable of the experiment is the group assignment (a binary variable which identifies the system assigned to each user) and the dependent variable is measured through online evaluation measures (see Section 6.5). To conduct a reliable experiment, it is fundamental that each user is assigned randomly to each group, this ensures that the effects observed in the OEC are actually due to the change between the control and treatment group (causality is established [421]). Sometimes, it might be useful to run an A/A test [236], e.g., an A/B test where all users are assigned to the same control group. This allows to check the experimental set-up: if everything is implemented correctly, the hypothesis that the new system A is better than the old system, again A, should be rejected with high confidence ($\geq 1 - \alpha$). Moreover, A/A tests allow to collect data to compute the power of the statistical test used to assess the experimental hypothesis (see Section 4.1).

A fundamental part of A/B testing experiments is the randomization algorithm, i.e., the algorithm that maps each user to the control or treatment group. According to Kohavi et al. [234], good randomization algorithms should have the following properties:

(1) Users must be assigned to the control and treatment groups with equal probability, thus a $50 - 50$ split will avoid any bias;
(2) User assignments must be consistent, i.e., if a user access the system multiple times, he/she should be assigned to the same group every time;
(3) There must not be correlation among experiments, which means that if a user is subject to multiple experiments, the probability of a user being assigned to the control or treatment group should not depend on his/her previous assignment.

Simple versions of randomization algorithms can be implemented with a standard pseudo-random number generator and a caching mechanism. The number generator needs to be seeded only once, when the experiment starts, this will ensure that properties (1) and (3) are satisfied. Then, caching the assignment with a database (server side) or cookies (client side) will ensure property (2). Alternatively, hash functions can be applied to the user and experiment identifiers, to generate uniformly distributed numbers in a given range. Then, this range is partitioned to assign users

[30] https://exp-platform.com/

to the control and treatment groups. MD5 hash function has been proven to satisfy all the 3 properties [234].

6.4 Interleaving

As illustrated in Figure 17b, interleaving [215] assigns users to both experimental conditions at the same time (both systems A and B). This type of controlled experiment merges the rankings returned by two IR systems in a single ranking, which is then presented to the users. The single merged or interleaved ranking should minimally differ from both candidates, to ensure that users can see results from both systems in an unbiased way.

Interleaving approaches have two main components: the *mixing policy* and the *scoring rule*. The mixing policy takes as input the rankings from A and B, and returns a single interleaved ranking I. The scoring function takes as input the original rankings, the interleaved ranking, and the collected observations, and returns a score, whose purpose is to allow a fair the comparison between the systems A and B.

The idea behind interleaving is that clicks can be used as preference judgements instead of absolute judgements. Indeed, clicks might not be a good choice to estimate relevance since they are inherently biased (e.g., position bias, popularity bias) and users might click on a document even when it is not relevant. However, clicks can be used as pairwise preference judgements: if a large number of users click more a document d_i from system A and less on a document d_j from system B, we can conclude that document d_i is preferred to document d_j. Therefore, if documents from A are clicked more often than documents from B, we can infer that system A is better than system B. Note that this approach assumes that the interleaved ranking is built in an unbiased way. For example, if documents from A are ranked always higher than documents from B in the interleaved ranking, then it is more likely that documents from A receive more clicks due to the position bias.

Joachims [215] was the first to propose an interleaving approach to compare online IR systems: *balanced interleaving*. As suggested by its name, balanced interleaving returns an interleaved ranking where documents from A and B are balanced: any top k results in the interleaved ranking I contain the top k_a results from A and the top k_b from B, where k_a and k_b are equal or differ at most by 1. Algorithm 1 details balanced interleaving as presented in [313]. The first operation is to randomly select the preferred algorithm ($AFirst$), which is always chosen when there are ties. Then, k_a and k_b are pointers to the documents in A and B, which are closest to the top of the corresponding input rankings but are not yet included in the interleaved list I. The algorithm stops when the pointers k_a and k_b reach the end of the rankings.

Both A/B testing and interleaving have advantages and disadvantages. A/B testing is easier to implement and can be applied to a wide range of experimental settings (not just with rankings). The main assumption behind A/B testing is that users are independent among each other, which is reasonable for online users of IR systems. However, A/B testing suffers from high variance, since each user with his/her own information need is exposed just to one condition. Thus we need to involve a large enough number of users and ensure proper randomization.

On the other side, with interleaving, the same user is exposed to both conditions, thus we can separate the variance due to different users and information needs from the effects due to the experimental condition. However, to ensure a fair comparison between systems, we need extra assumptions when we create the interleaved ranking and we need complex scoring functions, able to correctly interpret user preferences. Thus, interleaving can be less flexible than A/B testing.

As mentioned before, online evaluation is difficult to generalize and is hard to interpret, due to all biases that affect log data. Furthermore, online evaluation comes with a cost: if during the controlled experiment (A/B testing or interleaving), we show wrong or low quality results, we can affect the online experience of thousands of users, which can turn in a loss of income.

Algorithm 1: Balanced Interleaving

Data: Rankings $A = (a[1], a[2], \ldots)$ and $B = (b[1], b[2], \ldots)$
$I \leftarrow ()$; $k_a \leftarrow 1$; $k_b \leftarrow 1$;
$AFirst \leftarrow RandBit()$; // randomly decide which ranking gets priority
while $(k_a \leq |A|) \wedge (k_b \leq |B|)$ **do**
  if $(k_a < k_b) \vee ((k_a = k_b) \wedge (AFirst = 1))$ **then**
    if $a[k_a] \notin I$ **then**
      $I \leftarrow I + a[k_a]$; // append the document in A to I
      $k_a \leftarrow k_a + 1$;
    end
  else
    if $b[k_b] \notin I$ **then**
      $I \leftarrow I + b[k_b]$; // append the document in B to I
      $k_b \leftarrow k_b + 1$;
    end
  end
end
Result: Interleaved ranking I

6.5 Online Measures

Controlled experiments rely on the *Overall Evaluation Criterion (OEC)* to deem which system is better between A and B. OEC is a "quantitative measure of the experiment's objective" [234]. It can be a single measure or a set of different measures, where the recommendation is to define a weighted sum of multiple measures [329].

Measures for online controlled experiments can be categorized in two families: *absolute measures* and *relative measures* [200]. Absolute measures return a score which represents the system quality by itself, without need to compare or relate to other systems. This score can be used to compare multiple systems or the same system at different time periods. These types of measures are usually exploited with A/B testing experiments. On the other side, relative measures return a preference judgement, i.e., a score that needs to be related to another system. This measures answer to the question whether system A is better than system B, but they do not inform on the quality of each system when considered individually. Note that transitivity might not hold: if system A is better than system B, and system B is better than system C, we can not conclude that system A is better than system C. These types of measures are usually exploited by interleaving.

Moreover, online evaluation measures are further classified based on their scope or level: *document level*, *ranking level* and *session level* measures. Hofmann et al. [200] present a comprehensive survey of online evaluation measures, we list some of them next.

Absolute document level metrics:

- *Click-Through Rate (CTR)*: is the ratio between the number of clicks and the number of impressions, i.e., the total number of times a document is displayed on the SERP. It is one of the most simple online measures, commonly used as baseline. It is noisy and affected by position bias [217], i.e., documents at the beginning of the ranking (rank positions 1, 2 and 3) receive the majority of clicks.
- Dwell-time: is the time a user spends in examining a search results (e.g., reading a Web page). A threshold of 30 seconds is commonly set to define satisfied clicks [428].

- Click models: can be used to evaluate IR systems as in discussed in Section 5.1, but they can also be used to infer the relevance of documents from user clicks. Thus, inferred relevance judgements can become the input for any offline measure.

Absolute ranking level metrics:

- Click rank: is the rank position of the first document clicked in a ranking. The closer to the top the clicks occur in a ranking, the better the IR system.
- Click reciprocal rank: is inspired by RR and is the inverse of the rank position where the first click occurs. Is a variation of click rank, but it is bounded in $[0, 1]$ and the higher the score the better the system.
- CTR@k: is CTR aggregated for the first k positions in the ranking.
- Time to click: is the time difference between the first or last click and the impression time of the SERP [69, 161]. Since time represents a cost for users examining the list, the lower the time the better the IR system. This measure needs to be considered carefully, especially in the case of good abandonment (see below).
- Good abandonment: no interactions can also be interpreted as a good signal [253]. Indeed nowadays, online IR systems can provide an answer without requiring users to click anywhere.

Absolute session level measures: simple measures to evaluate online IR systems can be derived from document level and ranking level absolute measures, e.g., time to first click, session length, number of queries in a session, etc. However, these measures can be misleading when considered with respect to the whole session. For example, if a high number of queries with many clicks occur in the same session, it can be interpreted positively, since the user is engaging with the system, but it can also hide a negative signal, i.e., the user struggling in finding relevant documents. Therefore, some absolute session level measures has been proposed, which are able to account for multiple types of user interactions simultaneously:

- Learned measures: combine several signals at different levels, e.g., CTR, dwell time, number of queries, etc. These signals are used as features for *Machine Learning (ML)* models which predict the user satisfaction as a binary classification task.
- Loyalty measures [244]: are long term measures which account for the engagement over time of a user with the IR systems. User engagement refers to the "emotional, cognitive, and behavioral experience of a user with a technological resource" [244]. These measures are especially useful for commercial IR systems, because they measure the quality of the user experience and not just the user amount of interactions. Relying only on measures such as CTR or dwell time might be misleading since they do not necessarily relate to user engagement. Examples of loyalty measures are: the number of queries/sessions per user, the number of daily queries/sessions per user, and the absence time [120], i.e., the time of the user returning to the system. These measures are hard to apply: user habits take time to establish causing a slow change in loyalty measures.

Relative measures: are used combined with interleaving experiments, where they are encoded in the scoring function. In the context of relative document level measures, Joachims et al. [217] show that users' interactions with documents in the SERP are affected by the context, i.e., the neighbouring documents. They propose the *skip above click rule*: if a user skips a document at the beginning of the ranking, and clicks a document lower in the ranking, we can assume that the user prefers the clicked document to the skipped one. Indeed the user has to make an extra effort (skip a document) to actually reach the document he/she is interested into. Once more this model assumes that the user scans the SERP from top to bottom. In a proper randomized setting, where pair of

documents are displayed on the SERP in all possible orders, this rule can be used to infer pairwise judgements of documents and thus evaluate systems at document level [311].

Several interleaving approaches have been proposed in the context of relative ranking level measures. For balanced interleaving [215] or TeamDraft [312] the preferred ranking is the one that receives more clicks. Extensions of these algorithms were proposed to merge rankings from more than two systems [356] or to exploit historical data [201].

Finally, note that there is no relative session level measure. Indeed, it is not clear how different sessions can be interleaved to produce a single session.

Whether you run an A/B test or an interleaving experiment, the definition of the OEC is crucial. Using a single measure to define the OEC might be misleading and offers only a single perspective when evaluating a system. For example, relying only on CTR can lead to wrong conclusions because (1) a click on a document does not necessarily mean that the document is relevant; (2) it is easy to "shift clicks", i.e., clicks simply move from one area of the page to another [235]. On the other side, using too many measures can add too much complexity making it hard to combine all measures in a single OEC. A rule of thumb is to use no more than 5 different measures, focusing only on key measures [237].

When evaluating commercial systems, the OEC often needs to account for the company revenue. Therefore, we need to define a trade-off between measures based on users' interactions and/or engagement and monetary revenue. Funnel approaches can help in breaking down the evaluation procedure in different steps. An example is the PULSE framework [323], which stands for Page views, Uptime, Latency, Seven-day active users, and Earnings. One shortcoming of PULSE is that an increase for some of its measures, e.g., active users, can lead to misleading results, because these are low level measures that do not reflect the actual user engagement. The HEART framework [323]: Happiness, Engagement, Adoption, Retention and Task success, focuses mainly on users' experience and engagement. The PIRATE framework [273]: AARRR! Acquisition, Activation, Retention, Referral, and Revenue includes both users' experience and revenue.

As a final remark, note that improvements of IR offline measures, such as AP or nDCG (see Sections 3 and 5) do not necessarily translate in improvements of online measures or company revenue. Empirically, only $10\% - 30\%$ of good ideas result into statistically significant improvements, which lead to business changes [204, 225, 234, 269].

7 MEASUREMENT

This section presents the foundations of measurement and discusses them and their implications from the perspective of IR evaluation, with particular reference to evaluation measures and statistical significance testing.

Sections 7.1 and 7.2 provide an intuitive overview of the representational theory of measurement and introduce the central concept of *scale* of measurement. Section 7.3 explains the classification of the scales of measurement based on their properties. Sections 7.4 and 7.5 discuss which operations among the values of a scale and which statistical significance test should be admissible, based on the scale properties. Section 7.6 explains why the general discourse about scales and their properties matters for IR evaluation. Finally, Section 7.7 presents a formal theory of IR evaluation measures to derive scale properties for IR evaluation measures and to account for their implications.

Note that the problems of admissible operations and statistical tests is a very much debated issue, both inside and outside IR, and there are different viewpoints and strong opinions about what you should or should not do. The objective of this section is not to support any specific viewpoint but rather to make the reader aware about the potential issues concerning scale properties and present the different viewpoints in order to let her/him take informed decisions.

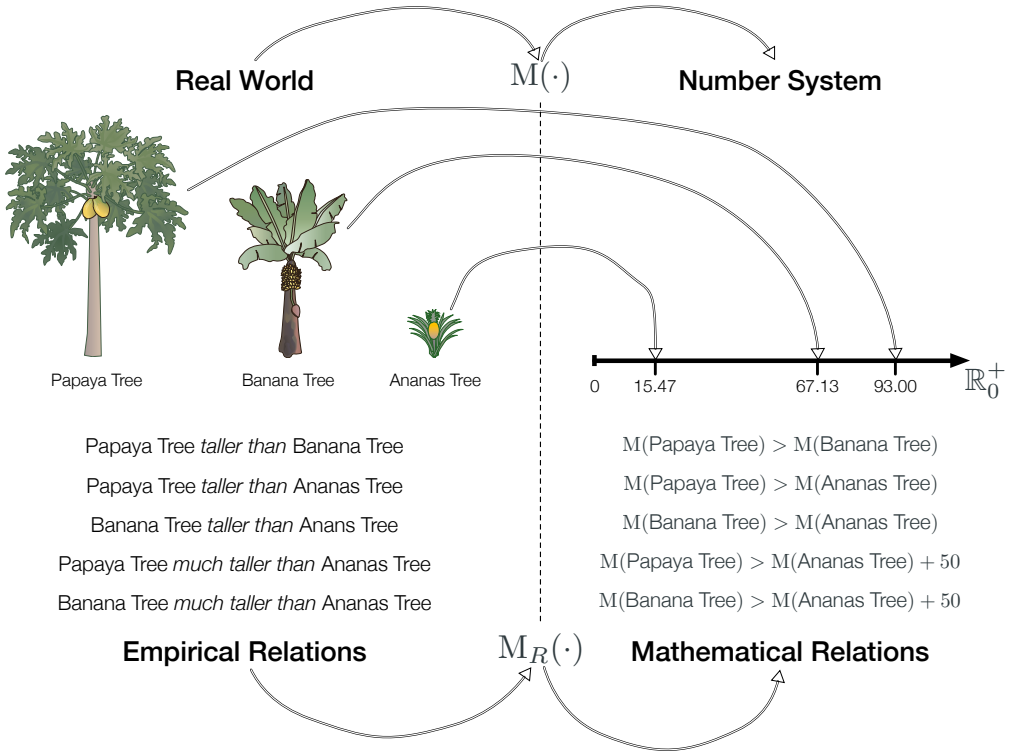


Fig. 18. Example of an empirical relations for the attribute "height" of a tree and its representation condition.

7.1 Overview

> **Measurement** is the process by which numbers (or symbols) are assigned to attributes of entities in the real world in such a way as to describe them accordingly to clearly defined rules.

The above definition of measurement by Fenton and Bieman [134] highlights several important facts about it. An *entity* is an object or an event existing in the real world, which is described by means of its identifying characteristics – the *attributes* – that allow us to distinguish one entity from another. We often define the attributes in terms of *numbers*, to easily process and work with them. In doing this, we need to preserve the empirical relationships among the (attributes of the) entities in the real world and properly translate them in the numerical domain.

Consider the example shown in Figure 18 about the attribute "heigth" of a tree, where the real world is constituted by just three entities: a Papaya tree, a Banana tree, and an Ananas tree. We can easily see that some trees are "taller than" others: for example, we can see that the Papaya tree is "taller than" the Banana and Ananas ones, while the the Banana tree is "taller than" the Ananas one. Moreover, we can have multiple relations on the same set of entities. For example, we can see that both a Papaya and a Banana tree are "much taller than" an Ananas one.

"Taller than" and "much taller than" are *empirical relations* for height (of a tree) and we can think at them as a *mapping* from the real world to a formal mathematical world. Indeed, they can be considered as a mapping from the set of trees to the set of real numbers, provided that, for

example, whenever a Papaya tree is "taller than" a Banana one, any measure of height assigns a higher number to the Banana tree than to the Papaya one.

7.2 The Representational Theory of Measurement

The *representational theory of measurement* [134, 238, 261, 378] is one of the most developed approaches to measurement, suitable for many areas of science ranging to physics and engineering, including *software engineering*, to psychology. As said, the basic idea is that real world *objects* (entities) have *attributes* which constitute their relevant features and induce a set of relationship among them. The set of objects E together with the relationships $R_1^E, R_2^E, \ldots$ among them comprise the so-called *Empirical Relational System (ERS)* $\mathbf{E} = \left\langle E, R_1^E, R_2^E, \ldots \right\rangle$. Then, we look for a mapping between the real word objects E and numbers N in such a way that the relationships $R_1^E, R_2^E, \ldots$ among the objects match with relationships $R_1^N, R_2^N, \ldots$ among numbers. The set of numbers N together with the relationships $R_1^N, R_2^N, \ldots$ constitutes the so-called *Numerical Relational System (NRS)* $\mathbf{N} = \left\langle N, R_1^N, R_2^N, \ldots \right\rangle$.

More precisely, the representational theory of measurement seeks for an *homomorphism* ϕ which maps E onto N in such a way that $\forall R_i^E, \forall e_1, e_2, \ldots, e_k \in E \mid (e_1, e_2, \ldots, e_k) \in R_i^E$ it holds that $\exists\, n_1 = \phi(e_1), n_2 = \phi(e_2), \ldots n_k = \phi(e_k) \in N \mid (n_1, n_2, \ldots, n_k) \in R_i^N$. The homomorphism ϕ is called a **scale (of measurement)**. Note that, in general, we seek for an homomorphism and not an isomorphism because two different real word objects might be mapped into the same number.

The most typical example is length. Suppose the ERS $\mathbf{E} = \left\langle E, \succsim, \circ \right\rangle$ is a set of rods with an order relationship $\succsim$ among rods and a concatenation operation $\circ$ among them. If the attribute under examination is the length of a rod, we can map the ERS to the NRS $\mathbf{N} = \left\langle \mathbb{R}_0^+, \geq, + \right\rangle$ such that $\forall e_1, e_2, e_3 \in E$ it holds $e_1 \succsim e_2 \Leftrightarrow \phi(e_1) \geq \phi(e_2)$ and $e_1 \circ e_2 \sim e_3 \Leftrightarrow \phi(e_1) + \phi(e_2) = \phi(e_3)$, that is if a rod is longer than another one the number assigned to the first one is bigger than the number assigned to the second one and the concatenation of two rods corresponds to the sum of the two numbers assigned to them.

The core of the representational theory of measurements is to seek for a *representation* theorem and a *uniqueness* theorem for the scale of measurement in order to fully define it.

The *representation theorem* ensures that if the ERS satisfies given properties, it is possible to construct an homomorphism to a certain NRS. In the previous example, the representation theorem defines which properties the order relation $\succsim$ and the concatenation $\circ$ have to satisfy in order to construct a real-valued function ϕ which is order preserving and additive. It is important to underline that the representational theory of measurement seeks for "operations" among real word objects – e.g., we can put two rods side by side to order them or we can lay two rods end by end to concatenate them – and if these "operations" satisfy given properties they can be reflected into corresponding operations among numbers, where numbers are just a proxy of what happens among real world objects but are much more convenient to manipulate.

In general, given an ERS and an NRS, it is possible to create more than one homomorphism between them. For example, it is possible to express length by using meters or yards and both of them are legitimate scales for length. The *uniqueness theorem* is concerned with determining which are the *permissible transformations* $\phi \rightarrow \phi'$ such that ϕ and ϕ' are *both* homomorphisms of the given ERS into the *same* NRS. In our example, any transformation $\phi' = \alpha\phi, \alpha > 0$ is permissible for length. Therefore, the uniqueness theorem guarantees that the "structure" of a scale of measurement is invariant to changes in the numerical assignment which preserve the relationships.

7.3 Classification of the Scales of Measurement

Stevens [375] introduced a classification of scales based on their permissible transformations, described below.

7.3.1 Nominal scale. It is used when entities of the real world can be placed into different classes or categories on the basis of their attribute under examination. The ERS consists only of different classes without any notion of ordering among them and any distinct numeric representation of the classes is an acceptable measure but there is no notion of magnitude associated with numbers. Therefore, any arithmetic operation on the numeric representation has no meaning.

The class of permissible transformations is the set of all *one-to-one mappings*, i.e., bijective functions: $\phi' = \mathrm{f}(\phi)$, since they preserve the distinction among classes.

Example 7.1 (Nominal Scale). Consider a classification of people by their country, e.g., France, Germany, Greece, Italy, Spain, etc. We could define the two following measurements:

$$\phi = \begin{cases} 5 & \text{if France} \\ 4 & \text{if Germany} \\ 3 & \text{if Greece} \\ 2 & \text{if Italy} \\ 1 & \text{if Spain} \\ \cdots & \text{if } \cdots \end{cases} \qquad \phi' = \begin{cases} 41 & \text{if France} \\ 13 & \text{if Germany} \\ -10 & \text{if Greece} \\ 23 & \text{if Italy} \\ 17 & \text{if Spain} \\ \cdots & \text{if } \cdots \end{cases}$$

both ϕ and ϕ' are valid measures, which can be related with a one-to-one mapping. Note that even if ϕ looks like being ordered, there is actually no meaning in the associated magnitudes and so it should not be confused with an ordinal scale (see below). Moreover, even if it is always possible to operate with numbers, this has no specific meaning. For example, using ϕ to perform $4 - 3 = 1$ would correspond to Germany $-$ Greece $\stackrel{?}{=}$ Spain. Similarly, using ϕ' to perform $13 - (-10) = 23$ would correspond to Germany $-$ Greece $\stackrel{?}{=}$ Italy, even in disagreement with the previous case.

7.3.2 Ordinal scale. It can be considered as a nominal scale where, in addition, there is a notion of ordering among the different classes or categories. The ERS consists of classes that are ordered with respect to the attribute under examination and any distinct numeric representation which preserves such ordering. Therefore, the magnitude of the numbers is used just to represent the ranking among classes. As a consequence, addition, subtraction or other mathematical operations have no meaning.

The class of permissible transformations is the set of all the *monotonic increasing functions*, since they preserve the ordering: $\phi' = \mathrm{f}(\phi)$.

Example 7.2 (Ordinal Scale). The European Commission Regulation 607/2009 [125] and the follow-up regulation 2019/33 [126] set the following increasing scale to classify sparkling wines on the basis of their sugar content:

- *pas dosé* (brut nature): sugar content is less than 3 grams per litre; let us call this range $s_0 = [0, 3]$;
- *extra brut*: sugar content is between 0 and 6 grams per litre; let us call this range $s_1 = [0, 6]$;
- *brut* : sugar content is less than 12 grams per litre; let us call this range $s_2 = [0, 12]$;
- *extra dry*: sugar content is between 12 and 17 grams per litre; let us call this range $s_3 = (12, 17]$;
- *sec* (dry): sugar content is between 17 and 32 grams per litre; let us call this range $s_4 = (17, 32]$;
- *demi-sec* (medium dry): sugar content is between 32 and 50 grams per litre; let us call this range $s_5 = (32, 50]$;

- *doux* (sweet): sugar content is greater than 50 grams per litre; let us call this range $s_6 = (50, 2000]$, where 2000 grams per litre is roughly the saturation of sugar in water, which is much higher than those of sugar in alcohol.

We can introduce two alternative ordinal scales ϕ and ϕ' of the above wine scale where ϕ is given by the maximum of a range[31] while ϕ' is given by a monotonic transformation $\phi' = \phi^2$:

$$\phi = \begin{cases} 3 & \text{if pas dosé} \\ 6 & \text{if extra brut} \\ 12 & \text{if brut} \\ 17 & \text{if extra dry} \\ 32 & \text{if sec} \\ 50 & \text{if demi-sec} \\ 2000 & \text{if doux} \end{cases} \qquad \phi' = \begin{cases} 9 & \text{if pas dosé} \\ 36 & \text{if extra brut} \\ 144 & \text{if brut} \\ 289 & \text{if extra dry} \\ 1024 & \text{if sec} \\ 2500 & \text{if demi-sec} \\ 4000000 & \text{if doux} \end{cases}$$

As in the case of the previous Example 7.1, mathematical operations have no specific meaning, even if, especially in the case of ϕ, we may be tempted to perform operations like $\frac{\text{brut}}{\text{extra brut}} = \frac{12}{6} = 2$ to express statements like "brut may be twice as sweet as extra brut". However, such statement cannot be expressed on the ϕ or ϕ' scale and it actually comes from implicitly changing scale to the *mass concentration* scale of the solution, which is a ratio scale (see below), where the division operation would make sense. Also addition and subtraction have no meaning, so brut − extra brut $= 12 - 6 = 6$ is not a way to express statements like "brut may have $6\,g/l$ of sugar more than extra brut", for the same reasons above. We could perform operations such as $\text{sgn}(\phi(e_1) - \phi(e_2))$ or $\text{sgn}(\phi'(e_1) - \phi'(e_2))$ but this would be just a more involute way of expressing the order among categories, which is the only property guaranteed by ordinal scales.

7.3.3 Interval scale. Besides relying on ordered classes, it also captures information about the size of the intervals that separate the classes. The ERS consists of classes that are ordered with respect to the attribute under examination and where the *size of the "gap"* among two classes is somehow understood. More precisely, fundamental to the definition of an interval scale is that *intervals must be equi-spaced*. An interval scale preserves order, as an ordinal one, and differences among classes have meaning – but not their ratio. Therefore, addition and subtraction are acceptable operations but not multiplication and division.

The class of permissible transformations is the set of all *affine transformations*: $\phi' = \alpha\phi + \beta,\ \alpha > 0$.

Note that while ratios of classes $\frac{\phi(e_1)}{\phi(e_2)}$ have no meaning on an interval scale, the ratio of differences among classes, i.e., the ratio of intervals, is allowed and invariant $\frac{\phi'(a) - \phi'(b)}{\phi'(c) - \phi'(d)} = \frac{[\alpha\phi(a) + \beta] - [\alpha\phi(b) + \beta]}{[\alpha\phi(c) + \beta] - [\alpha\phi(d) + \beta]} = \frac{\phi(a) - \phi(b)}{\phi(c) - \phi(d)}$.

Example 7.3 (Interval Scale). A typical example of interval scale is temperature, which can be expressed on either the Fahrenheit or the Celsius scale, where the affine transformation $F = \frac{9}{5}C + 32$ allows us to pass from one to the other. When talking about temperature it does not make sense to say that 20 °C is twice as hot as 10 °C, i.e., multiplication and division are not allowed. You can also note that the division operation is not invariant to the transformation, since $\frac{20\,^\circ\text{C}}{10\,^\circ\text{C}} = 2$ but $\frac{68\,^\circ\text{F}}{50\,^\circ\text{F}} = 1.36$. However, it makes sense to say that the increase between 10 °C and 20 °C is the same as the increase between 20 °C and 30 °C, i.e., addition and subtractions are allowed; you can also note that the

[31] Note that the EU regulations define intervals that are not disjoint and we follow the official definitions. This does not influence the definition of the ordinal scales above, since they are based on the maximum of each interval, which is unique and strictly increasing for each of them.

subtraction operation is invariant to the transformation since $30\,^\circ\text{C} - 20\,^\circ\text{C} = 20\,^\circ\text{C} - 10\,^\circ\text{C} = 10\,^\circ\text{C}$ and $86\,^\circ\text{F} - 68\,^\circ\text{F} = 68\,^\circ\text{F} - 50\,^\circ\text{F} = 18\,^\circ\text{F}$. Moreover, the ratio of intervals $\frac{20\,^\circ\text{C} - 10\,^\circ\text{C}}{30\,^\circ\text{C} - 20\,^\circ\text{C}} = 1$ is invariant to the transformation $\frac{68\,^\circ\text{F} - 50\,^\circ\text{F}}{86\,^\circ\text{F} - 68\,^\circ\text{F}} = 1$.

Central to the notion of temperature is the fact that the size of the "gap" has the same meaning all over the scale; indeed, 1 degree represents the same amount of thermal energy over the whole scale, i.e., the gaps are *equi-spaced*.

7.3.4 Ratio scale. It allows us to compute ratios among the different classes. The ERS consists of classes that are ordered, where there is a notion of "gap" among two classes and where the "proportion" among two classes is somehow understood. It preserves order and differences as well as ratios. Therefore, all the arithmetic operations are allowed.

The class of permissible transformations is the set of all *linear transformations*: $\phi' = \alpha\phi$, $\alpha > 0$.

Example 7.4 (Ratio Scale). A typical example of ratio scale is length which can be expressed on different scales, e.g., meters or yards, which can all be mapped one into another via a similarity transformation. For example, to pass from kilometers (ϕ) to miles (ϕ'), we have the following transformation $\phi' = 0.62\phi$.

Another example of ratio scale is the absolute temperature on the Kelvin scale where there is a zero element, which represents the absence of any thermal motion.

7.4 Admissible Statistical Operations

Stevens moved a step forward and linked the notion of scale with that of admissible statistical operations which can be carried out with that scale:

- *Nominal scale*: the only allowable operation is counting number of items in each class, that is, in statistical terms, mode and frequency.
- *Ordinal scale*: besides the operations already allowed for nominal scales, median, quantiles, and percentiles are appropriate, since there is a notion of ordering.
- *Interval scale* besides the operations already allowed for ordinal scales, mean and standard deviation are allowable since they depend just on sum and subtraction[32].
- *Ratio scale*: besides the operations already allowed for interval scales, geometric and harmonic mean, as well as coefficient of variation, are allowable since they depend on multiplication and division.

These prescriptions originated several debates over the decades. Lord [258, p. 751] argued that "since the numbers don't remember where they come from, they always behave the same way, regardless" and so any operation should be allowed even on "football numbers", i.e., a nominal scale; Gaito [170] reinforced this argument by distinguishing between the realm of the measurement theory, where Stevens's restrictions should apply, and the realm of the statistical theory, where these restrictions should not be applied, since other assumptions, such as normal distribution of the data, are those actually needed. Townsend and Ashby [391] replied back showing cases where performing operations inadmissible for a given scale of measurement may mislead the conclusions drawn by statistical tests. O'Brien [296] discussed the type of errors introduced when using ordinal data for representing an underlying continuous variable, classifying them into pure transformation errors, pure categorization errors, pure grouping errors, and random measurement errors. Velleman and Wilkinson [401] summarized the previous debate and argumented that once you are in the

[32]Note that when we talk about admissible operations, we mean operations between items of the scale. So, for example, a mean involves summing items of the scale, e.g., temperature, and this is possible on an interval scale. The fact that a mean also requires a division by the number N of items added together is not in contrast with saying that only addition and subtraction are allowed, since N is not an item of the scale.

numerical realm every operation is admissible among numbers. Recently, Scholten and Borsboom [355] made a case of flaws in the original Lord's argument and, as a striking consequence, Lord's experiment would not be a counterargument to Stevens's restrictions but it would rather comply with them. In a very recent textbook, Sauro and Lewis [351] firmly supported Lord's view, at least in the case of ordinal scales, but with the caveat to not make claims on the outcomes of a statistical test that violate the underlying scale. So, for example, if you are on ordinal scale and you detected a significant effect using a test which would require a ratio scale, you should not claim that that effect is twice as big as another effect but just that it is significant.

7.5 Statistical Significance Testing

Siegel [360] and Senders [358] have discussed the implications of Stevens' classification and permissible operations in the case of statistical inference and parametric and nonparametric statistical significance tests. We consider the following tests:

- **Sign Test**: it requires samples to be on an *ordinal scale*, since it needs to determine the sign of their difference or, equivalently, which one is greater.
- **Wilcoxon Rank Sum Test** (or Mann-Whitney U Test): it requires samples to be on an *ordinal scale*, since it needs to order them for determining their rank.
- **Wilcoxon Signed Rank Test**: it requires samples to be on an *interval scale*, since it regards the ranks of the differences, for which intervals must be equi-spaced.
- **Student's t Test**: it requires samples to be on an *interval scale*, since it needs to compute means and variances.
- **ANOVA**: it requires samples to be on an *interval scale*, since it needs to compute means and variances.
- **Kruskal-Wallis Test**: it requires samples to be on an *ordinal scale*, since it needs to order them for determining their rank.
- **Friedman Test**: it requires samples to be on an *ordinal scale*, since it needs to order them for determining their rank.

As in the case of Stevens' permissible operations, defining which statistical significance tests should be permitted on the basis of the scale properties of the investigated variables raised a lot of discussion and controversy. Anderson [23], along the line of reasoning of Lord, argued that statistical significance tests should be used regardless of scale limitations. Gardner [172] summarizes much of the discussion up to that point, leaning towards not worrying too much about scale assumptions. Gardner suggests that, if and when lack of compliance to measurement scale requirements biases the outcomes of significance tests, transformations can be applied to turn ordinal scales into more interval-like ones. For example, averaging the ranks of each score, as proposed by Gaito [169], or using a more complex set of rules, as developed by Abelson and Tukey [1]. Ware and Benson [419] replied to Gardner's positions by further revising the pro and con arguments and concluding that parametric significance tests should be used only when dealing with interval and ratio scales while, in the case of ordinal scales, nonparametric significance tests should be adopted. Townsend and Ashby [391] further investigated the issue, highlighting some serious pitfalls you may fall in, when ignoring the scale assumptions.

We can summarise the discussion with the conclusions of Marcus-Roberts and Roberts [271, p. 391]:

> The appropriateness of a statistical test of a hypothesis is just a matter of whether the population and sampling procedure satisfy the appropriate statistical model, and is not influenced by the properties of the measurement scale used. However, if we want to draw conclusions about a population which say something basic about the population,

> rather than something which is an accident of the particular scale of measurement used, then we should only test meaningful hypotheses, and meaningfulness is determined by the properties of the measurement scale in connection with the distribution of the population.

and Hand [186, p. 471]:

> Restrictions on statistical operations arising from scale type are more important in model fitting and hypothesis testing contexts than in model generation or hypothesis generation contexts.

7.6 Why the Measurement Theory Matters to IR Evaluation?

The measurement theory and permissible operations affect IR evaluation in different ways. We present some of them next.

Averaging System Performance. The most common and basic operation we perform to understand whether a system A is better than a system B is to average their performance over a set of topics and compare these aggregated scores. According to Stevens's prescriptions, this would require IR evaluation measures to be, at least, interval scales. As it happened for other areas over the decades (see Section 7.4), this prescription originated a lot of debate in IR as well.

Robertson [321] was the first to discuss the admissibility of the use of the geometric mean from the Stevens's perspective in the context of the TREC Robust track. In particular, Robertson focused on the fact that *Mean Average Precision (MAP)* and *Geometric Mean Average Precision (GMAP)* may lead to different conclusions – e.g., blind feedback is beneficial according to MAP but detrimental according to GMAP. In this respect, Robertson [321, p. 80] observed that:

> If the interval assumption is not valid for the original measure nor for any specific transformation of it, then *any* monotonic transformation of the measure is *just as good a measure* as the untransformed version. If we believe that the interval assumption is good for the original measure, that would give the arithmetic mean some validity over and above the means of transformed versions. If, however, we believe that the interval assumption might be good for one of the transformed versions, we should perhaps favour the transformed version over the original. But if there is no particular reason to believe the interval assumption for any version, then all versions are equally valid. If they differ, it is because they measure different things.

Since both AP and the log-transformation of AP (implied by the geometric mean) are not interval scales, Robertson concluded that no preference could be granted to MAP or GMAP in terms of (intrinsic) validity of their findings. In this way Robertson takes a neutral stance with respect to the debate on whether certain operations should be permitted or not on the basis of the scale properties.

Fuhr [168] took a firm position about averaging and, in particular, argued that *Mean Reciprocal Rank (MRR)* [364] should not be computed because:

(1) in general, RR is just an ordinal scale and, according to Stevens means cannot be computed for a ordinal scales;
(2) in particular, RR has some counter-intuitive behaviour.

On the other hand, Sakai [341] disagreed with Fuhr:

(1) in general, on the fact that means should not be computed for an ordinal scale, using arguments similar to those discussed in Section 7.4;
(2) in particular, on the use of RR which Sakai finds quite useful from a practical point of view.

Ferrante et al. [141, 143, 145] used the representational theory of measurement to formally model IR evaluation measures (see Section 7.7 for a brief summary of it), showing that most IR measures are not an interval scale. Ferrante et al. [136, 139] conducted extensive experimentation to assess the impact on averaging, and other operations, when IR measures depart from being an interval scale. Moffat [283] questioned the idea that most IR measures are not interval scales, advocating for the existence of interval scales which are not equi-spaced and, in turn, would allow for modelling, among other, RR and its averaging. Ferrante et al. [137] replied to Moffat's arguments not entering in the discussion on whether RR should be used or not but rather clarifying how the framework provided by the representational theory of measurement works, what are its implications in terms properties of the scales and operations over them, and why it would be worth to delve deeper into its applications to IR. Moffat [284] then reiterated his arguments in support of RR and averaging. Giner [176] further advanced the formalization of IR measures through the representational theory of measurement, confirming the theoretical results by Ferrante et al. [145] and highlighting some of their limitations, and framed this debate in the difference of views between what Michel [277, 278] calls a *representational paradigm* versus an *operational paradigm*.

As already anticipated above, the purpose of this section is neither to take a position in support of any of these stances nor to prescribe to use/to not use any evaluation measure. The objective is rather to provide readers with an understanding of the fundamentals of the representational theory of measurement, so that they can appreciate the different angles and issues entailed by measuring in IR and, potentially, further advance the knowledge in this area. Moreover, this debate is an interesting and useful example of how knowledge creation in research and science does not necessarily proceed along a straightforward path, which is the impression you might have when studying theories and models once they have been consolidated over the years, but, on the contrary, it may originate and evolve by means of discussions and rather different opinions.

Statistical Significance Testing. Statistical significance testing has a long story of adoption and investigation in IR, from the early uses of t-test reported by Salton and Lesk [345], to the discussion on the compliance with the distribution assumptions of significance tests by van Rijsbergen [399], to advocating for a more wide-spread adoption of different types of significance tests by Carterette [68], Hull [208], Sakai [335], Savoy [352], to surveys on the current state of adoption of significance tests by Sakai [337]. Again, according to the discussion in Section 7.5, we should use parametric or nonparametric tests in accordance with the scale properties of the adopted IR evaluation measures or, at least, be conscious and highlight the limitations that violating these assumptions may produce.

Score Standardization. Several authors have proposed the use of score transformation and standardisation techniques, such as z-score by Webber et al. [420] and other types of linear (and non-linear) transformations by Sakai [336], Urbano et al. [396], in order to compare performance across collections and to reduce the impact of few topics skewing the performance distribution. However, according to Stevens's prescriptions, at least an interval scale would be required to perform such transformations.

Topic Difficulty. Topic difficulty [66] is another central theme in IR because of its importance for adapting the behaviour of a system to the topic at hand. Voorhees [403, 405], in the TREC Robust tracks, explored how to evaluate and compare systems designed to deal with difficult topics and proposed to use the geometric mean, instead of the arithmetic one, for *Average Precision (AP)* [60]. However, the use of a geometric mean further raises the requirements for the evaluation measures, even calling for a ratio scale.

7.7 A Formal Theory of IR Evaluation Measures

7.7.1 Early Attempts. van Rijsbergen [398] was the first to tackle the issue of the foundations of measurement for IR by exploiting the representational theory of measurement. He observed that [398, pp. 365–366]

> The problems of measurement in information retrieval differ from those encountered in the physical sciences in one important respect. In the physical sciences there is usually an empirical ordering of the quantities we wish to measure. For example, we can establish empirically by means of a scale in which masses are equal, and which are greater or lesser than others. Such a situation docs not hold in information retrieval. In the case of the measurement of effectiveness by precision and recall, there is no *absolute* sense in which one can say that one particular pair of precision/recall values is better or worse than some other pair, or, for that matter that they are comparable at all.

Later on, van Rijsbergen [400, p. 33] further stressed this issue: "There is no empirical ordering of retrieval effectiveness and therefore any measure of retrieval effectiveness will be by necessity artificial".

van Rijsbergen addressed this issue by exploiting the *additive conjoint measurement* [238, 262]. Additive conjoint measurement was a new part of the measurement theory developed as a reaction to the views of Campbell [64, 65] and the conclusions of the Ferguson Committee of British Association for the Advancement of Science [135], where Campbell was an influential member. The committee considered the *additive property*, i.e., the concatenation operation mentioned in Section 7.1, as fundamental to science and proper measurement. As a consequence, measurement of psychological attributes, which is lacking such additive property, was not possible in a proper scientific way. As explained by Michel [278, p. 67]

> Conjoint measurement involves a situation in which two variables (A and B) are noninteractively [e.g., non additively] related to a third (X). It is not required that any of the variables be already quantified, although it is necessary that the values of X be orderable, and that values of A and B be independently identifiable (at least at a classificatory level). Then, via the order on P, ordinal and additive relations on A, B, and X may be derived

Typical examples from physics are the momentum of an object, which is affected by its mass and velocity, or the density, which is affected by its mass and volume [238].

van Rijsbergen considered retrieval effectiveness as the "orderable X" mentioned above and took precision P and recall R as the two variables A and B. In particular, he demostrated that on the relational structure $(R \times P, \succsim)$ it was possible to define an additive conjoint measurement and to derive actual measures of retrieval effectiveness from it. Note that, in this way, he avoided the need to explicitly define what an ordering by retrieval effectiveness is and he considered that precision and recall contribute independently to retrieval effectiveness. The problem of how to order runs in the ERS has been addressed some years later by Ferrante et al. [141, 143, 145]. More subtly, van Rijsbergen treats precision and recall as two attributes which can be jointly exploited to order retrieval effectiveness but, each of them, is already a measure and quantification of retrieval effectiveness and, thus, this brings some circularity in the reasoning.

Bollmann and Cherniavsky [43, 44] built on the conjoint measurement work by van Rijsbergen and applied it to further study under which conditions the *MZ-metric* [194]. In particular, Bollmann and Cherniavsky leveraged what they called *transformational viewpoints*, i.e., elementary transformations of the runs.

In the case of set-based measures, Bollmann [42] showed that measures complying with a monotonicity and an Archimedean axiom are a linear combination of the number of relevant

retrieved documents and the number of not relevant not retrieved documents. Bollmann further showed how this could be related to collections and sub-collections.

7.7.2 Current Studies. Ferrante et al. [141, 143, 145] leveraged the representational theory of measurement for developing a formal theory of IR evaluation measures, which allows us to determine the scale properties of an evaluation measure. In particular, they defined an ERS for system runs and used two basic operations to determine an ordering of runs. The two basic operations are *swap*, i.e., swapping a relevant with a not-relevant document in a ranking, and *replacement*, i.e., substituting a relevant document with a not-relevant one. In this way, they demonstrated that there exists a *partial order of runs* such that: when runs are comparable, all the measures agree on the same way of ordering them; however, when runs are not comparable, measures may disagree on how to order them. By using properties of the partial orders and theorems from the representational theory of measurement, they were able to define an interval scale measure ϕ. Moreover, they check whether there is any linear transformation between such measure ϕ and IR evaluation measures, in order to determine if the latter are interval scales too.

In particular, with respect to the identified order of runs, Ferrante et al. have defined the following interval scale measures:

- the *Set-Based Total Order (SBTO)* measure

$$\mathrm{SBTO}(\hat{r}) = \sum_{i=1}^{N} \binom{\hat{r}_t[i] + N - i}{N - i + 1}$$

  where N is the length of the run;
- the *Rank-Based Total Order (RBTO)* measure

$$\mathrm{RBTO}(\hat{r}) = \sum_{i=1}^{N} \hat{r}_t[i](c+1)^{N-i}$$

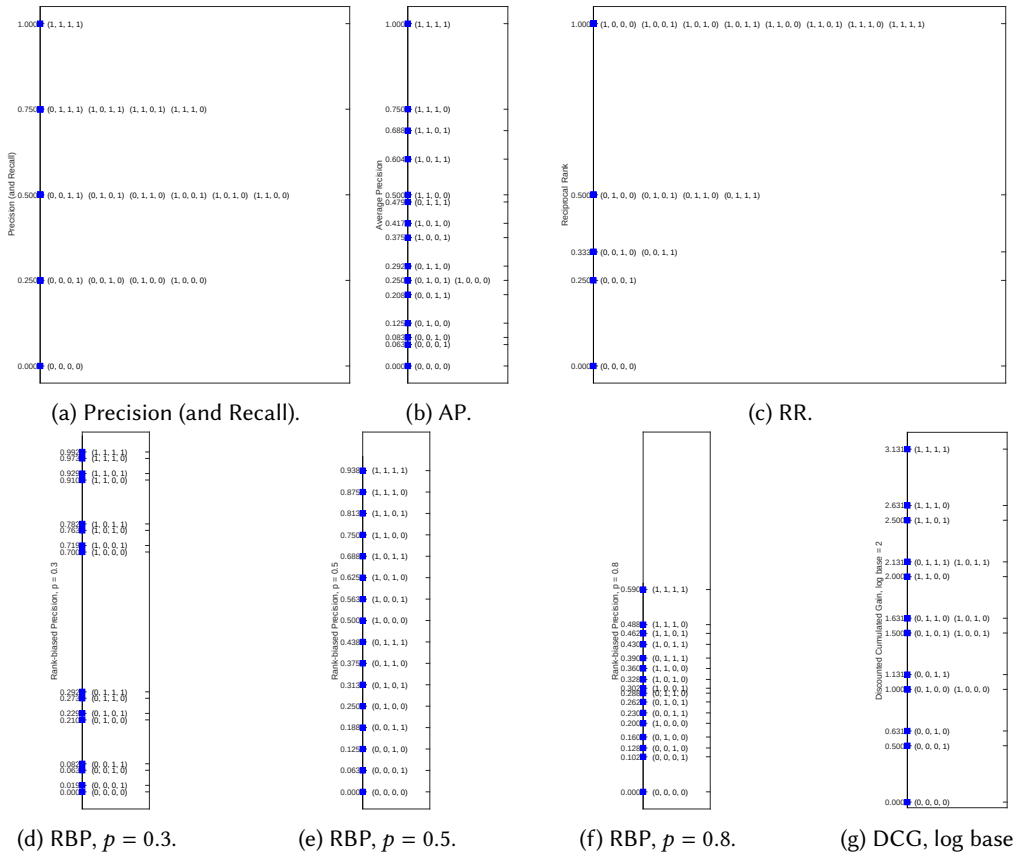
  where N is the length of the run and c is the total number of relevance degrees.

By leveraging these measures and with respect to the identified order of runs, Ferrante et al. found that, for a single topic:

- set-based evaluation measures:
  - binary relevance: precision, recall, F-measure are interval scales;
  - multi-graded relevance: *Generalized Precision (gP)* and *Generalized Recall (gR)* are interval scales only if the relevance degrees are on a ratio scale;
- rank-based evaluation measures:
  - binary relevance: *Rank-Biased Precision (RBP)* [288] is an interval scale only for $p = 1/2$; *Average Precision (AP)* is not an interval scale;
  - multi-graded relevance: *Graded Rank-Biased Precision (gRBP)* is an interval scale only for $p = G/(G+1)$, where G is the normalized smallest gap between the gain of two consecutive relevance degrees, and the relevance degrees themselves are on a ratio scale; *Discounted Cumulated Gain (DCG)* [214] and *Expected Reciprocal Rank (ERR)* [70] are not interval scales.

Ferrante et al. [136] point on how the main reason for IR measures not being interval scales is that they are not equi-spaced. In this respect, Figure 19 shows how different measures – namely, Precision (and Recall[33]), AP, RR, RBP with $p \in \{0.3, 0.5, 0.8\}$, and DCG with log base 2 – order and space all the possible runs of length $N = 4$ assuming a recall base $RB = 4$.

[33]Note that in this specific case Precision and Recall yield to the same scores because the length of the run $N = 4$ and the recall base $RB = 4$ are the same.



(a) Precision (and Recall). (b) AP. (c) RR.

(d) RBP, $p = 0.3$. (e) RBP, $p = 0.5$. (f) RBP, $p = 0.8$. (g) DCG, log base 2.

Fig. 19. Ordering and spacing of the all the possible runs of length $N = 4$ by different evaluation measures. Each blue square corresponds to a score of a given measure. On the right of the square, the run corresponding to that score is reported; in case of tied runs, i.e., runs for which the measures produces the same score, they are all listed on the right of the square.

We can observe that only Precision (Recall) and RBP with $p = 0.5$ produce equi-spaced values, while all the other measures violate this assumption, required to obtain an interval scale. We can also notice how IR measures behave differently in violating the equi-spacing assumption. RBP with $p \in \{0.3, 0.8\}$ and DCG follow a somehow regular pattern, where scores are not equi-spaced but they are in some way evenly clustered and they are symmetric if you fold the figure along its middle horizontal axis. On the other hand, AP and RR follow a much more irregular and not symmetric pattern.

We can also note how these measures spread values in their range differently. Precision (and Recall) and DCG spread their values all over the possible range while this is not always the case with RBP. Indeed, RBP assumes runs of infinite length and normalizes by the $\frac{1}{1-p}$ factor. However, we deal with runs of limited length and the $\frac{1}{1-p}$ factor is an overestimation, the bigger the overestimation the bigger is the value of p and the smaller is the length of the run – this is more clearly visible in the case of RBP with $p = 0.8$ in Figure 19f. Finally, AP, RBP with $p = 0.3$, and RR, i.e., those measures farther from being interval scales, leave large portions of their possible range completely unused. In particular, AP leaves one quarter of its range unused, in the top part roughly corresponding

Table 3. Tukey HSD test using the Kruskal-Wallis test and ANOVA. Each cell contains the number of significantly different pairs detected and, within parenthesis, the ratio with respect to the total number of system pairs.

Set-based measures, Binary Relevance – T08, 8,256 system pairs compared		
	Significantly Different Pairs	
Measure Pair	**Kruskal-Wallis Test**	**ANOVA**
Precision	1,566 (18.97%)	2,785 (33.73%)
Recall	1,748 (21.17%)	3,259 (39.47%)
F-measure	1,721 (20.85%)	3,081 (37.32%)
SBTO	1,566 (18.97%)	2,785 (33.73%)
Rank-based measures, Binary Relevance – T08, 8,256 system pairs compared		
	Significantly Different Pairs	
Measure Pair	**Kruskal-Wallis Test**	**ANOVA**
RBP $p = 1/2$	1,677 (20.31%)	2,861 (34.65%)
RBP $p = 0.2$	1,675 (20.29%)	2,198 (26.62%)
RBP $p = 0.8$	1,783 (21.60%)	3,476 (42.10%)
AP	1,824 (22.09%)	3,320 (40.21%)
RBTO	1,677 (20.31%)	2,861 (34.65%)

to the first quartile of the possible values; RR leaves one half of its range unused, in the top part roughly corresponding to the first and second quartiles of the possible values; and, finally, RBP with $p = 0.3$ leaves half of its range empty, in the middle part roughly corresponding to the second and third quartile of the possible values.

Giner [176] further advances the application of the representational theory of measurement to IR evaluation measures, by deepening the axiomatization of orderings of runs in the empirical world, which he calls *Axiomatic Model of Preferences (AMP)*, from the view point of the lattice theory [111, 180], and by studying the structural properties of these AMPs, how IR measures can be defined over them, and their derived properties.

7.8 Implications on statistical significance testing

Ferrante et al. [139] conducted a preliminary investigation on the impact on violating (or not) the scale assumptions for statistical significance tests. They considered the Kruskal-Wallis test, which is suitable for both ordinal and interval scales, and ANOVA, which is suitable only for interval scales. Being a parametric test, ANOVA is more powerful than the Kruskal-Wallis test and, generally speaking, it is able to spot more differences among the compared systems. They applied the Tukey HSD correction to both Kruskal-Wallis test and ANOVA in order to properly deal with multiple comparisons.

Table 3 reports the results of the Tukey HSD test and the number of significantly different pairs detected for both the Kruskal-Wallis test and ANOVA for different evaluation measures.

All the set-based measures are on an interval scale and so they are suitable for being used with both the Kruskal-Wallis test and ANOVA. We can observe that, as expected, they all detect a comparable number of significantly different pairs and that this number increases when ANOVA is used, since it is a more powerful test than Kruskal-Wallis.

In the case of rank-based measures, RBTO and RBP with $p = 1/2$ are interval scales and they match the scale assumptions behind both the Kruskal-Wallis test and ANOVA. RBP with $p = 0.2$ is

an ordinal scale and, therefore, it matches the scale assumptions for the Kruskal-Wallis test, but not for ANOVA. We can note how, in the case of the Kruskal-Wallis test, it detects more or less the same number of significantly different pairs while for ANOVA, being provided with a less powerful scale than the one assumed, it detects less significantly different pairs. When it comes to RBP with $p = 0.8$ and AP, they are neither ordinal nor interval scales and we can observe that they detect a higher number of significantly different pairs.

Overall, this preliminary study suggests that violating the scale assumptions may have an impact on the number of significantly different pairs detected.

7.8.1 Other Studies. Busin and Mizzaro [62], Maddalena and Mizzaro [265] and Amigó and Mizzaro [21] proposed a unifying framework for ranking, classification, and clustering measures, which is rooted in the representational theory of measurement as well. They considered scales but as a way of mapping between relevance judgements (assessor scales) and *Retrieval Status Value (RSV)* (system scales). Moreover, they introduced axioms over measures, instead of studying which are the scales actually used by IR evaluation measures and their impact on actual experiments. Amigó et al. [17] investigated the formal definition and properties of measures for data mining tasks while Amigó et al. [15] focused on measures for fairness in a Recommender Systems scenario.

Even if not specifically focused on scales and their relationship to IR evaluation measures, there is a bulk of research on studying which constraints define the core properties of evaluation measures: Amigó et al. [16, 18–20, 22] and Sebastiani [357] face this issue from a formal and theoretical point of view, applying it to various tasks such as ranking, filtering, diversity and quantification, while Moffat [282] adopts a more numerical approach.

8 TRENDS AND RESEARCH ISSUES

As it emerges from the previous sections, it is plenty of open research issues about experimental evaluation: how to improve pooling; how to develop evaluation measures which incorporate credible and realistic user models; how to bridge the gap between offline and online evaluation and develop offline measures able to predict online user behaviour; how to improve A/B testing and interleaving; and, how to avoid bias in online evaluation, just to mention a few possibilities.

However, in this section, we would like to emphasize four fundamental problems which are central to the scientific method and transversal to all the topics which have been discussed in this chapter, namely *evaluation of complex tasks*, *reproducibility*, *meaningfulness*, and *Large Language Models (LLMs)* and generative *Artificial Intelligence (AI)*.

Evaluation of Complex Tasks. Evaluation measures rely on the notion of relevance, i.e., the pertinence of a document to the user's information need. Although being intuitive, the concept of relevance is rather complex and highly subjective, since only the final user can determine whether a document is relevant or not. A number of studies investigated the concept of relevance [47, 99, 281, 350], concluding that relevance is multidimensional and dynamic. Moreover, depending on the task and the application domain, some aspects might be more important than other.

The multidimensional nature of relevance calls for new evaluation measures and frameworks able to account for multiple criteria simultaneously [268]. For example, in the consumer health domain, a good IR system should promote relevant, credible, and correct information while avoiding misinformation, i.e., relevant, credible, but incorrect information. The TREC Health Misinformation Track [3, 89] exploited measures such as *Convex Aggregating Measure (CAM)* [255], the *Multidimensional Measure (MM)* [304] framework and Compatibility [84, 91] to account for relevance, credibility and correctness simultaneously. Similarly, CLEF e-health [177, 178, 231, 377] exploited *Understandability-biased Rank-Biased Precision (uRBP)* [436] to account for both topicality,

i.e., relevance, and understandability of the retrieved documents, and *Credibility-biased Rank-Biased Precision (cRBP)*, to account for topicality and credibility.

The availability of log data allows to personalize the evaluation process, i.e., personalized relevance judgements are inferred from log data. This calls for an evaluation framework for *Personalized Information Retrieval (PIR)* systems able to incorporate aspects such as context, user satisfaction and/or perception, usability, etc. [291, 381]. Many approaches are user centered, i.e., they exploit user studies to evaluate the overall quality of the system [229]. These can be validated with log based approaches, accounting for variables such as the dwell time, return rate, etc. [244]. A middle solution between user studies and log data is to simulate the user behaviour, for example White et al. [423] generated simulated search paths corresponding to different search scenarios.

Finally, bias and fairness represent further aspects that are important to consider during the evaluation process, especially when personalization or log data are used in the evaluation process. In recent years, ML models have been widely exploited in decision-making, resulting in the consequential decisions being biased or unfair towards specific groups [280]. IR is not an exception, since increasingly complex ML models are used to automatically rank documents. In the IR context, fairness has been defined in terms of fair exposure: items should receive an amount of attention proportional to their worthiness, which can be estimated with relevance [40]. Singh and Joachims [363] assume that exposure mostly depends on the rank position where a document is displayed. They propose a probabilistic framework to define fairness of exposure, thus to determine when a disparate treatment occurs and measure its extent. On a similar line of work, the TREC Fair Ranking track [38, 39] aims at designing an evaluation protocol for fair ranking. To measure individual and group exposure, the track organizers exploited expected exposure [119], a modification of the cascade user model behind ERR (see Section 5.1).

Reproducibility. Whatever your (philosophical) stance about what science is [240, 308, 310], there is wide agreement that experimentation is at the core of the modern scientific method and that the possibility of carrying out the (same) experiments more than once is central to science in terms of validation and confirmation of the findings. Moreover, the fourth paradigm [197] opened a completely new way of making scientific discoveries, based on computational and data-intensive approaches, where experimentation originated by computation is getting even more prominent.

We are today facing the so-called *reproducibility crisis* [32, 299] across all areas of science, where researchers fail to reproduce and confirm previous experimental findings. This crisis obviously involves also the more recent computational and data-intensive sciences [162, 293], including hot areas such as artificial intelligence and machine learning [174]. For example, Baker [32] reports that roughly 70% of researchers in physics and engineering fail to reproduce someone else's experiments and roughly 50% fail to reproduce even their own experiments.

IR is not an exception and researchers are paying more and more attention to what the reproducibility crisis may mean for the field [147–149, 151], even more with the raise of the new deep learning and neural approaches [101, 146].

Even in the most favorable context of offline evaluation and Cranfield-based evaluation campaigns, where repeatability of experiments should be taken for granted, several issues arise, among which: it is often difficult (if not impossible) to re-run someone else's experiments, even if the full source code is available, because of, e.g., not properly described deep tunings; if an approach relies on online resources, like Wikipedia, or ephemeral resources, like tweets, it becomes practically unfeasible to re-run exactly the same experiment; it is even more difficult to know whether an observed effect will still hold if you change topics, document corpora, or both. This situation is even harder in the case of industrial data [63] or user-oriented evaluation where, just to name a few, sensitivity and privacy concerns, differences among user cohorts, the impossibility of recreating experimental

data arise. And least but not last, how do we know when reproduced is reproduced? The current attitude is some sort of "close enough": researchers put any reasonable effort to understand how an approach was implemented and how an experiment was conducted and, after some (several) iterations, when they obtain performance scores which somehow resemble the original ones, they decide that an experimental result is reproduced. Breuer et al. [53, 54] and Maistro et al. [267] have started to investigate this issue and proposed some objective measures to quantify reproducibility but this applies just to offline evaluation while the issue becomes more and more challenging as you move towards user-oriented evaluation and interactive IR.

Meaningfulness. In Section 7, we presented the foundations of measurement, the notion of scale properties, the consequent limitations to the performed operations and statistical tests, and how all of this matters and affects IR evaluation. In Section 7, we also discussed how much these limitations have been debated over the decades, both within and outside IR, and how there is not definitive agreement on whether you should stick with them or not.

However, both Hand [186] and Michel [277, 278] argued that the problem is not what operations you can perform with numbers but what kind of *inference* you wish to make from those operations and how much such inference has to be indicative of what actually happens among real world objects. Already Adams et al. [4, pp. 99-100] explicitly stated that:

> Statistical operations on measurements of a given scale are not appropriate or inappropriate *per se* but only relative to the kinds of statements made about them. The criterion of *appropriateness* for a statement about a statistical operation is that the statement be *empirically meaningful* in the sense that its truth or falsity must be invariant under permissible transformations of the underlying scale[34].

These statements opened the way to the development of a full (formal) theory of *meaningfulness* [133, 292, 320], which is a central concept to clearly shape and define the questions discussed above: according to the adopted measurement scales, what processing, manipulation, and analyses can be conducted and what can we tell about the conclusions drawn from such processing?

As observed by Ferrante et al. [136, 137], whatever stance you wish to take about whether (or not) operations should be constrained by scale properties, from the discussion so far, it clearly emerges that IR needs further and systematic investigation about the implications and impact of derogating from compliance with scale properties. Moreover, most of the debate is just about averaging values and does not tackles the implications for statistical significance testing. Finally, and more importantly, we completely lack a thorough discussion on and any adoption of the notion of *meaningfulness* in IR and this is quite striking for a discipline so strongly rooted in experimentation and so much based on inference.

Large Language Models and Generative AI. Generative LLMs, both proprietary such as *Generative Pre-trained Transformer (GPT)* [300, 302], and open source such as *Large Language Model Meta AI (LLaMA)* [389, 390] and its derivatives [76, 384] are being successfully applied to a wide range of different tasks, covering multiple media and modalities. As a consequence, they are gaining more and more attention from researchers, industry, and also the general public and *prompting*, i.e. seeking for a a textual input and instructions which make a LLM perform the desired task, has become a new area of investigation. Obviously, LLMs have an impact also on evaluation since they call for two fundamental questions: (i) how can we evaluate their quality, reliability, reasoning capabilities and more? and, (ii) how can we employ them to improve our evaluation methodologies?

[34]For example, the statement "A mouse weights more than an elephant" is meaningful even if it is clearly false; indeed, its truth value, i.e., false, does not change whatever weight scale you use (kilograms, pounds, and so on).

When it comes to the first question, current evaluation procedures tend to consider commonsense reasoning [41, 83, 112, 279, 333, 380, 433], world knowledge [219, 243], reading comprehension [77, 82], math capabilities [95], and coding tasks [74]. Some popular aggregated benchmarks are *Massive Multitask Language Understanding (MMLU)* [195], *BIG-Bench Hard (BBH)* [374] and *Artificial General Intelligence (AGI)* Eval [434]. Chen et al. [73], developed a dataset, in both Chinese and English, to evaluate how well LLMs avoid to hallucinate by exploiting *Retrieval-Augmented Generation (RAG)* [250]. RAG is a two-step technique where, in the first step, an IR system retrieves snippets relevant to a user query and, in the second step, these snippets are used to augment the prompt to the LLM. Along these lines, Gao et al. [171] developed a datasets for evaluating how well LLMs generate text with citations, improving their factual correctness and verifiability. Kamalloo et al. [221] proposed a dataset for building end-to-end generative information-seeking models that are capable of retrieving candidate quotes and generating attributed explanations. Rashkin et al. [317] developed a dataset and a two-stage annotation pipeline to evaluate attribution of LLMs. On a slightly different stance, Gienapp et al. [175] proposed a general framework for what they called *generative ad-hoc IR*, blending together, in possibly multi-turn iterations, RAG and retroactively retrieving references for a generated statement, similarly to claim verification.

Large-scale evaluation campaigns have started to focus on LLMs, as well, organizing specific tasks focused on their evaluation. TREC 2024 is exploring LLMs mainly in three tracks: (i) the NeuCLIR track on neural cross-language IR organizes a task on query-driven report generation, in a RAG-like fashion and in a multilingual document setting; the new biomedical generative retrieval track explores different combinations of generative LLM and RAG in a biomedicine scenario; finally, the new RAG track specifically focuses on developing effective and trustworthy methodologies for the evaluation of RAG systems. CLEF 2024 investigates LLMs in two labs. The ELOQUENT lab[35] [224] assesses certain quality aspects of content generated by LLMs, seeking answers to the following questions: can an LLM assess itself if it is capable to process data in some application domain of interest? Can an LLM be used to evaluate the output of other LLMs to detect hallucinated or factually incorrect information? Will an LLM output the same content independent of input variation which is equivalent in content but non-identical in form or style? Can an LLM be used to detect if some piece of text is written by a human author or generated by an LLM? The Monster Track lab[36] [150] focuses on the versatility of LLMs: it acts as a meta-challenge across a selection of tasks chosen from other evaluation labs running in CLEF by asking participants to develop a generative AI or LLM-based system that is run against all these tasks with no or just minimal task adaptation.

When it comes to the second question, i.e. how LLMs can be employed for improving IR evaluation methodologies, Thomas et al. [386] used LLMs to deploy large-scale relevance labelling at Bing, observing that LLMs can be effective, with accuracy as good as human labellers, but finding that changes in the prompt, even simple paraphrases, can impact on accuracy. As part of the outcomes of the Dagstuhl Seminar 23031 on "Frontiers of Information Access Experimentation for Research and Education" [35, 36], Faggioli et al. [129, 130] discuss possible ways for LLMs to support relevance judgments along with the concerns and issues that arise, foreseeing a full human–machine collaboration spectrum which ranges from fully automated relevance judgments via LLM

[35] https://eloquent-lab.github.io/
[36] https://monsterclef.dei.unipd.it/

to fully manual ones, and revamping a long-standing discussion in the field about automated relevance judgement creation [326].

Overall, all these issues – evaluation of complex tasks, reproducibility, meaningfulness, and evaluation of and with LLMs – go back to one of the core questions for IR experimental evaluation, discussed in Section 2.1: what is the *quality*, *validity*, and *generalizability* of our experimental findings? We should keep striving to properly answer them.

ACKNOWLEDGMENTS

This paper was partially supported by the EU Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893667.

REFERENCES

[1] R. P. Abelson and J. W. Tukey. 1959. Efficient Conversion Of Non-Metric Information Into Metric Information. In *Proc. of the Social Statistics Section of the American Statistical Association.* American Statistical Association, Washington, USA, 226–230.

[2] I. Abraham, O. Alonso, V. Kandylas, R. Patel, S. Shelford, and A. Slivkins. 2016. How Many Workers to Ask?: Adaptive Exploration for Collecting High Quality Labels, See [305], 473–482.

[3] M. Abualsaud, C. Lioma, M. Maistro, M. D. Smucker, and G. Zuccon. 2020. Overview of the TREC 2019 Decision Track, See [414].

[4] E. W. Adams, R. F. Fagot, and R. E. Robinson. 1965. A theory of appropriate statistics. *Psychometrika* 30 (June 1965), 99–127.

[5] E. Agirre, G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. 2009. CLEF 2008: Ad Hoc Track Overview. In *Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers*, C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, and A. Peñas (Eds.). Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany, 15–37.

[6] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. 2009. Diversifying Search Results, See [30], 5–14.

[7] J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel (Eds.). 2009. *Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009).* ACM Press, New York, USA.

[8] J. Allan, D. K. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. M. Voorhees. 2018. TREC 2017 Common Core Track Overview. In *The Twenty-Sixth Text REtrieval Conference Proceedings (TREC 2017)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-324, Washington, USA.

[9] J. Allan, D. K. Harman, E. Kanoulas, and E. M. Voorhees. 2019. TREC 2018 Common Core Track Overview, See [413].

[10] O. Alonso. 2013. Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval* 16, 2 (April 2013), 101–120.

[11] O. Alonso. 2019. *The Practice of Crowdsourcing.* Morgan & Claypool Publishers, USA.

[12] O. Alonso and S. Mizzaro. 2012. Using Crowdsourcing for TREC Relevance Assessment. *Information Processing & Management* 48, 6 (November 2012), 1053–1066.

[13] E. Amigó, M. Carrillo de Albornoz, J. andAlmagro-Cádiz, J. Gonzalo, J. andRodríguez-Vidal, and M. F. Verdejo. 2017. EvALL: Open Access Evaluation for Information Access Systems, See [223], 1301–1304.

[14] E. Amigó, P. Castells, J. Gonzalo, B. A. Carterette, J. S. Culpepper, and G. Kazai (Eds.). 2022. *Proc. 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022).* ACM Press, New York, USA.

[15] E. Amigó, Y. Deldjoo, S. Mizzaro, and A. Bellogín. 2023. A unifying and general account of fairness measurement in recommender systems. *Information Processing & Management* 60, 1 (January 2023), 103115:1–103115:19.

[16] E. Amigó, J. Gonzalo, J. Artiles, and M. F. Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12, 4 (August 2009), 461–486.

[17] E. Amigó, J. Gonzalo, and S. Mizzaro. 2023. What is My Problem? Identifying Formal Tasks and Metrics in Data Mining on the Basis of Measurement Theory. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 35, 2 (February 2023), 2147–2157.

[18] E. Amigo, J. Gonzalo, S. Mizzaro, and J. Carrillo de Albornoz. 2020. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.). Association for Computational Linguistics, USA, 3938–3949.

[19] E. Amigó, J. Gonzalo, and M. F. Verdejo. 2013. A General Evaluation Measure for Document Organization Tasks, See [218], 643–652.
[20] E. Amigó, J. Gonzalo, M. F. Verdejo, and D. Spina. 2019. A comparison of filtering evaluation metrics based on formal constraints. *Information Retrieval Journal* 22, 6 (December 2019), 581–619.
[21] E. Amigó and S. Mizzaro. 2020. On the nature of information access evaluation metrics: a unifying framework. *Information Retrieval Journal* 23, 3 (June 2020), 318–386.
[22] E. Amigó, D. Spina, and J. Carrillo-de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric, See [98], 625–634.
[23] NN. H. Anderson. 1961. Scales and Statistics: Parametric and Nonparametric. *Psychological Bulletin* 58, 4 (1961), 305–316.
[24] M. Angelini, N. Ferro, B. Larsen, H. Müller, G. Santucci, G. Silvello, and T. Tsikrika. 2014. Measuring and Analyzing the Scholarly Impact of Experimental Evaluation Initiatives. In *Proc. 10th Italian Research Conference on Digital Libraries (IRCDL 2014)*, M. Agosti, T. Catarci, and F. Esposito (Eds.), Vol. 38. Procedia Computer Science, Vol. 38, 133–137.
[25] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. EvaluatIR: an Online Tool for Evaluating and Comparing IR Systems, See [7], 833.
[26] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998, See [75], 601–610.
[27] J. A. Aslam, E. Yilmaz, and V. Pavlu. 2005. The Maximum Entropy Method for Analyzing Retrieval Measures, See [31], 27–34.
[28] A. Ayanso and R. Yoogalingam. 2009. Profiling Retail Web Site Functionalities and Conversion Rates: A Cluster Analysis. *International Journal Electronic Commerce* 14, 1 (2009), 79–114.
[29] L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra (Eds.). 2019. *Advances in Information Retrieval. Proc. 41st European Conference on IR Research (ECIR 2019) – Part I.* Lecture Notes in Computer Science (LNCS) 11437, Springer, Heidelberg, Germany.
[30] R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, and B. B. Cambazoglu (Eds.). 2009. *Proc. 2nd ACM International Conference on Web Searching and Data Mining (WSDM 2009).* ACM Press, New York, USA.
[31] R. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait (Eds.). 2005. *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005).* ACM Press, New York, USA.
[32] M. Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533 (May 2016), 452–454.
[33] D. Banks, P. Over, and N.-F. Zhang. 1999. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval* 1, 1-2 (May 1999), 7–34.
[34] M. Bashir, J. Anderton, J. Wu, M. Ekstrand-Abueg, P. B. Golbus, V. Pavlu, and J. A. Aslam. 2013. Northeastern University Runs at the TREC12 Crowdsourcing Track, See [411].
[35] C. Bauer, B. Carterette, N. Ferro, N. Fuhr, J. Beel, T. Breuer, C. L. A. Clarke, A. Crescenzi, G. Demartini, G. M. Di Nunzio, L. Dietz, G. Faggioli, B. Ferwerda, M. Fröbe, M. Hagen, A. Hanbury, C. Hauff, D. Jannach, N. Kando, E. Kanoulas, B. P. Knijnenburg, U. Kruschwitz, M. Li, M. Maistro, L. Michiels, A. Papenmeier, M. Potthast, P. Rosso, A. Said, P. Schaer, C. Seifert, D. Spina, B. Stein, N. Tintarev, J. Urbano, H. Wachsmuth, M. C. Willemsen, and J. Zobel. 2023. Report on the Dagstuhl Seminar on Frontiers of Information Access Experimentation for Research and Education. *SIGIR Forum* 57, 1 (June 2023), 7:1–7:28.
[36] C. Bauer, B. A. Carterette, N. Ferro, N. Fuhr, and G. Faggioli (Eds.). 2023. *Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education.* Dagstuhl Reports, Volume 13, Number 1, pages 68–154. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany.
[37] Y. Benjamini and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300.
[38] A. J. Biega, F. Diaz, M. D. Ekstrand, S. Feldman, and S. Kohlmeier. 2021. Overview of the TREC 2020 Fair Ranking Track, See [415].
[39] A. J. Biega, F. Diaz, M. D. Ekstrand, and S. Kohlmeier. 2020. Overview of the TREC 2019 Fair Ranking Track, See [414].
[40] A. J. Biega, K. P. Gummadi, and G. Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings, See [98], 405–414.
[41] Y. Bisk, R. Zellers, R. Le Bras, J. Gao, and Y. Choi. 2020. Shared Tasks as Tutorials: A Methodical Approach. In *Proc. 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*, P. Stone, F. Rossi, V. Conitzer, and F. Sha (Eds.), Vol. 34. Proc. 34th AAAI Conference on Artificial Intelligence (AAAI 2020), 7432–7439.
[42] P. Bollmann. 1984. Two Axioms for Evaluation Measures in Information Retrieval. In *Proc. of the Third Joint BCS and ACM Symposium on Research and Development in Information Retrieval*, C. J. van Rijsbergen (Ed.). Cambridge University Press, UK, 233–245.

[43] P. Bollmann and V. S. Cherniavsky. 1980. Measurement-theoretical investigation of the MZ-metric. In *Proc. 3rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1980)*, C. J. van Rijsbergen (Ed.). ACM Press, New York, USA, 256–267.

[44] P. Bollmann and V. S. Cherniavsky. 1981. Restricted Evaluation in Information Retrieval. In *Proc. 4th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1981)*, C. J. Crouch, W. S. Cooper, and J. Herr (Eds.). ACM Press, New York, USA, 15–21.

[45] A. Bondarenko, M. Fröbe, J. Kiesel, F. Schlatt, V. Barriere, B. Ravenet, L. Hemamou, S. Luck, J. H. Reimer, B. Stein, M. Potthast, and M. Hagen. 2023. Overview of Touché 2023: Argument and Causal Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, A. Aliannejadi, M. Vlachos, G. Faggioli, and N. Ferro (Eds.). Lecture Notes in Computer Science (LNCS) 14163, Springer, Heidelberg, Germany, 507–530.

[46] C. E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilità.* Number 8 in Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze. Libreria internazionale Seeber, Firenze, Italia.

[47] P. Borlund. 2003. The Concept of Relevance in IR. *Journal of the American Society for Information Science and Technology (JASIST)* 54, 10 (August 2003), 913–925.

[48] M. Braschler. 2001. CLEF 2000 – Overview of Results. In *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000)*, C. Peters (Ed.). Lecture Notes in Computer Science (LNCS) 2069, Springer, Heidelberg, Germany, 89–101.

[49] M. Braschler. 2002. CLEF 2001 – Overview of Results. In *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross–Language Evaluation Forum (CLEF 2001) Revised Papers*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck (Eds.). Lecture Notes in Computer Science (LNCS) 2406, Springer, Heidelberg, Germany, 9–26.

[50] M. Braschler. 2003. CLEF 2002 – Overview of Results. In *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross–Language Evaluation Forum (CLEF 2002) Revised Papers*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck (Eds.). Lecture Notes in Computer Science (LNCS) 2785, Springer, Heidelberg, Germany, 9–27.

[51] M. Braschler. 2004. CLEF 2003 – Overview of Results. In *Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers*, C. Peters, M. Braschler, J. Gonzalo, and M. Kluck (Eds.). Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany, 44–63.

[52] M. Braschler, G. M. Di Nunzio, N. Ferro, and C. Peters. 2005. CLEF 2004: Ad Hoc Track Overview and Results Analysis. In *Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004) Revised Selected Papers*, C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini (Eds.). Lecture Notes in Computer Science (LNCS) 3491, Springer, Heidelberg, Germany, 10–26.

[53] T. Breuer, N. Ferro, N. Fuhr, M. Maistro, T. Sakai, P. Schaer, and I. Soboroff. 2020. How to Measure the Reproducibility of System-oriented IR Experiments. In *Proc. 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, Y. Chang, X. Cheng, J. Huang, Y. Lu, J. Kamps, V. Murdock, J.-R. Wen, A. Diriye, J. Guo, and O. Kurland (Eds.). ACM Press, New York, USA, 349–358.

[54] T. Breuer, N. Ferro, M. Maistro, and P. Schaer. 2021. repro_eval: A Python Interface to Reproducibility Measures of System-Oriented IR Experiments, See [198], 481–486.

[55] T. Breuer, J. Keller, and P. Schaer. 2022. ir_metadata: An Extensible Metadata Schema for IR Experiments, See [14], 3078–3089.

[56] M. Buckland and F. Gey. 1994. The relationship between Recall and Precision. *Journal of the American Society for Information Science and Technology (JASIST)* 45, 1 (January 1994), 12–19.

[57] C. Buckley. 2005. The SMART project at TREC, See [190], 301—320.

[58] C. Buckley and E. M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, E. Yannakoudakis, N. J. Belkin, M.-K. Leong, and P. Ingwersen (Eds.). ACM Press, New York, USA, 33–40.

[59] C. Buckley and E. M. Voorhees. 2004. Retrieval Evaluation with Incomplete Information. In *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, M. Sanderson, K. Järvelin, J. Allan, and P. Bruza (Eds.). ACM Press, New York, USA, 25–32.

[60] C. Buckley and E. M. Voorhees. 2005. Retrieval System Evaluation, See [190], 53–78.

[61] J. Burstein, C. Doran, and T. Solorio (Eds.). 2019. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019).* Association for Computational Linguistics, USA.

[62] L. Busin and S. Mizzaro. 2013. Axiometrics: An Axiomatic Approach to Information Retrieval Effectiveness Metrics. In *Proc. 4th International Conference on the Theory of Information Retrieval (ICTIR 2013)*, O. Kurland, D. Metzler, C. Lioma, B. Larsen, and P. Ingwersen (Eds.). ACM Press, New York, USA, 22–29.

[63] J. Callan and A. Moffat. 2012. Panel on Use of Proprietary Data. *SIGIR Forum* 46, 2 (December 2012), 10–18.
[64] N. R. Campbell. 1920. *Physics: The Elements.* Cambridge University Press, UK.
[65] N. R. Campbell. 1928. *An account of the principles of measurement and calculation.* Longmans, Green, Lonndono, UK.
[66] D. Carmel and E. Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval.* Morgan & Claypool Publishers, USA.
[67] B. A. Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation, See [263], 903–912.
[68] B. A. Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 4:1–4:34.
[69] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. 2012. Large-scale Validation and Analysis of Interleaved Search Evaluation. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 1–41.
[70] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance, See [75], 621–630.
[71] O. Chapelle and Y. Zhang. 2009. A Dynamic Bayesian Network Click Modelfor Web Search Ranking. In *Proc. 18th International Conference on World Wide Web (WWW 2009)*, J. Quemada, G. León, Y. Maarek, and W. Nejdl (Eds.). ACM Press, New York, USA, 1–10.
[72] H.-H. Chen, W.-J. Duh, H.-H. Huang, M. P. Kato, J. Mothe, and B. Poblete (Eds.). 2023. *Proc. 46th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023).* ACM Press, New York, USA.
[73] J. Chen, H. Lin, X. Han, and L. Sun. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. *arXiv.org, Computation and Language (cs.CL)* arXiv:2309.01431 (September 2023).
[74] M. Chen, Tworek. J., H. Jun, Q. Yuan, H. Pondé de Oliveira Pinto, J. Kaplan, H. Edwards, B. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, P. Petroski Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, E Herbert-Voss, W. Hebgen Guss, A. Nichol, A. Paino, A. Tezak, A. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. 2021. Training Verifiers to Solve Math Word Problems. *arXiv.org, Machine Learning (cs.LG)* arXiv:2107.03374 (July 2021).
[75] D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin (Eds.). 2009. *Proc. 18th International Conference on Information and Knowledge Management (CIKM 2009).* ACM Press, New York, USA.
[76] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/.
[77] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (Eds.). Association for Computational Linguistics, USA, 2924–2936.
[78] Z. Chu, T. Sakai, Q. Ai, and Y. Liu. 2023. Chuweb21D: A Deduped English Document Collection for Web Search Tasks. In *Proc. 1st International ACM SIGIR Conference on Information Retrieval in the Asia Pacific (SIGIR-AP 2023)*, Y. Liu, A. Moffat, Q. Ai, X. Huang, T. Sakai, and J. Zobel (Eds.). ACM Press, New York, USA.
[79] T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani (Eds.). 2008. *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008).* ACM Press, New York, USA.
[80] A. Chuklin, I. Markov, and M. de Rijke. 2015. *Click Models for Web Search.* Morgan & Claypool Publishers, USA.
[81] A. Chuklin, P. Serdyukov, and M. de Rijke. 2013. Click Model-Based Information Retrieval Metrics, See [218], 493–502.
[82] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions, See [61], 2924–2936.
[83] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv.org, Artificial Intelligence (cs.AI)* arXiv:1803.05457 (March 2018).
[84] C. L.A. Clarke, A. Vtyurina, and M. D. Smucker. 2020. Offline Evaluation without Gain. In *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, and K. Berberich (Eds.). ACM, 185–192.
[85] C. L. A. Clarke, N. Craswell, and I. Soboroff. 2004. Overview of the TREC 2004 Terabyte Track, See [409].
[86] C. L. A. Clarke, N. Craswell, and I. Soboroff. 2010. Overview of the TREC 2009 Web Track. In *The Eighteenth Text REtrieval Conference Proceedings (TREC 2009)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-278, Washington, USA.
[87] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. 2011. Overview of the TREC 2010 Web Track. In *The Nineteenth Text REtrieval Conference Proceedings (TREC 2010)*, E. M. Voorhees and L. P. Buckland (Eds.). National

Institute of Standards and Technology (NIST), Special Publication 500-294, Washington, USA.
[88] C. L. A. Clarke, . Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation, See [79], 659–666.
[89] C. L. A. Clarke, S. Rizvi, M. D. Smucker, M. Maistro, and G. Zuccon. 2021. Overview of the TREC 2020 Health Misinformation Track, See [415].
[90] C. L. A. Clarke, F. Scholer, and I. Soboroff. 2005. Overview of the TREC 2005 Terabyte Track, See [410].
[91] C. L. A. Clarke, M. D. Smucker, and A. Vtyurina. 2020. Offline Evaluation by Maximum Similarity to an Ideal Ranking, See [110], 225–234.
[92] C. W. Cleverdon. 1962. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems.* Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK.
[93] C. W. Cleverdon. 1967. The Cranfield Tests on Index Languages Devices. *Aslib Proceedings* 19, 6 (1967), 173–194.
[94] C. W. Cleverdon. 1972. On the inverse relationship of recall and precision. *Journal of Documentation* 28, 3 (1972), 195–201.
[95] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv.org, Machine Learning (cs.LG)* arXiv:2110.14168 (November 2021).
[96] K. Collins-Thompson, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. 2014. TREC 2013 Web Track Overview. In *The Twenty-Second Text REtrieval Conference Proceedings (TREC 2013)*, E. M. Voorhees (Ed.). National Institute of Standards and Technology (NIST), Special Publication 500-302, Washington, USA.
[97] K. Collins-Thompson, C. Macdonald, P. N. Bennett, and E. M. Voorhees. 2015. TREC 2014 Web Track Overview. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*, E. M. Voorhees and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-308, Washington, USA.
[98] K. Collins-Thompson, Q. Mei, B. Davison, Y. Liu, and E. Yilmaz (Eds.). 2018. *Proc. 41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018).* ACM Press, New York, USA.
[99] W. S. Cooper. 1971. A Definition of Relevance for Information Retrieval. *Inf. Storage Retr.* 7, 1 (1971), 19–37.
[100] P. C. Cozby and S. C. Bates. 2018. *Methods in Behavioral Research* (13th ed.). McGraw-Hill Education, New York, USA.
[101] M. Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics (TACL)* 6 (2018), 241–252.
[102] N. Craswell, M. Bhaskar, E. Yilmaz, D. Campos, and J. Lin. 2022. Overview of the TREC 2021 Deep Learning Track. In *The Thirtieth Text REtrieval Conference Proceedings (TREC 2021)*, I. Soboroff and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 550-335, Washington, USA.
[103] N. Craswell, M. Bhaskar, E. Yilmaz, D. Campos, J. Lin, E. M. Voorhees, and I. Soboroff. 2023. Overview of the TREC 2022 Deep Learning Track. In *The Thirty-First Text REtrieval Conference Proceedings (TREC 2022)*, I. Soboroff and A. Ellis (Eds.). National Institute of Standards and Technology (NIST), Special Publication 550-338, Washington, USA.
[104] N. Craswell and D. Hawking. 2002. Overview of the TREC-2002 Web Track. In *The Eleventh Text REtrieval Conference (TREC 2002)* , E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-251, Washington, USA.
[105] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. 2003. Overview of the TREC 2003 Web Track. In *The Twelfth Text REtrieval Conference (TREC 2003)*, E. M. Voorhees and L. P. Buckland (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-255, Washington, USA.
[106] N. Craswell, B. Mitra, D. Campos, E. Yilmaz, and E. M. Voorhees. 2020. Overview of the TREC 2019 Deep Learning Track, See [414].
[107] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos. 2021. Overview of the TREC 2020 Deep Learning Track, See [415].
[108] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. 2008. An Experimental Comparisonof Click Position-Bias Models. In *Proc. 1st ACM International Conference on Web Searching and Data Mining (WSDM 2008)*, M. Najork, A. Broder, and S. Chakrabarti (Eds.). ACM Press, New York, USA, 87–94.
[109] A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Broder, M. J. Zaki, S. Candan, A. Labrinidis, A. Schuster, and H. Wang (Eds.). 2018. *Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018).* ACM Press, New York, USA.
[110] M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux (Eds.). 2020. *Proc. 29th International Conference on Information and Knowledge Management (CIKM 2020).* ACM Press, New York, USA.
[111] B. A. Davey and H. A. Priestley. 2002. *Introduction to Lattices and Order* (2nd ed.). Cambridge University Press, Cambridge, UK.
[112] E. Davis. 2024. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Computing Surveys (CSUR)* 56, 4 (April 2024), 81:1–81:41.
[113] M. de Rijke. 2018. Retrieval as Interaction. Tony Kent Strix Annual Memorial Lecture. London, UK. Video of the lecture available at https://www.youtube.com/watch?v=Zb6YGoiPt8M.

[114] G. M. Di Nunzio and N. Ferro. 2005. DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, A. Rauber, S. Christodoulakis, and A. Min Tjoa (Eds.). Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 483–484.

[115] G. M. Di Nunzio, N. Ferro, G. J. F. Jones, and C. Peters. 2006. CLEF 2005: Ad Hoc Track Overview. In *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers*, C. Peters, F. C. Gey, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke (Eds.). Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany, 11–36.

[116] G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. 2007. CLEF 2006: Ad Hoc Track Overview. In *Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross–Language Evaluation Forum (CLEF 2006). Revised Selected Papers*, C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, and M. Stempfhuber (Eds.). Lecture Notes in Computer Science (LNCS) 4730, Springer, Heidelberg, Germany, 21–34.

[117] G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. 2008. CLEF 2007: Ad Hoc Track Overview. In *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross–Language Evaluation Forum (CLEF 2007). Revised Selected Papers*, C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, and D. Santos (Eds.). Lecture Notes in Computer Science (LNCS) 5152, Springer, Heidelberg, Germany, 13–32.

[118] F. Diaz, M. Bhaskar, M. D. Ekstrand, A. J. Biega, and B. A. Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure, See [110], 275–284.

[119] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure, See [110], 275–284.

[120] G. Dupret and M. Lalmas. 2013. Absence Time and User Engagement: Evaluating Ranking Functions, See [249], 173–182.

[121] M. Dussin and N. Ferro. 2009. Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas (Eds.). Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany, 63–74.

[122] B. Efron and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, USA.

[123] L. Egghe. 2008. The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing & Management* 44, 2 (March 2008), 856–876.

[124] C. G. Eickhoff, C. Harris and A. P. de Vries. 2012. Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments, See [196], 871–880.

[125] European Commission. 2009. Commission Regulation (EC) No 607/2009 of 14 July 2009 laying down certain detailed rules for the implementation of Council Regulation (EC) No 479/2008 as regards protected designations of origin and geographical indications, traditional terms, labelling and presentation of certain wine sector products. *Official Journal of the European Union, OJ L 193, 24.7.2009* 52 (July 2009), 60–139.

[126] European Commission. 2019. Commission Delegated Regulation (EC) No 2019/33 of 17 October 2018 supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as regards applications for protection of designations of origin, geographical indications and traditional terms in the wine sector, the objection procedure, restrictions of use, amendments to product specifications, cancellation of protection, and labelling and presentation. *Official Journal of the European Union, OJ L 9, 11.1.2019* 62 (January 2019), 2–45.

[127] A. Fabris, A. Esuli, A. Moreo, and F. Sebastiani. 2023. Measuring Fairness Under Unawareness of Sensitive Attributes: A Quantification-Based Approach. *Journal of Artificial Intelligence Research (JAIR)* 76 (April 2023), 1117–1180.

[128] A. Fabris, G. Silvello, G. A. Susto, and A. J. Biega. 2023. Pairwise Fairness in Ranking as a Dissatisfaction Measure. In *Proc. 16th ACM International Conference on Web Searching and Data Mining (WSDM 2023)*, T.-S. Chua, H. Lauw, L. Si, E. Terzi, and P. Tsaparas (Eds.). ACM Press, New York, USA, 280–288.

[129] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proc. 9th ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2023)*, M. Yoshioka, J. Kiseleva, and M. Aliannejadi (Eds.). ACM Press, New York, USA, 39–50.

[130] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. 2024. Who determines what is relevant? Humans or AI? Why not both! A Spectrum of Human-AI Collaboration in Assessing Relevance. *Communications of the ACM (CACM)* (2024).

[131] G. Faggioli and N. Ferro. 2021. System Effect Estimation by Sharding: A Comparison between ANOVA Approaches to Detect Significant Differences, See [198], 33–46.

[132] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, and F. Scholer. 2021. An Enhanced Evaluation Framework for Query Performance Prediction. In *Advances in Information Retrieval. Proc. 43rd European Conference on IR Research (ECIR 2021) – Part I*, D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani (Eds.). Lecture Notes in

Computer Science (LNCS) 12656, Springer, Heidelberg, Germany, 115–129.
[133] J. C. Falmagne and L. Narens. 1983. Scales and Meaningfulness of Quantitative Laws. *Synthese* 55, 3 (June 1983), 287–325.
[134] N. E. Fenton and J. Bieman. 2014. *Software Metrics: A Rigorous & Practical Approach* (3rd ed.). Chapman and Hall/CRC, USA.
[135] A. Ferguson, C. S. Myers, R. J. Bartlett, H. Banister, F. C. Bartlett, W. Brown, N. R. Campbell, K. J. W. Craik, J. Drever, J. Guild, R. A. Houstoun, J. O. Irwin, G. W. C. Kaye, S. J. F. Philpott, L. F. Richardson, J. H. Shaxby, T. Smith, R. H. Thouless, and W. S. Tucker. 1940. Quantitative estimates of sensory events: final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science* 2 (1940), 331–349.
[136] M. Ferrante, N. Ferro, and N. Fuhr. 2021. Towards Meaningful Statements in IR Evaluation. Mapping Evaluation Measures to Interval Scales. *IEEE Access* 9 (2021), 136182–136216.
[137] M. Ferrante, N. Ferro, and N. Fuhr. 2022. Response to Moffat's Comment on "Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales". *arXiv.org, Information Retrieval (cs.IR)* arXiv:2212.11735 (December 2022).
[138] M. Ferrante, N. Ferro, and E. Losiouk. 2019. Stochastic Relevance for Crowdsourcing, See [29], 755–762.
[139] M. Ferrante, N. Ferro, and E. Losiouk. 2020. How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal* 23, 3 (June 2020), 289–317.
[140] M. Ferrante, N. Ferro, and M. Maistro. 2014. Injecting User Models and Time into Precision via Markov Chains. In *Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2014)*, S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, and K. Järvelin (Eds.). ACM Press, New York, USA, 597–606.
[141] M. Ferrante, N. Ferro, and M. Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In *Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2015)*, J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, and Y. Zhang (Eds.). ACM Press, New York, USA, 21–30.
[142] M. Ferrante, N. Ferro, and M. Maistro. 2017. AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. *ACM Transactions on Information Systems (TOIS)* 36, 2 (September 2017), 20:1–20:38.
[143] M. Ferrante, N. Ferro, and S. Pontarollo. 2017. Are IR Evaluation Measures on an Interval Scale?, See [222], 67–74.
[144] M. Ferrante, N. Ferro, and S. Pontarollo. 2018. Modelling Randomness in Relevance Judgments and Evaluation Measures. In *Advances in Information Retrieval. Proc. 40th European Conference on IR Research (ECIR 2018)*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury (Eds.). Lecture Notes in Computer Science (LNCS) 10772, Springer, Heidelberg, Germany, 197–209.
[145] M. Ferrante, N. Ferro, and S. Pontarollo. 2019. A General Theory of IR Evaluation Measures. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31, 3 (March 2019), 409–422.
[146] M. Ferrari Dacrema, P. Cremonesi, and D. Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proc. 13th ACM Conference on Recommender Systems (RecSys 2019)*, T. Bogers, A. Said, P. Brusilovsky, and D. Tikk (Eds.). ACM Press, New York, USA, 101–109.
[147] N. Ferro. 2017. Reproducibility Challenges in Information Retrieval Evaluation. *ACM Journal of Data and Information Quality (JDIQ)* 8, 2 (February 2017), 8:1–8:4.
[148] N. Ferro, N. Fuhr, and A. Rauber. 2018. Introduction to the Special Issue on Reproducibility in Information Retrieval: Evaluation Campaigns, Collections, and Analyses. *ACM Journal of Data and Information Quality (JDIQ)* 10, 3 (October 2018), 9:1–9:4.
[149] N. Ferro, N. Fuhr, and A. Rauber. 2018. Introduction to the Special Issue on Reproducibility in Information Retrieval: Tools and Infrastructures. *ACM Journal of Data and Information Quality (JDIQ)* 10, 4 (November 2018), 14:1–14:4.
[150] N. Ferro, J. Gonzalo, J. Karlgren, and H. Müller. 2023. The CLEF Monster Track: One Lab to Rule Them All, See [274].
[151] N. Ferro and D. Kelly. 2018. SIGIR Initiative to Implement ACM Artifact Review and Badging. *SIGIR Forum* 52, 1 (June 2018), 4–10.
[152] N. Ferro, Y. Kim, and M. Sanderson. 2019. Using Collection Shards to Study Retrieval Performance Effect Sizes. *ACM Transactions on Information Systems (TOIS)* 37, 3 (May 2019), 30:1–30:40.
[153] N. Ferro and C. Peters. 2010. CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. In *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers*, C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda (Eds.). Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany, 13–35.
[154] N. Ferro and C. Peters (Eds.). 2019. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*. The Information Retrieval Series, Vol. 41. Springer International Publishing, Germany.
[155] N. Ferro and M. Sanderson. 2017. Sub-corpora Impact on System Effectiveness, See [223], 901–904.
[156] N. Ferro and M. Sanderson. 2019. Improving the Accuracy of System Performance Estimation by Using Shards, See [307], 805–814.

[157] N. Ferro and M. Sanderson. 2022. How do you Test a Test? A Multifaceted Examination of Significance Tests. In *Proc. 15th ACM International Conference on Web Searching and Data Mining (WSDM 2022)*, K. S. Candan, H. Liu, L. Akoglu, X. L. Dong, and J. Tang (Eds.). ACM Press, New York, USA, 280–288.
[158] N. Ferro and G. Silvello. 2017. 3.5K runs, 5K topics, 3M assessments and 70M measures: What trends in 10 years of Adhoc-*ish* CLEF? *Information Processing & Management* 53, 1 (January 2017), 175–202.
[159] R. A. Fisher. 1925. *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh, UK.
[160] R. A. Fisher. 1935. *The Design of Experiments.* Oliver & Boyd, Edinburgh, UK.
[161] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.
[162] J. Freire, N. Fuhr, and A. Rauber (Eds.). 2016. *Report from Dagstuhl Seminar 16041: Reproducibility of Data-Oriented Experiments in e-Science.* Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany.
[163] M. Fricke. 2009. The Knowledge Pyramid: a Critique of the DIKW Hierarchy. *Journal of Information Science* 35, 2 (2009), 131–142.
[164] M. Friedman. 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association* 32, 200 (December 1937), 675–701.
[165] M. Friedman. 1939. A Correction: The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association* 34, 205 (March 1939), 109.
[166] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, and M. Potthast. 2023. The Information Retrieval Experiment Platform, See [72], 2826–2836.
[167] N. Fuhr. 2012. Salton Award Lecture: Information Retrieval As Engineering Science. *SIGIR Forum* 46, 2 (December 2012), 19–28.
[168] N. Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (December 2017), 32–41.
[169] J. Gaito. 1959. Non-Parametric Methods in Psychological Research. *Psychological Reports* 5, 1 (March 1959), 115–125.
[170] J. Gaito. 1980. Measurement Scales and Statistics: Resurgence of an Old Misconception. *Psychological Bulletin* 87, 3 (1980), 564–567.
[171] T. Gao, H. Yen, J. Yu, and D. Chen. 2023. Enabling Large Language Models to Generate Text with Citations. *arXiv.org, Computation and Language (cs.CL)* arXiv:2305.14627 (May 2023).
[172] P. L. Gardner. 1975. Scales and Statistics. *Review of Educational Research* 45, 1 (Winter 1975), 43–57.
[173] J. D. Gibbons and S. Chakraborti. 2011. *Nonparametric Statistical Inference* (5th ed.). Chapman & Hall/CRC, Taylor and Francis Group, Boca Raton (FL), USA.
[174] E. Gibney. 2020. This AI researcher is trying to ward off a reproducibility crisis. *Nature* 577 (January 2020), 14.
[175] L. Gienapp, H. Scells, N. Deckers, J. Bevendorff, S. Wang, J. Kiesel, S. Syed, M. Fröbe, G. Zucoon, B. Stein, M. Hagen, and M. Potthast. 2023. Evaluating Generative Ad Hoc Information Retrieval. *arXiv.org, Information Retrieval (cs.IR)* arXiv:2311.04694 (November 2023).
[176] F. Giner. 2024. Information Retrieval Evaluation Measures Defined on Some Axiomatic Models of Preferences. *ACM Transactions on Information System (TOIS)* (2024), 1–34.
[177] L. Goeuriot, L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, R. Spijker, J. R. M. Palotti, and G. Zuccon. 2017. CLEF 2017 eHealth Evaluation Lab Overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eighth International Conference of the CLEF Association (CLEF 2017)*, G. J. F. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro (Eds.). Lecture Notes in Computer Science (LNCS) 10456, Springer, Heidelberg, Germany, 291–303.
[178] L. Goeuriot, H. Suominen, L. Kelly, A. Miranda-Escalada, M. Krallinger, Z. Liu, G. Pasi, G. González Sáez, M. Viviani, and C. Xu. 2020. Overview of the CLEF eHealth Evaluation Lab 2020. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)*, A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, and N. Ferro (Eds.). Lecture Notes in Computer Science (LNCS) 12260, Springer, Heidelberg, Germany, 255–271.
[179] C. Grady and M. Lease. 2010. Crowdsourcing Document Relevance Assessment with Mechanical Turk. In *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, C. Callison-Burch and M. Dredze (Eds.). The Association for Computational Linguistics (ACL), USA, 172–179.
[180] G. Grätzer. 2003. *General Lattice Theory* (2nd ed.). Birkhäuser Basel, Germany.
[181] F. Guo, C. Liu, and Y. M. Wang. 2009. Efficient Multiple-Click Models in Web Search, See [30], 124–131.
[182] Y. Guo and F. Farooq (Eds.). 2018. *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2018).* ACM Press, New York, USA.
[183] A. Halfaker, O. Keyes, D. Kluver, J. Thebault-Spieker, T. T. Nguyen, K. Shores, A. Uduwage, and M. Warncke-Wang. 2015. User Session Identification Based on Strong Regularities in Inter-activity Time. In *Proc. 24th International Conference on World Wide Web (WWW 2015)*, A. Gangemi, S. Leonardi, A. Panconesi, K. Gummadi, and C. Zhai (Eds.).

ACM Press, New York, USA, 410–418.
[184] L. Han, E. Maddalena, A. Checco, C. Sarasua, U. Gadiraju, K. Roitero, and G. Demartini. 2020. Crowd Worker Strategies in Relevance Judgment Tasks. In *Proc. 13th ACM International Conference on Web Searching and Data Mining (WSDM 2020)*, J. Caverlee, X. Hu, M. Lalmas, and W. Wang (Eds.). ACM Press, New York, USA, 241–249.
[185] L. Han, K. Roitero, U. Gadiraju, C. Sarasua, A. Checco, E. Maddalena, and G. Demartini. 2021. The Impact of Task Abandonment in Crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 33, 5 (May 2021), 2266–2279.
[186] D. J. Hand. 1996. Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159, 3 (1996), 445–492.
[187] D. K. Harman. 1994. Overview of the Third Text REtrieval Conference (TREC-3), See [188], 1–19.
[188] D. K. Harman (Ed.). 1994. *The Third Text REtrieval Conference (TREC-3)*. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA.
[189] D. K. Harman. 2011. *Information Retrieval Evaluation.* Morgan & Claypool Publishers, USA.
[190] D. K. Harman and E. M. Voorhees (Eds.). 2005. *TREC. Experiment and Evaluation in Information Retrieval.* MIT Press, Cambridge (MA), USA.
[191] C. Harris and P. Srinivasan. 2013. Using Hybrid Methods for Relevance Assessment in TREC Crowd'12, See [411].
[192] D. Hawking. 2000. Overview of the TREC-9 Web Track. In *The Ninth Text REtrieval Conference (TREC-9)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-249, Washington, USA, 87–103.
[193] D. Hawking and N. Craswell. 2001. Overview of the TREC-2001 Web Track. In *The Tenth Text REtrieval Conference (TREC 2001)* , E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-250, Washington, USA, 61–67.
[194] M. H. Heine. 1973. Distance between sets as an objective measure of retrieval effectiveness. *Information Storage and Retrieval* 9, 3 (March 1973), 181–198.
[195] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proc. 9th International Conference on Learning Representations (ICLR 2021)*, S. Mohamed, K. Hofmann, A. Oh, N. Murray, and I. Titov (Eds.). OpenReview.net, https://openreview.net/group?id=ICLR.cc/2021/Conference.
[196] W. Hersh, J. Callan, Y. Maarek, and M. Sanderson (Eds.). 2012. *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012).* ACM Press, New York, USA.
[197] T. Hey, S. Tansley, and K. Tolle (Eds.). 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, USA.
[198] D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani (Eds.). 2021. *Advances in Information Retrieval. Proc. 43rd European Conference on IR Research (ECIR 2021) – Part II.* Lecture Notes in Computer Science (LNCS) 12657, Springer, Heidelberg, Germany.
[199] Y. Hochberg and A. C. Tamhane. 1987. *Multiple Comparison Procedures.* John Wiley & Sons, USA.
[200] K. Hofmann, L. Li, and F. Radlinski. 2016. Online Evaluation for Information Retrieval. *Foundations and Trends in Information Retrieval (FnTIR)* 10, 1 (June 2016), 1–117.
[201] K. Hofmann, A. Schuth, S. Whiteson, and M. de Rijke. 2013. Reusing Historical Interaction Data for Faster Online Learning to Rank for IR, See [249], 183–192.
[202] K. Hofmann, S. Whiteson, and M. de Rijke. 2011. Balancing Exploration and Exploitation in Learning to Rank Online. In *Advances in Information Retrieval. Proc. 33rd European Conference on IR Research (ECIR 2011)*, P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Mudoch (Eds.). Lecture Notes in Computer Science (LNCS) 6611, Springer, Heidelberg, Germany, 251–263.
[203] K. Hofmann, S. Whiteson, A. Schuth, and M. de Rijke. 2014. Learning to Rank for Information Retrieval fromUser Interactions. *ACM SIGWEB Newsletter* Issue Spring (2014), 5:1–5:7.
[204] C. Holland. 2005. *Breakthrough Business Results With MVT: A Fast, Cost-Free, "Secret Weapon" for Boosting Sales, Cutting Expenses, and Improving Any Business Process.* John Wiley & Sons, New York, USA.
[205] F. Hopfgartner, A. Hanbury, H. Müller, I. Eggel, K. Balog, T. Brodt, G. V. Cormack, J. Lin, J. Kalpathy-Cramer, N. Kando, M. P. Kato, A. Krithara, T. Gollub, M. Potthast, E. Viegas, and S. Mercer. 2018. Evaluation-as-a-Service for the Computational Sciences: Overview and Outlook. *ACM Journal of Data and Information Quality (JDIQ)* 10, 4 (November 2018), 15:1–15:32.
[206] M. Hosseini, I. J. Cox, N. Milić-Frayling, G. Kazai, and V. Vinay. 2012. On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In *Advances in Information Retrieval. Proc. 32nd European Conference on IR Research (ECIR 2012)*, R. Baeza-Yaetes, A. P. de Vries, H. Zaragoza, B. B. Cambazoglu, V. Murdock, R. Lempel, and F. Silvestri (Eds.). Lecture Notes in Computer Science (LNCS) 7224, Springer, Heidelberg, Germany, 182–194.
[207] J. C. Hsu. 1996. *Multiple Comparisons. Theory and methods.* Chapman and Hall/CRC, USA.

[208] D. A. Hull. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, R. Korfhage, E. Rasmussen, and P. Willett (Eds.). ACM Press, New York, USA, 329–338.

[209] P. Ingwersen and K Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer, Heidelberg, Germany.

[210] P. G. Ipeirotis and E. Gabrilovich. 2014. Quizz: Targeted Crowdsourcing with a Billion (Potential) Users. In *Proc. 23rd International Conference on World Wide Web (WWW 2014)*, C.-W. Chung, A. Broder, K. Shim, and T. Suel (Eds.). ACM Press, New York, USA, 143–154.

[211] P. K. Ito. 1980. Robustness of ANOVA and MANOVA test procedures. In *Handbook of Statistics – Analysis of Variance*, P. R. Krishnaiah (Ed.), Vol. 1. Elsevier, The Netherlands, 199–236.

[212] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. 2007. Defining a Session on Web Search Engines. *J. Assoc. Inf. Sci. Technol.* 58, 6 (2007), 862–871. https://doi.org/10.1002/asi.20564

[213] B. J. Jansen, A. Spink, and V. Kathuria. 2006. How to Define Searching Sessions on Web Search Engines. In *Advances in Web Mining and Web Usage Analysis, 8th International Workshop on Knowledge Discovery on the Web, WebKDD 2006, Philadelphia, PA, USA, August 20, 2006, Revised Papers (Lecture Notes in Computer Science)*, O. Nasraoui, M. Spiliopoulou, J. Srivastava, B. Mobasher, and B. M. Masand (Eds.), Vol. 4811. Springer, 92–109. https://doi.org/10.1007/978-3-540-77485-3_6

[214] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (October 2002), 422–446.

[215] T. Joachims. 2002. Optimizing Search Engines using Clickthrough Data. In *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, O. Zaïane, R. Goebel, D. Hand, D. Keim, and R. Ng (Eds.). ACM Press, New York, USA, 133–142.

[216] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback, See [31], 154–161.

[217] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Transactions on Information Systems (TOIS)* 25, 2 (2007), 7.

[218] G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai (Eds.). 2013. *Proc. 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2013).* ACM Press, New York, USA.

[219] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, R. Barzilay and K. Min-Yen (Eds.). Association for Computational Linguistics, USA, 1601–1611.

[220] H. J. Jung and M. Lease. 2015. A Discriminative Approach to Predicting Assessor Accuracy. In *Advances in Information Retrieval. Proc. 37th European Conference on IR Research (ECIR 2015)*, N. Fuhr, A. Rauber, G. Kazai, and A. Hanbury (Eds.). Lecture Notes in Computer Science (LNCS) 9022, Springer, Heidelberg, Germany.

[221] E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, and J. Lin. 2023. HAGRID: A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution. *arXiv.org, Computation and Language (cs.CL)* arXiv:2307.16883 (July 2023).

[222] J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz (Eds.). 2017. *Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2017).* ACM Press, New York, USA.

[223] N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White (Eds.). 2017. *Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017).* ACM Press, New York, USA.

[224] J. Karlgren, L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, M. Sahlgren, and A. Talman. 2023. ELOQUENT CLEF shared tasks for evaluation of generative language model quality, See [274].

[225] A. Kaushik. 2006. Experimentation and Testing: A Primer. Occam's Razor. https://www.kaushik.net/avinash/experimentation-and-testing-a-primer/. Last accessed: 2022-03-23.

[226] G. Kazai, J. Kamps, M. Koolen, and N. Milić-Frayling. 2011. Crowdsourcing for Book Search Evaluation: Impact of HIT Design on Comparative System Ranking, See [263], 205–214.

[227] G. Kazai, J. Kamps, and N. Milić-Frayling. 2011. The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy. In *Proc. 20th International Conference on Information and Knowledge Management (CIKM 2011)*, I. Ounis, I. Ruthven, B. Berendt, A. P. de Vries, and F. Wenfei (Eds.). ACM Press, New York, USA, 2583–2586.

[228] J. Kekäläinen and K. Järvelin. 2002. Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science and Technology (JASIST)* 53, 13 (November 2002), 1120—1129.

[229] D. Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval (FnTIR)* 3, 1–2 (2009), 1–224.

[230] D. Kelly and J. Teevan. 2003. Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum* 37, 2 (September 2003), 18–28.
[231] L. Kelly, H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, and J. Palotti. 2019. Overview of the CLEF eHealth Evaluation Lab 2019. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)*, F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro (Eds.). Lecture Notes in Computer Science (LNCS) 11696, Springer, Heidelberg, Germany, 322–339.
[232] S. Kharazmi, F. Scholer, D. Vallet, and M. Sanderson. 2016. Examining Additivity and Weak Baselines. *ACM Transactions on Information Systems (TOIS)* 34, 4 (June 2016), 23:1–23:18.
[233] I. King, K.-T. Chen, O. Alonso, and M. Larson. 2016. Special Issue: Crowd in Intelligent Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (May 2016).
[234] R. Kohavi, T. Crook, R. Longbotham, B. Frasca, R. Henne, J. L. Ferres, and T. Melamed. 2009. Online Experimentation at Microsoft. In *Proc. of the 3rd International Workshop on Data Mining Case Studies (DMCS 2009)*, P. van der Putten, G. Melli, and B. Kitts (Eds.). ACM Press, New York, USA, 11–22.
[235] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. 2014. Seven Rules of Thumb for Web Site Experimenters. In *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2014)*, S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, and R. Ghani (Eds.). ACM Press, New York, USA, 1857–1866.
[236] R. Kohavi and R. Longbotham. 2011. Unexpected Results in Online Controlled Experiments. *SIGKDD Explorations* 12, 2 (March 2011), 31–35.
[237] R. Kohavi, D. Tang, and Y. Xu. 2020. *Trustworthy Online Controlled Experiments. A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, UK.
[238] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of Measurement. Additive and Polynomial Representations*. Vol. 1. Academic Press, New York, USA.
[239] W. H. Kruskal and W. A. Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* 47, 260 (December 1952), 583–621.
[240] T. S. Kuhn. 1996. *The Structure of Scientific Revolutions* (3rd ed.). University of Chicago Press, USA.
[241] M. Kunaver and T. Požrl. 2017. Diversity in recommender systems – A survey. *Knowledge-Based Systems* 123 (May 2017), 154–162.
[242] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin, New York, USA.
[243] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics (TACL)* 7 (2019), 452–466.
[244] M. Lalmas, H. O'Brien, and E. Yom-Tov. 2014. *Measuring User Engagement*. Morgan & Claypool Publishers.
[245] B. Larsen. 2019. The Scholarly Impact of CLEF 2010-2017, See [154], 547–554.
[246] E. Law, P. N. Bennett, and E. Horvitz. 2011. The Effects of Choice in Routing Relevance Judgments, See [263], 1127–1128.
[247] M. Lease and E. Yilmaz. 2013. Crowdsourcing for Information Retrieval: Introduction to the Special Issue. *Information Retrieval* 16, 2 (April 2013), 91–100.
[248] E. L. Lehmann. 1993. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association* 88, 424 (December 1993), 1242–1249.
[249] S. Leonardi, A. Panconesi, P. Ferragina, and A. Gionis (Eds.). 2013. *Proc. 6th ACM International Conference on Web Searching and Data Mining (WSDM 2013)*. ACM Press, New York, USA.
[250] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proc. 33rd Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.). https://proceedings.neurips.cc/paper_files/paper/2020, 9459–9474.
[251] D. Li and M. de Rijke. 2023. Extending Label Aggregation Models with a Gaussian Process to Denoise Crowdsourcing Labels, See [72], 729–738.
[252] D. Li, Z. Ren, and E. Kanoulas. 2021. CrowdGP: a Gaussian Process Model for Inferring Relevance from Crowd Annotations. In *Proc. The Web Conference 2021 (WWW 2021)*, J. Leskovec, M. Grobelnik, M. Najork, J. Tang, and Z. Leila (Eds.). ACM Press, New York, USA, 1821–1832.
[253] J. Li, S. Huffman, and A. Tokuda. 2009. Good Abandonment in Mobile and PC Internet Search, See [7], 43–50.
[254] X. Li, X. S. Wang, M. Garofalakis, I. Soboroff, T. Suel, and M. Wang (Eds.). 2014. *Proc. 23rd International Conference on Information and Knowledge Management (CIKM 2014)*. ACM Press, New York, USA.

[255] C. Lioma, J. G. Simonsen, and B. Larsen. 2017. Evaluation Measures for Relevance and Credibility in Ranked Lists, See [222], 91–98.
[256] A. Lipani, D. E. Losada, G. Zuccon, and M. Lupu. 2021. Fixed-Cost Pooling Strategies. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 33, 4 (April 2021), 1503–1522.
[257] T.-Y. Liu. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval (FnTIR)* 3, 3 (March 2009), 225–331.
[258] F. M. Lord. 1953. On the Statistical Treatment of Football Numbers. *American Psychologist* 8, 12 (1953), 750–751.
[259] D. E. Losada, J. Parapar, and A. Barreiro. 2016. Feeling Lucky? Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation. In *Proc. 2016 ACM Symposium on Applied Computing (SAC 2016)*, S. Ossowski (Ed.). ACM Press, New York, USA, 1027–1034.
[260] D. E. Losada, J. Parapar, and A. Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management* 53, 5 (September 2017), 1005–1025.
[261] R. D. Luce, D. H. Krantz, P. Suppes, and A. Tversky. 1990. *Foundations of Measurement. Representation, Axiomatization, and Invariance.* Vol. 3. Academic Press, New York, USA.
[262] R. D. Luce and J. W. Tukey. 1964. Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. *Journal of Mathematical Psychology* 1, 1 (January 1964), 1–27.
[263] W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, and W. B. Croft (Eds.). 2011. *Proc. 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011).* ACM Press, New York, USA.
[264] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, and N. Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *Proc. 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai, A. Bellogín, and M. Yoshioka (Eds.). ACM Press, New York, USA, 2429–2436.
[265] E. Maddalena and S. Mizzaro. 2014. Axiometrics: Axioms of Information Retrieval Effectiveness Metrics. In *Proc. 6th International Workshop on Evaluating Information Access (EVIA 2014)*, S. Mizzaro and R. Song (Eds.). National Institute of Informatics, Tokyo, Japan, 17–24.
[266] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)* 35, 3 (January 2017), 19:1–19:32.
[267] M. Maistro, T. Breuer, P. Schaer, and N. Ferro. 2023. An in-depth investigation on the behavior of measures to quantify reproducibility. *Information Processing & Management* 60, 3 (May 2023), 103332:1–103332:39.
[268] M. Maistro, L. C. Lima, J. G. Simonsen, and C. Lioma. 2021. Principled Multi-Aspect Evaluation Measures of Rankings. In *Proc. 30th International Conference on Information and Knowledge Management (CIKM 2021)*, G. Demartini, G. Zuccon, S. Culpepper, Z. Huang, and H. Tong (Eds.). ACM Press, New York, USA.
[269] J. Manzi. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-error for Business, Politics, and Society.* Basic Books, New York, USA.
[270] A. Marcus and A. Parameswaran. 2015. Crowdsourced Data Management: Industry and Academic Perspectives. *Foundations and Trends in Databases (FnTDB)* 6, 1–2 (December 2015), 1–161.
[271] H. M. Marcus-Roberts and F. S. Roberts. 1987. Meaningless Statistics. *Journal of Educational and Behavioral Statistics* 12, 4 (Winter 1987), 383–394.
[272] S. Maxwell and H. D. Delaney. 2004. *Designing Experiments and Analyzing Data. A Model Comparison Perspective* (2nd ed.). Lawrence Erlbaum Associates, Mahwah (NJ), USA.
[273] D. McClure. 2007. Startup Metrics for Pirates: AARRR!!! https://www.slideshare.net/dmc500hats/startup-metrics-for-pirates-long-version. Last accessed: 2022-03-23.
[274] G. McDonald, C. Macdonald, I. Ounis, N. Tonellotto, and G. Nazli (Eds.). 2023. *Advances in Information Retrieval. Proc. 46th European Conference on IR Research (ECIR 2024).* Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany.
[275] R. McGill, J. W. Tukey, and W. A. Larsen. 1978. Variations of Box Plots. *The American Statistician* 32, 1 (February 1978), 12–16.
[276] W. Mendenhall and T: Sincich. 2012. *A Second Course in Statistics. Regression Analysis* (7th ed.). Prentice Hall, USA.
[277] J. Michel. 1986. Measurement Scales and Statistics: A Clash of Paradigms. *Psychological Bulletin* 100, 3 (1986), 398–407.
[278] J. Michel. 1990. *An Introduction to the Logic of Psychological Measurement.* Lawrence Erlbaum Associates Inc., Mahwah (NJ), USA.
[279] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *arXiv.org, Computation and Language (cs.CL)* arXiv:1809.02789 (September 2018).
[280] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (March 2021), 141–163.
[281] S. Mizzaro. 1997. Relevance: The Whole History. *Journal of the American Society for Information Science and Technology (JASIST)* 48, 9 (September 1997), 810–832.

[282] A. Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *Proc. 9th Asia Information Retrieval Societies Conference (AIRS 2013)*, R. E. Banchs, F. Silvestri, T.-Y. Liu, M. Zhang, S. Gao, and J. Lang (Eds.), Vol. 8281. Lecture Notes in Computer Science (LNCS) 8281, Springer, Heidelberg, Germany, 1–12.

[283] A. Moffat. 2022. Batch Evaluation Metrics in Information Retrieval: Measures, Scales, and Meaning. *IEEE Access* 10 (2022), 105564–105577.

[284] A. Moffat. 2023. Categorical, Ratio, and Professorial Data: The Case for Reciprocal Rank. *arXiv.org, Information Retrieval (cs.IR)* (December 2023).

[285] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Transactions on Information Systems (TOIS)* 35, 3 (June 2017), 24:1–24:38.

[286] A. Moffat, J. Mackenzie, P. Thomas, and L. Azzopardi. 2022. A Flexible Framework for Offline Effectiveness Metrics, See [14], 578–587.

[287] A. Moffat, P. Thomas, and F. Scholer. 2013. Users Versus Models: What Observation Tells Us About Effectiveness Metrics. In *Proc. 22nd International Conference on Information and Knowledge Management (CIKM 2013)*, A. Iyengar, Q. He, J. Pei, R. Rastogi, and W. Nejdl (Eds.). ACM Press, New York, USA, 659–668.

[288] A. Moffat and J. Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (December 2008), 2:1–2:27.

[289] Y. Moshfeghi and A. F. Huertas-Rosero. 2022. A Game Theory Approach for Estimating Reliability of Crowdsourced Relevance Assessments. *ACM Transactions on Information Systems (TOIS)* 40, 3 (July 2022), 60:1–19:29.

[290] Y. Moshfeghi, H. F. Huertas Rosero, and J. M. Jose. 2016. A Game-Theory Approach for Effective Crowdsource-Based Relevance Assessment. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (May 2016), 55:1–55:XXX.

[291] C. Mulwa, S. Lawless, M. Sharp, and V. Wade. 2011. The Evaluation of Adaptive and Personalised Information Retrieval Systems: a Review. *Int. J. Knowl. Web Intell.* 2, 2/3 (2011), 138–156.

[292] L. Narens. 2002. *Theories of Meaningfullness.* Lawrence Erlbaum Associates, Mahwah (NJ), USA.

[293] National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science.* The National Academies Press, Washington, USA.

[294] D. Newman. 1939. The Distribution of Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. *Biometrika* 31, 2 (July 1939), 20–30.

[295] J. Neyman and E. S. Pearson. 1928. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. *Biometrika* 20A, 1/2 (July 1928), 175–240.

[296] R. M. O'Brien. 1985. The Relationship Between Ordinal Measures and Their Underlying Values: Why All the Disagreement? *Quality & Quantity* 19, 3 (June 1985), 265–277.

[297] S. Olejnik and J. Algina. 2003. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods* 8, 4 (December 2003), 434–447.

[298] H. Oosterhuis and M. de Rijke. 2021. Unifying Online and Counterfactual Learning to Rank: A Novel Counterfactual Estimator that Effectively Utilizes Online Interventions. In *Proc. 14th ACM International Conference on Web Searching and Data Mining (WSDM 2021)*, L. Lewin-Eytan, D. Carmel, E. Yom-Tov, E. Agichtein, and E. Gabrilovich (Eds.). ACM Press, New York, USA, 463–471.

[299] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (August 2015), 943–952.

[300] OpenAI. 2023. GPT-4 Technical Report. *arXiv.org, Computation and Language (cs.CL)* arXiv:2303.08774 (March 2023).

[301] D. Otero, J. Parapar, and N. Ferro. 2023. How Discriminative Are Your Qrels? How To Study the Statistical Significance of Document Adjudication Methods. In *Proc. 32nd International Conference on Information and Knowledge Management (CIKM 2023)*, I. Frommholz, F. Hopfgartner, M. Lee, M. Oakes, M. Lalmas, M. Zhang, and R. Santos (Eds.). ACM Press, New York, USA, 1960–1970.

[302] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. 2022. Training language models to follow instructions with human feedback. In *Proc. 36th Annual Conference on Neural Information Processing Systems (NeurIPS 2022)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.). https://proceedings.neurips.cc/paper_files/paper/2022.

[303] A. Overwijk, C. Xiong, and J. Callan. 2022. ClueWeb22: 10 billion web documents with rich information, See [14], 3360–3362.

[304] J. Palotti, G. Zuccon, and A. Hanbury. 2018. MM: A New Framework for Multidimensional Evaluation of Search Engines, See [109], 1699–1702.

[305] R. Perego, F. Sebastiani, J. Aslam, I. Ruthven, and J. Zobel (Eds.). 2016. *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016).* ACM Press, New York, USA.

[306] I. Pillai, I. Fumera, and F. Roli. 2013. Multi-label classification with a reject option. *Pattern Recognition* 46, 8 (August 2013), 2256–2266.

[307] B. Piwowarski, M. Chevalier, E. Gaussier, Y. Maarek, J.-Y. Nie, and F. Scholer (Eds.). 2019. *Proc. 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019).* ACM Press, New York, USA.

[308] K. Popper. 2002. *The Logic of Scientific Discovery* (2nd ed.). Routledge, Taylor & Francis Group, UK.

[309] M. Potthast, T. Gollub, M. Wiegmann, and b. Stein. 2019. TIRA Integrated Research Architecture, See [154], 123–160.

[310] W. V. Quine. 1998. *From Stimulus to Science.* Harvard University Press, Cambridge (MA), USA.

[311] F. Radlinski and T Joachims. 2006. Minimally Invasive Randomization for Collecting Unbiased Preferences from Clickthrough Logs. *CoRR* abs/cs/0605037 (2006). arXiv:cs/0605037 http://arxiv.org/abs/cs/0605037

[312] F. Radlinski, R. Kleinberg, and T. Joachims. 2008. Learning Diverse Rankings with Multi-armed Bandits. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008) (ACM International Conference Proceeding Series)*, W. W. Cohen, A. McCallum, and S. T. Roweis (Eds.), Vol. 307. ACM, 784–791. https://doi.org/10.1145/1390156.1390255

[313] F. Radlinski, M. Kurup, and T. Joachims. 2008. How Does Clickthrough Data Reflect Retrieval Quality?. In *Proc. 17th International Conference on Information and Knowledge Management (CIKM 2008)*, J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury (Eds.). ACM Press, New York, USA, 43–52.

[314] A. Rai and M. D. Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison, See [14], 726–736.

[315] T. V. Rampisela, M. Maistro, T. Ruotsalo, and C. Lioma. 2024. Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study. *ACM Transactions on Information System (TOIS)* (2024), 1–55.

[316] L. Rashidi, J. Zobel, and A. Moffat. 2023. The Impact of Judgment Variability on the Consistency of Offline Effectiveness Measures. *ACM Transactions on Information Systems (TOIS)* 42, 1 (August 2023), 19:1–19:31.

[317] H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics* (August 2023), 1–64.

[318] V. C. Raykar and S. Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research* 13 (February 2012), 491–518.

[319] V. C. Raykar, L. H. Zhao, G. Hermosillo Valadez, C. Florin, L. Bogoni, and L. Moy. 2010. Learning From Crowds. *Journal of Machine Learning Research* 11 (April 2010), 1297–1322.

[320] F. S. Roberts. 1985. Applications of the Theory of Meaningfulness to Psychology. *Journal of Mathematical Psychology* 29, 3 (September 1985), 311–332.

[321] S. Robertson. 2006. On GMAP: and Other Transformations, See [429], 78–83.

[322] S. Robertson. 2008. A New Interpretation of Average Precision, See [79], 689–690.

[323] K. Rodden, H. B. Hutchinson, and X. Fu. 2010. Measuring the User Experience on a Large Scale: User-centered Metrics for Web Applications. In *Proc. of the 28th International Conference on Human Factors and Computing Systems (CHI 2010)*, E. D. Mynatt, D. Schoner, G. Fitzpatrick, S. E. Hudson, W. K. Edwards, and T. Rodden (Eds.). ACM Press, New York, USA, 2395–2398.

[324] K. Roitero, E. Maddalena, S. Mizzaro, and F. Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Information Processing & Management* 58, 6 (November 2021), 102688:1–102688:23.

[325] K. Roteiro, A. Brunello, G. Serra, and S. Mizzaro. 2020. Effectiveness evaluation without human relevance judgments: A systematic analysis of existing methods and of their combinations. *Information Processing & Management* 57, 2 (March 2020), 102149:1–102149:20.

[326] K. Roteiro, A. Brunello, G. Serra, and S. Mizzaro. 2020. s-grams: Defining generalized n-grams for information retrieval. *Information Processing & Management* 57, 2 (March 2020), 102149:1–102149:20.

[327] B. R. Rowe, D. W. Wood, A. L. Link, and D. A. Simoni. 2010. *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program.* RTI Project Number 0211875, RTI International, USA. http://trec.nist.gov/pubs/2010.economic.impact.pdf.

[328] J. Rowley. 2007. The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science* 33, 2 (2007), 163–180.

[329] R. K. Roy. 2001. *Design of Experiments Using the Taguchi Approach: 16 Steps to Product and Process Improvement.* John Wiley & Sons, New York, USA.

[330] P. Ruiz, P. Morales-Álvarez, R. Molina, and A. K. Katsaggelos. 2019. Learning from crowds with variational Gaussian processes. *Pattern Recognition* 88 (April 2019), 298–311.

[331] A. Rutherford. 2011. *ANOVA and ANCOVA. A GLM Approach* (2nd ed.). John Wiley & Sons, New York, USA.

[332] I. Ruthven and D. Kelly (Eds.). 2011. *Interactive Information Seeking, Behaviour and Retrieval.* Facet Publishing, UK.

[333] K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi. 2021. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Communications of the ACM (CACM)* 64, 9 (September 2021), 99–106.

[334] T. Sakai. 2014. Metrics, Statistics, Tests. In *Bridging Between Information Retrieval and Databases - PROMISE Winter School 2013, Revised Tutorial Lectures*, N. Ferro (Ed.). Lecture Notes in Computer Science (LNCS) 8173, Springer, Heidelberg, Germany, 116–163.
[335] T. Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1 (June 2014), 3–12.
[336] T. Sakai. 2016. A Simple and Effective Approach to Score Standardisation. In *Proc. 2nd ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR 2016)*, B. A. Carterette, H. Fang, M. Lalmas, and J.-Y. Nie (Eds.). ACM Press, New York, USA, 95–104.
[337] T. Sakai. 2016. Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015, See [305], 5–14.
[338] T. Sakai. 2016. Topic set size design. *Information Retrieval Journal* 19, 3 (June 2016), 256–283.
[339] T. Sakai. 2018. *Laboratory Experiments in Information Retrieval*. The Information Retrieval Series, Vol. 40. Springer Singapore.
[340] T. Sakai. 2019. How to Run an Evaluation Task, See [154], 71–102.
[341] T. Sakai. 2020. On Fuhr's Guideline for IR Evaluation. *SIGIR Forum* 54, 1 (June 2020), p14:1–p14:8.
[342] T. Sakai, D. W. Oard, and N. Kando (Eds.). 2020. *Evaluating Information Retrieval and Access Tasks – NTCIR's Legacy of Research Impact*. The Information Retrieval Series, Vol. 43. Springer International Publishing, Germany.
[343] T. Sakai and R. Song. 2011. Evaluating Diversified Search Results Using Per-intent Graded Relevance, See [263], 1043–1052.
[344] T. Sakai, S. Tao, Z. Chu, M. Maistro, Y. Li, N. Chen, N. Ferro, J. Wang, I. Soboroff, and Y. Liu. 2022. Overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) Task. In *Proc. 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)*, M. P. Kato, T. Yamamoto, and Z. Dou (Eds.). National Institute of Informatics, Tokyo, Japan, 231–242.
[345] G. Salton and M. E. Lesk. 1968. Computer Evaluation of Indexing and Text Processing. *Journal of the ACM (JACM)* 15, 1 (January 1968), 8–36.
[346] M. Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval (FnTIR)* 4, 4 (2010), 247–375.
[347] M. Sanderson, M. Lestari Paramita, P. Clough, and E. Kanoulas. 2010. Do User Preferences and Evaluation Measures Line Up?. In *Proc. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, F. Crestani, S. Marchand-Maillet, E. N. Efthimiadis, and J. Savoy (Eds.). ACM Press, New York, USA, 555–562.
[348] M. Sanderson and J. Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, See [31], 162–169.
[349] T. Saracevic. 1968. *Comparative Systems Laboratory Final Technical Report, An Inquiry into Testing of Information Retrieval Systems Part II: Analysis of Results*. Technical Report. Case Western Reserve University, USA.
[350] T. Saracevic. 1975. RELEVANCE: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science and Technology (JASIST)* 26, 6 (November/December 1975), 321–343.
[351] J. Sauro and J. R. Lewis. 2016. *Quantifying the User Experience: Practical Statistics for User Research* (2nd ed.). Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.
[352] J. Savoy. 1997. Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management* 33, 44 (1997), 495–512.
[353] S. M. Scariano and J. M. Davenport. 1987. The Effects of Violations of Independence Assumptions in the One-Way ANOVA. *The American Statistician* 41, 2 (1987), 123–129.
[354] H. Scheffe. 1953. A Method for Judging all Contrasts in the Analysis of Variance. *Biometrika* 40, 1/2 (June 1953), 87–104.
[355] A. Z. Scholten and D. Borsboom. 2009. A reanalysis of Lord's statistical treatment of football numbers. *Journal of Mathematical Psychology* 53, 2 (April 2009), 69–75.
[356] A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. 2014. Multileaved Comparisons for Fast Online Evaluation, See [254], 71–80.
[357] F. Sebastiani. 2020. Evaluation measures for quantification: an axiomatic approach. *Information Retrieval Journal* 23, 3 (June 2020), 255–288.
[358] V. L. Senders. 1958. *Measurement and statistics: a basic text emphasizing behavioral science applications*. Oxford University Press, New York, USA.
[359] X. Shen, B. Tan, and C. Zhai. 2005. Context-Sensitive Information Retrieval Using Implicit Feedback, See [31], 43–50.
[360] S. Siegel. 1956. *Nonparametric Statistics: For the Behavioral Science*. McGraw-Hill, New York, USA.
[361] G. Silvello, G. Bordea, N. Ferro, P. Buitelaar, and T. Bogers. 2017. Semantic Representation and Enrichment of Information Retrieval Experimental Data. *International Journal on Digital Libraries (IJDL)* 18, 2 (June 2017), 145–172.

[362] A. Singh and T. Joachims. 2018. Fairness of Exposure in Rankings, See [182], 2219–2228.
[363] A. Singh and T. Joachims. 2018. Fairness of Exposure in Rankings, See [182], 2219–2228.
[364] A. Singhal, J. Choi, D. Hindle, and F. C. N. Pereira. 1997. AT&T at TREC-6: SDR Track. In *The Sixth Text REtrieval Conference (TREC-6)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-240, Washington, USA, 227–232.
[365] A. Slivkins. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends in Machine Learning (FnTML)* 12, 1–2 (2019), 1–286.
[366] M. D. Smucker, J. Allan, and B. A. Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proc. 16th International Conference on Information and Knowledge Management (CIKM 2007)*, M. J. Silva, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. and Falcão (Eds.). ACM Press, New York, USA, 623–632.
[367] M. D. Smucker and C. L. A. Clarke. 2012. Stochastic Simulation of Time-Biased Gain. In *Proc. 21st International Conference on Information and Knowledge Management (CIKM 2012)*, X. Chen, G. Lebanon, H. Wang, and M. J. Zaki (Eds.). ACM Press, New York, USA, 2040–2044.
[368] M. D. Smucker and C. L. A. Clarke. 2012. Time-Based Calibration of Effectiveness Measures, See [196], 95–104.
[369] I. Soboroff, S. Huang, and D. Harman. 2019. TREC 2018 News Track Overview, See [413].
[370] I. Soboroff, S. Huang, and D. Harman. 2020. TREC 2019 News Track Overview, See [414].
[371] K. Spärck Jones. 1974. Automatic indexing. *Journal of Documentation* 30, 4 (1974), 393–432.
[372] K. Spärck Jones (Ed.). 1981. *Information Retrieval Experiment*. Butterworths, London, United Kingdom.
[373] K. Spärck Jones and C. J. van Rijsbergen. 1975. Report on the need for and provision of an 'ideal' information retrieval test collection. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.
[374] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie

Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)* (2023).

[375] S. S. Stevens. 1946. On the Theory of Scales of Measurement. *Science, New Series* 103, 2684 (June 1946), 677–680.

[376] Student. 1908. The Probable Error of a Mean. *Biometrika* 6, 1 (March 1908), 1–25.

[377] H. Suominen, L. Kelly, L. Goeuriot, A. Névéol, L. Ramadier, A. Robert, E. Kanoulas, R. Spijker, L. Azzopardi, D. Li, Jimmy, J. Palotti, and G. Zuccon. 2018. Overview of the CLEF eHealth Evaluation Lab 2018. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Nineth International Conference of the CLEF Association (CLEF 2018)*, P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J.-Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro (Eds.). Lecture Notes in Computer Science (LNCS) 11018, Springer, Heidelberg, Germany, 286–301.

[378] P. Suppes, D. H. Krantz, R. D. Luce, and A. Tversky. 1989. *Foundations of Measurement. Geometrical, Threshold, and Probabilistic Representations*. Vol. 2. Academic Press, New York, USA.

[379] J. M. Tague-Sutcliffe and J. Blustein. 1994. A Statistical Analysis of the TREC-3 Data, See [188], 385–398.

[380] A. Talmor, J. Herzig, N. Lourie, and J. Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge, See [61], 4149–4158.

[381] T. Tamine-Lechani, M. Boughanem, and M. Daoud. 2010. Evaluation of Contextual Information Retrieval Effectiveness: Overview of Issues and Research. *Knowl. Inf. Syst.* 24, 1 (2010), 1–34.

[382] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang. 2019. Domain-Weighted Majority Voting for Crowdsourcing. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* 30, 1 (January 2019), 163–174.

[383] S. Tao, N. Chen, T. Sakai, Z. Chu, H. Arai, I. Soboroff, N. Ferro, and M. Maistro. 2023. Overview of the NTCIR-17 FairWeb-1 Task. In *Proc. 17th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-17)*, M. P. Kato, T. Yamamoto, and Z. Dou (Eds.). National Institute of Informatics, Tokyo, Japan.

[384] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. 2023. Alpaca: A Strong, Replicable Instruction-Following Model. https://crfm.stanford.edu/2023/03/13/alpaca.html.

[385] P. Thomas, F. Scholer, and A. Moffat. 2013. What Users Do: The Eyes Have It. In *Information Retrieval Technology – Proc. 9th Asia Information Retrieval Symposium (AIRS 2013)*, R. E. Banchs, F. Silvestri, T.-Y. Liu, M. Zhang, S. Gao, and J. Lang (Eds.). Lecture Notes in Computer Science (LNCS) 8281, Springer, Heidelberg, Germany, 416–427.

[386] P. Thomas, S. Spielman, N. Craswell, and B. Mitra. 2023. Large language models can accurately predict searcher preferences. *arXiv.org, Information Retrieval (cs.IR)* arXiv:2309.10621 (September 2023).

[387] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee. 2011. The Scholarly Impact of TRECVid (2003–2009). *Journal of the American Society for Information Science and Technology (JASIST)* 62, 4 (April 2011), 613–627.

[388] T. Tian, J. Zhu, and Y. Qiaoben. 2019. Max-Margin Majority Voting for Learning from Crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 10 (October 2019), 2480–2494.

[389] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv.org, Computation and Language (cs.CL)* arXiv:2302.13971 (February 2023).

[390] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. Singh Koura, M.-H. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv.org, Computation and Language (cs.CL)* arXiv:2307.09288 (July 2023).

[391] J. T. Townsend and F. G. Ashby. 1984. Measurement Scales and Statistics: The Misconception Misconceived. *Psychological Bulletin* 96, 2 (1984), 394–401.

[392] T. Tsikrika, A. Garcia Seco de Herrera, and H. Müller. 2011. Assessing the Scholarly Impact of ImageCLEF. In *Multilingual and Multimodal Information Access Evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011)*, P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke (Eds.). Lecture Notes in Computer Science (LNCS) 6941, Springer, Heidelberg, Germany, 95–106.

[393] T. Tsikrika, B. Larsen, H. Müller, S. Endrullis, and E. Rahm. 2013. The Scholarly Impact of CLEF (2000–2009). In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF 2013)*, P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein (Eds.). Lecture Notes in Computer Science (LNCS) 8138, Springer, Heidelberg, Germany, 1–12.

[394] J. W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (June 1949), 99–114.

[395] J. W. Tukey. 1991. The Philosophy of Multiple Comparisons. *Statistical Science* 6, 1 (February 1991), 100–116.

[396] J. Urbano, H. Lima, and A. Hanjalic. 2019. A New Perspective on Score Standardization, See [307], 1061–1064.

[397] D. van Dijk, M. Ferrante, N. Ferro, and E. Kanoulas. 2019. A Markovian Approach to Evaluate Session-based IR Systems, See [29], 621–635.

[398] C. J. van Rijsbergen. 1974. Foundations of Evaluation. *Journal of Documentation* 30, 4 (1974), 365–373.

[399] C. J. van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworths, London, England.

[400] C. J. van Rijsbergen. 1981. Retrieval effectiveness, See [372], 32–43.

[401] P. F. Velleman and L. Wilkinson. 1993. Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician* 47, 1 (February 1993), 65–72.

[402] E. M. Voorhees. 2004. Overview of the TREC 2004 Robust Track, See [409].

[403] E. M. Voorhees. 2004. Overview of the TREC 2004 Robust Track, See [409].

[404] E. M. Voorhees. 2005. Overview of the TREC 2005 Robust Retrieval Track, See [410].

[405] E. M. Voorhees. 2005. Overview of the TREC 2005 Robust Retrieval Track, See [410].

[406] E. M. Voorhees. 2009. Topic Set Size Redux, See [7], 806–807.

[407] E. M. Voorhees. 2018. On Building Fair and Reusable Test Collections using Bandit Techniques, See [109], 407–416.

[408] E. M. Voorhees. 2019. The Evolution of Cranfield, See [154], 45–69.

[409] E. M. Voorhees and L. P. Buckland (Eds.). 2004. *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*. National Institute of Standards and Technology (NIST), Special Publication 500-261, Washington, USA.

[410] E. M. Voorhees and L. P. Buckland (Eds.). 2005. *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*. National Institute of Standards and Technology (NIST), Special Publication 500-266, Washington, USA.

[411] E. M. Voorhees and L. P. Buckland (Eds.). 2013. *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*. National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA.

[412] E. M. Voorhees and C. Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. Hyon Myaeng (Eds.). ACM Press, New York, USA, 316–323.

[413] E. M. Voorhees and A. Ellis (Eds.). 2019. *The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018)*. National Institute of Standards and Technology (NIST), Special Publication 500-331, Washington, USA.

[414] E. M. Voorhees and A. Ellis (Eds.). 2020. *The Twenty-Eighth Text REtrieval Conference Proceedings (TREC 2019)*. National Institute of Standards and Technology (NIST), Special Publication 1250, Washington, USA.

[415] E. M. Voorhees and A. Ellis (Eds.). 2021. *The Twenty-Ninth Text REtrieval Conference Proceedings (TREC 2020)*. National Institute of Standards and Technology (NIST), Special Publication 1266, Washington, USA.

[416] E. M. Voorhees and D. K. Harman. 1998. Overview of the Seventh Text REtrieval Conference (TREC-7). In *The Seventh Text REtrieval Conference (TREC-7)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-242, Washington, USA, 1–24.

[417] E. M. Voorhees and D. K. Harman. 1999. Overview of the Eigth Text REtrieval Conference (TREC-8). In *The Eighth Text REtrieval Conference (TREC-8)*, E. M. Voorhees and D. K. Harman (Eds.). National Institute of Standards and Technology (NIST), Special Publication 500-246, Washington, USA, 1–24.

[418] E. M. Voorhees, D. Samarov, and I. Soboroff. 2017. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)* 36, 2 (September 2017), 12:1–12:21.

[419] W. B. Ware and J. Benson. 1975. Appropriate Statistics and Measurement Scales. *Science Education* 59, 4 (October/December 1975), 575–582.

[420] W. Webber, A. Moffat, and J. Zobel. 2008. Score Standardization for Inter-Collection Comparison of Retrieval Systems, See [79], 51–58.

[421] C. H. Weiss. 1997. *Evaluation: Methods for Studying Programs and Policies*. Prentice Hall.

[422] R. W. White. 2016. *Interactions with Search Systems*. Cambridge University Press, Cambridge, UK.

[423] R. W. White, I. Ruthven, J. M. Jose, and C. J. van Rijsbergen. 2005. Evaluating Implicit Feedback Models Using Searcher Simulations. *ACM Transactions on Information Systems (TOIS)* 23, 3 (July 2005), 325–361.

[424] F. Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (December 1945), 80–83.

[425] H. Wu, Y. Zhang, C. Ma, F. Lyu, B. He, B. Mitra, and X. Liu. 2023. Result Diversification in Search and Recommendation: A Survey. *arXiv.org, Information Retrieval (cs.IR)* arXiv:2212.14464 (July 2023).

[426] K. Yang and J. Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Pro. 29th International Conference on Scientific and Statistical Database Management (SSDBM 2017)*, A. Choudhary, K. Wu, and B. Dong (Eds.). ACM Press, New York, USA, 1–6.

[427] E. Yilmaz and J. A. Aslam. 2006. Estimating Average Precision With Incomplete and Imperfect Judgments, See [429], 102–111.

[428] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. 2014. Relevance and Effort: An analysis of Document Utility, See [254], 91–100.

[429] P. S. Yu, V. Tsotras, E. A. Fox, and C.-B. Liu (Eds.). 2006. *Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006)*. ACM Press, New York, USA.

[430] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yaetes. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proc. 26th International Conference on Information and Knowledge Management (CIKM 2017)*, E.-P. Lim, M. Winslett, J. S. Culpepper, E. Lo, J. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, and V. S. Tseng (Eds.). ACM Press, New York, USA, 1569–1578.

[431] M. Zehlike, K. Yang, and J. Stoyanovich. 2023. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Computing Surveys (CSUR)* 55, 6 (June 2023), 118:1–118:36.

[432] M. Zehlike, K. Yang, and J. Stoyanovich. 2023. Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems. *ACM Computing Surveys (CSUR)* 55, 6 (June 2023), 117:1–117:41.

[433] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, A. Korhonen, D. Traum, and L. Màrquez (Eds.). Association for Computational Linguistics, USA, 4791–4800.

[434] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv.org, Computation and Language (cs.CL)* arXiv:2304.06364 (September 2023).

[435] J. Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel (Eds.). ACM Press, New York, USA, 307–314.

[436] G. Zuccon. 2016. Understandability Biased Evaluation for Information Retrieval. In *Advances in Information Retrieval. Proc. 38th European Conference on IR Research (ECIR 2016)*, N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, and G. Silvello (Eds.). Lecture Notes in Computer Science (LNCS) 9626, Springer, Heidelberg, Germany, 280–292.