Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives

CHRISTINE BAUER, Paris Lodron University Salzburg, Austria
EVA ZANGERLE, University of Innsbruck, Austria
ALAN SAID, University of Gothenburg, Sweden

Recommender systems research and practice are fast-developing topics with growing adoption in a wide variety of information access scenarios. In this paper, we present an overview of research specifically focused on the emphevaluation of recommender systems. We perform a systematic literature review, in which we analyze 57 papers spanning six years (2017–2022). Focusing on the processes surrounding evaluation, we dial in on the methods applied, the datasets utilized, and the metrics used. Our study shows that the predominant experiment type in research on the evaluation of recommender systems is offline experimentation and that online evaluations are primarily used in combination with other experimentation methods, e.g., an offline experiment. Furthermore, we find that only a few datasets (MovieLens, Amazon review dataset) are widely used, while many datasets are used in only a few papers each. We observe a similar scenario when analyzing the employed performance metrics—a few metrics are widely used (precision, nDCG, and Recall), while many others are used in only a few papers. Overall, our review indicates that beyond-accuracy qualities are rarely assessed. Our analysis shows that the research community working on evaluation has focused on the development of evaluation in a rather narrow scope, with the majority of experiments focusing on a few metrics, datasets, and methods.

CCS Concepts: • **Information systems** → **Recommender systems**; **Evaluation of retrieval results**; • **Human-centered computing** → **HCI design and evaluation methods**.

Additional Key Words and Phrases: evaluation, survey, systematic literature review, recommender systems

1 INTRODUCTION

Recommender systems aim to alleviate choice overload by providing personalized item recommendations to users. In the development and maintenance of these systems, evaluating their performance is crucial. This work provides an overview of research specifically focused on the *evaluation* of recommender systems from 2017 to 2022. While evaluation is a significant aspect of the recommender systems field, our systematic literature review concentrates on research that specifically addresses the evaluation of recommender systems, covering papers that delve into methodological evaluation issues. This includes, for instance, papers describing research on new evaluation methods or metrics, papers analyzing how the design and implementation of the evaluation can impact the outcome of an analysis, research highlighting flaws in evaluation—or how evaluation can be improved. On the contrary, works that, for instance, propose a new recommendation model and validate it through evaluation or in other ways use evaluation to gauge the performance of a recommender system, thus, fall outside of the scope of this literature review.

The evaluation of recommender systems has been explored in previous works, but no systematic literature review has comprehensively examined datasets, metrics, or experiment types, and performed a quantitative

Authors' addresses: Christine Bauer, christine.bauer@plus.ac.at, Paris Lodron University Salzburg, Salzburg, 5020, Jakob-Haringer-Strasse 1, Austria; Eva Zangerle, eva.zangerle@uibk.ac.at, University of Innsbruck, Technikerstr. 21A, Innsbruck, 6020, Austria; Alan Said, University of Gothenburg, Sweden, alansaid@acm.org.

2770-6699/2023/10-ART
https://doi.org/10.1145/3629170

analysis of the reviewed literature. One notable study by Herlocker et al. [50] focuses on collaborative filtering systems and proposes various recommendation tasks, such as identifying good items or recommending in sequence. The work also discusses the suitability of datasets and metrics for evaluating recommendation-specific tasks prevalent during that era of recommender systems research. More recently, Gunawardana et al. [45] provide an extensive overview of the evaluation processes involved in assessing recommender systems. The study examines a wide range of properties that impact user experience and explores methods for measuring these properties, encompassing the entire evaluation pipeline from research hypotheses and experimental design to metrics for quantification. Taking a specialized approach, Pu et al. [78] presents a survey on recommender system evaluation from the users' perspective. The research particularly focuses on the initial preference elicitation process, preference refinement, and the final presentation of recommendations. From the survey results, Pu et al. [78] distills a set of usability and user interface design guidelines for user-centered evaluation of recommender systems. Beel et al. [14, 15] surveyed evaluation approaches in the field of research paper recommender systems and found that 69% of the papers featured an offline evaluation while 21% do not provide an evaluation. A survey conducted by Ihemelandu and Ekstrand [51] examines the use of statistical inference in recommender systems research and reveals that 59% of the surveyed papers did not perform significance testing. The authors argue for the inclusion of statistical inference tests in recommender systems evaluation while also acknowledging the associated challenges. More recently, Zangerle and Bauer [96] present a survey on the evaluation of recommender systems, introducing the "Framework for EValuating Recommender systems" (FEVR). This framework conceptualizes the evaluation space of recommender systems, providing a systematic overview of essential evaluation aspects and their application. The proposed FEVR framework encompasses a wide variety of facets required for evaluating recommender systems, accommodating comprehensive evaluations that address the multi-faceted dimensions found in this domain.

In addition to survey papers, several works offer critical retrospectives and analyses of evaluation procedures and setups. For example, Ferrari Dacrema et al. [40, 41] critically analyze the performance of neural recommendation approaches published from 2015 and 2018. They compare these approaches against well-tuned, non-neural baseline methods, such as nearest-neighbor or content-based approaches, and find that the simpler methods outperform 11 out of the 12 analyzed approaches. These findings suggest that limited progress has been made due to weak baselines and insufficient optimization of their parameters. Similarly, Rendle et al. [79] analyze the use of baselines in research, focusing on the MovieLens 10M and the Netflix Prize datasets. They compare the reported results of baselines with the results obtained through a re-run of the baselines, revealing substantial divergences, particularly for the MovieLens 10M dataset. They then introduce stronger and well-tuned baselines, which outperform the proposed methods. Following the same line of investigation, Ludewig et al. [66] perform a similar analysis of evaluation for session-based recommendation approaches. They compare neural sequential recommendation approaches from 2016 to 2019 with well-tuned baseline approaches, such as nearest-neighbor. Like previous works, they conclude that the claimed progress is mostly illusory, attributing it to weak baselines that are insufficiently or not at all tuned. Ludewig et al. [66] argue that this limitation is a critical drawback in current evaluation practices.

The goal of our study is to provide a quantitative snapshot of the landscape of research on the evaluation of recommender systems over the past six years. Through a systematic literature review [57] of major conferences and journals from 2017 to 2022, we analyze the evaluation methods, datasets, and metrics employed in the recommender systems community. Initially screening 339 papers, we apply defined inclusion and exclusion criteria to narrow down our review to a final sample of 57 papers. Our focus lies on three key aspects of recommender systems evaluation: (1) experiment type (offline experiments, user study, online experiment), (2) datasets, and (3) evaluation metrics.

This paper is structured as follows: In Section 2, we detail the stepwise procedure for the systematic literature review. In Section 3, we present the results of our analysis with a focus on experiment type, datasets, and

evaluation metrics. Finally (Section 4), we discuss the findings of this review and provide an outlook on future work.

2 MATERIAL AND METHODS

Our approach to identifying papers that are concerned with the evaluation of recommender systems relies on a systematic literature review [57]. A systematic literature review represents a systematic search for papers on a predefined topic and the analysis of the respective paper landscape. In this section, we outline the stepwise procedure for searching, filtering, categorizing, and analyzing the papers, which is visualized in Fig. 1 and described in detail in the following subsections.

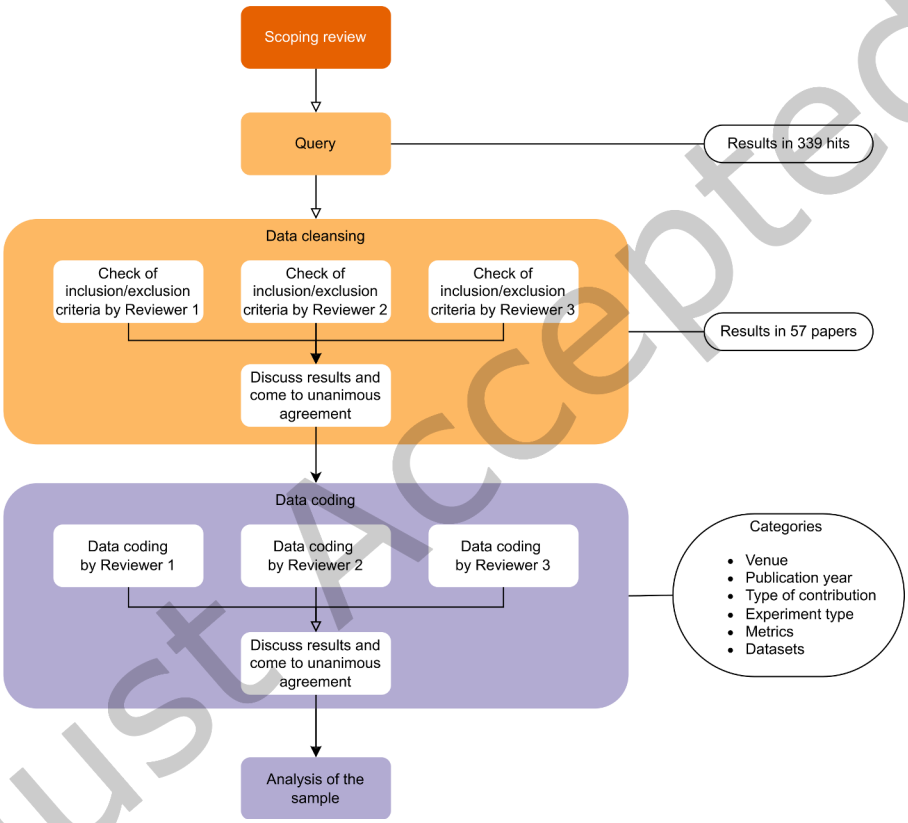

Fig. 1. Stepwise procedure for searching, filtering, categorizing, and analyzing the surveyed papers.

2.1 Literature Search

For data collection, we rely on the systematic literature review procedure as outlined in the guidelines by Kitchenham et al. [57]. To develop and pursue an effective search strategy, we performed a so-called scoping review on relevant published literature. In this scoping review, we, for instance, identified that the keyword *recommendation systems* is used interchangeably with *recommender systems*, with the latter being more common in the research community centered around the ACM Conference on Recommender Systems (RecSys), while

both alternatives are used broadly in other research outlets. Moreover, as our paper aims to cover research that revolves around methodological issues of evaluation, we identified that a search with the keywords *reproducible* or *reproducibility* has strong overlaps with a search for the keyword *evaluation*, but also yields additional hits. Similarly, using the keywords *method* or *methodology* has proven useful to identify additional works. Further, we identified that some papers were miscategorized (e.g., as a short paper instead of research paper), necessitating the use of a broader query followed by manual filtering.

The search strategy to identify eligible papers to be included in our sample consisted of several consecutive stages. As the ACM Digital Library[1] does not only contain papers published by ACM but also by other publishers, we could use this library to search for papers in the main established conferences and journals where research on recommender system evaluation is published. Besides the main conference on recommender systems–RecSys—, this embraces conferences such as SIGIR, CIKM, UMAP, and WSDM. Journals include for instance, TOIS, UMUAI, and CSUR.

Accordingly, we sampled papers that we found in the ACM Digital Library (The ACM Guide to Computing Literature), which describes as "the most comprehensive bibliographic database in existence today focused exclusively on the field of computing"[2]. For reasons of reproducibility, we consider papers in an encapsulated time frame of six years, for which we can assume that the employed databases and search engines have already completed indexing the papers from conferences and journals (*2017–2022*). As our literature review is concerned with research on the evaluation of recommender systems, we searched for papers that were indexed with the keywords *recommend** (to cover both, *recommender systems* and *recommendation systems*), and either *evalua** (to cover *evaluation* and *evaluability*) or *reproducib** (to cover *reproducible* and *reproducibility*) or *method* or *methodology*. For papers appearing in the ACM Conference on Recommender Systems, we presume that the keywords *recommender systems* or *recommendation systems* are not necessarily used; hence, for papers appearing in RecSys, we relied solely on the keywords *evalua** or *reproducib** or *method* or *methodology*. Altogether, this resulted in the following query:[3]

```
"query": {
        Keyword:(recommend*)
        AND
        Keyword:(reproducib* OR method OR methodology OR evalua*)
    OR
        ContentGroupTitle:("ACM Conference on Recommender Systems")
        AND
        Keyword:(reproducib* OR method OR methodology OR evalua*)
}
"filter": { E-Publication Date: (01/01/2017 TO 12/31/2022) }
```

This query returns a total of 339 hits (as of 10 June 2023).

We note that the query did not return any papers from the conferences CHI, CSCW, and IUI. To validate this result, for each conference separately, we searched for papers with the respective keywords without time

[1] https://dl.acm./org

[2] https://libraries.acm.org/digital-library/acm-guide-to-computing-literature

[3] https://dl.acm.org/action/doSearch?fillQuickSearch=false&target=advanced&expand=all&AfterMonth=1&AfterYear=2017&BeforeMonth=12&BeforeYear=2022&AllField=Keyword%3A%28recommend*%29+AND+Keyword%3A%28reproducib*+OR+method+OR+methodology+OR+evalua*%29+OR+ContentGroupTitle%3A%28%22ACM+Conference+on+Recommender+Systems%22%29+AND+Keyword%3A%28reproducib*+OR+method+OR+methodology+OR+evalua*%29

restriction. The latest papers on the evaluation of recommender systems at CSCW and IUI were published in 2013, and at CHI in 2016.

2.2 Data Cleansing and Selection of Papers for the Sample

We retrieved the 339 papers and reviewed them against the ex-ante-defined inclusion and exclusion criteria described below.

A paper was included if it fulfilled *each and every* of the following criteria (ex-ante inclusion criteria):

(A) The paper revolves around methodological issues of the evaluation of recommender systems.
(B) The paper is a full research paper.
(C) The paper is published within the time range from 01/01/2017 until and including 12/31/2022.

A paper was excluded if *any* of the following criteria were met (ex-ante exclusion criteria):

(a) The paper is not a research paper.
(b) The paper is a short paper, an abstract, a demo paper, a tutorial paper, or a workshop paper.[4]
(c) The paper is not concerned with recommender systems.
(d) The paper does not make a contribution regarding the evaluation of recommender systems.

Next, three reviewers independently screened the retrieved 339 papers against these inclusion and exclusion criteria by examining titles and abstracts, as well as the results and methodology sections. Any disagreement on paper selection was resolved by discussions to reach unanimous consensus among the three reviewers. These discussions resulted in the formulation of more specific inclusion criteria, further specifying the ex-ante inclusion criterion (A) that a paper is included if it "revolves around methodological issues of the evaluation of recommender systems". Hence, the ex-ante inclusion criterion (A) was considered fulfilled if *any* of the following criteria was fulfilled (ex-post inclusion criteria):

(A.1) The paper provides a literature survey on the evaluation of recommender systems.
(A.2) The paper introduces one or more novel metrics of evaluation.
(A.3) The paper analyzes metrics of evaluation.
(A.4) The paper contributes an extensive critical evaluation across a set of approaches.
(A.5) The paper contributes a conceptual framework for evaluation.
(A.6) The paper contributes a framework for evaluation in the form of a toolkit.
(A.7) The paper contributes a novel evaluation model; e.g., related to off-policy learning.
(A.8) The paper proposes a novel sampling approach for (offline) evaluation.
(A.9) The paper contributes to evaluation by analyzing sampling approaches.
(A.10) The paper demonstrates or discusses how the results inform the evaluation of recommender systems.

Further, the ex-ante inclusion criterion (A) was *not* considered fulfilled if *any* of the following criteria was fulfilled (ex-post exclusion criteria):[5]

(A.i) The paper proposes a recommendation model with or without validating it through evaluation but does not contribute to methodological issues of evaluation.
(A.ii) The paper presents an exploratory evaluation of a recommender system but does not contribute to methodological issues of evaluation.
(A.iii) The paper presents an experiment but does not contribute to methodological issues of evaluation.
(A.iv) The paper analyses recommendation approaches but does not contribute to methodological issues of evaluation.

[4]We note that we did not consider the search criterion *research paper* in the query because essential full papers were not returned by the query due to miscategorization as a short paper in the database (e.g., [10]).
[5]Note, these are also a further specification of the ex-ante exclusion criterion (d).

(A.v) The paper studies psychological effects influencing the design and development of recommender systems.

This data cleansing and selection procedure led to the exclusion of 282 papers (see Appendix). The remaining 57 papers make up our final sample resulting from the query. Table 1 provides an overview of all papers in the sample.

Table 1. Surveyed papers, sorted by venue (alphabetically) and year.

Papers	Venues	Year
Saraswat et al. [84]	AIML Systems	2021
Jannach [52]	ARTR	2023
Eftimov et al. [38]	BDR	2021
Sonboli et al. [88], Zhu et al. [99]	CIKM	2021
Ekstrand [39]	CIKM	2020
Alhijawi et al. [5], Sánchez and Bellogín [83], Zangerle and Bauer [96]	CSUR	2022
Jin et al. [54]	HAI	2021
Belavadi et al. [16]	HCII	2021
Peska and Vojtas [77]	HT	2020
Ostendorff et al. [75]	ICADL	2021
Afolabi and Toivanen [2]	IJEHMC	2020
Bellogín et al. [17]	IRJ	2017
Latifi et al. [62]	ISCI	2022
Carraro and Bridge [23]	JIIS	2022
Krichene and Rendle [60], Li et al. [63], McInerney et al. [69]	KDD	2020
Dehghani Champiri et al. [36]	KIS	2019
Latifi and Jannach [61]	RecSys	2022
Dallmann et al. [35], Narita et al. [73], Parapar and Radlinski [76], Saito et al. [82]	RecSys	2021
Cañamares and Castells [22], Kouki et al. [59], Sun et al. [90], Symeonidis et al. [91]	RecSys	2020
Ferrari Dacrema et al. [41]	RecSys	2019
Yang et al. [95]	RecSys	2018
Xin et al. [94]	RecSys	2017
Ali et al. [6]	Scientometrics	2021
Diaz and Ferraro [37], Silva et al. [87]	SIGIR	2022
Anelli et al. [10], Li et al. [64], Lu et al. [65]	SIGIR	2021
Balog and Radlinski [11], Mena-Maldonado et al. [70]	SIGIR	2020
Cañamares and Castells [21]	SIGIR	2018
Cañamares and Castells [20]	SIGIR	2017
Chen et al. [25]	TheWebConf	2019
Al Jurdi et al. [4]	TKDD	2021
Guo et al. [47]	TOCHI	2022
Zhao et al. [98]	TOIS	2022
Ferrari Dacrema et al. [40], Mena-Maldonado et al. [71]	TOIS	2021
Anelli et al. [9]	UMAP	2022
Frumerman et al. [42]	UMAP	2019
Bellogín and Said [19]	UMUAI	2021
Said and Bellogín [80]	UMUAI	2018
Chin et al. [26], Kiyohara et al. [58]	WSDM	2022
Cotta et al. [31]	WSDM	2019
Gilotte et al. [44]	WSDM	2018

2.3 Review of the Selected Papers in Full Text (Coding)

For each paper, we obtained meta-information on the paper from the citation information, i.e., author, year, title, type of venue—conference or journal—and venue name. In addition, to address the main purpose of this paper, we extracted the following information from the full text: experiment type, used dataset(s), used metric(s), and type of contribution. To this end, three reviewers examined the full text of the papers and extracted the respective information. Concerning datasets and metrics, the respective information was extracted directly from the full text of the papers. Concerning the experiment type, we relied on the established differentiation

between offline experiment, user study, and online experiment [96]: Offline evaluation refers to a computational evaluation without human subjects being involved in the evaluation process; user studies refer to evaluations (in live or laboratory settings) with a set of human participants that carry out tasks as defined by the researcher; and online evaluations refer to field experiments where users carry out their self-selected tasks in a real-world setting. For the type of contribution, the categorization scheme was developed inductively from raw data. The categorization scheme allowed each paper to belong to exclusively one type of contribution. An overview of the types is presented in Table 2; the specified types are benchmark, framework, metrics, model, and survey respectively. The initial inter-rater reliability was at an acceptable level (Krippendorff's $\alpha = 0.8214$). Disagreement was resolved by discussions to reach unanimous consensus (Krippendorff's $\alpha = 1$).

Table 2. The five types used to describe the type of contribution made in the reviewed literature.

Types of Contribution	Description
Benchmark	Providing an extensive critical evaluation across a (wide) set of approaches or datasets
Framework	Introducing a framework for evaluation, which may take the form of a toolkit or a conceptual framework
Metrics	Analyzing existing or introducing novel metrics of evaluation
Model	Introducing a novel recommendation or evaluation model
Survey	A literature survey

In all phases of extracting and categorizing data, all authors were engaged. Where disagreement emerged in rare cases, the authors discussed the categorization in question, drawing upon domain expertise on a case-by-case basis, until unanimous consensus was established.

3 RESULTS

In this section, we first give a general overview of papers on the evaluation of recommender systems in the analyzed time frame 2017–2022 (Section 3.1). Then, we detail the types of contributions to the discourse (Section 3.2). Further, we provide an overview of the experiment types used in the papers (Section 3.3). Section 3.4 provides an overview and discussion of the datasets used. In Section 3.5, we detail the metrics used and discussed in the papers.

3.1 General Overview

Most papers on evaluation in recommender systems are published at RecSys—the main conference concerning the research topic *recommender systems*—(12) and at SIGIR (9)— the main conference concerning the closely related research topic of *information retrieval*—(Fig. 2). Notably, as can be seen from Fig. 2, papers on the evaluation of recommender systems are published in a wide scale of venues (12 conference venues and 13 journal venues) where it is often only one paper at the respective venue in the set time frame of our review. The majority of papers on evaluation are published at conferences (39 papers) compared to 18 papers published in journals. Further, from Fig. 2, we see that there is a clear concentration across conference venues (RecSys and SIGIR), whereas papers on evaluation are particularly scattered across journal venues.

Concerning the temporal evolution of evaluation papers, we observe an increasing number of papers on the evaluation of recommender systems in the analyzed time frame 2017–2022 (Fig. 3). Starting in 2017, there were only 3 papers on the evaluation of recommender systems published, while this number peaked in 2021 with 19 papers on that topic. While there is a continuous upward trend of papers on that topic in conference

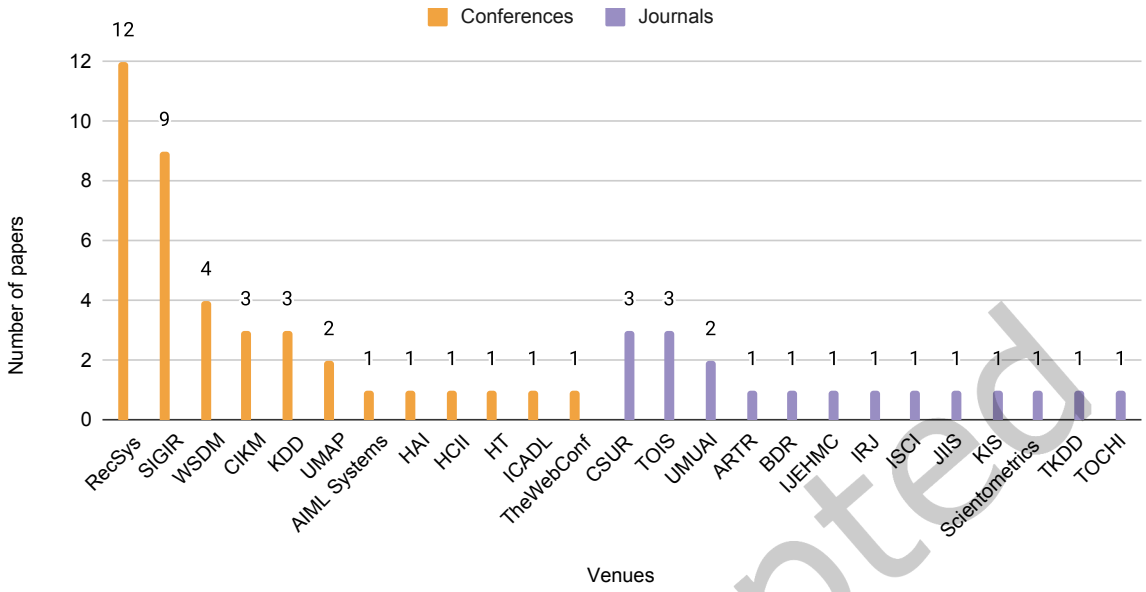

Fig. 2. Number of papers per venue, sorted by venue type (journals vs. conferences) and number of papers.

venues, there is a sharp increase of papers on that topic in journal venues (only one journal paper in the years 2017–2020, respectively; then 6 and 8 journal papers in 2021 and 2022, respectively). We note that two of the journal papers published in 2021 (Ferrari Dacrema et al. [40] and Mena-Maldonado et al. [71]) are extended versions of previously published conference papers (Ferrari Dacrema et al. [41] from 2019 and Mena-Maldonado et al. [70] 2020 respectively). Further, the increase of journal papers on evaluation in the years 2020 and 2021 aligns with the COVID-19 pandemic, during which all conferences were either canceled or held online; which points to having led researchers to focus on journal submissions instead of conferences.

3.2 Type of contribution

This section provides a detailed overview of the types of papers included in the literature review. The types as specified in Table 2 (i.e., benchmark, framework, metrics, model, and survey) were inferred according to the description in Section 2.3.

Fig. 4 provides an overview of the number of papers per type of contribution in our sample. Most of the papers in our sample contribute to models (19); these papers provide a conceptual and empirical basis for improved recommendation or evaluation models. Considerably fewer papers (13) investigate metrics. Nine papers provide a survey, another 9 papers provide benchmarks of various approaches and 7 papers propose frameworks.

Among the model papers, the majority focus on evaluation models, specifically on issues related to off-policy learning [23, 31, 44, 58, 69, 73, 82, 95], which helps to obtain unbiased estimates for improved offline evaluation [55]. Cañamares and Castells [20] propose a probabilistic reformulation of memory-based collaborative filtering. While the core contribution of that work is a recommendation model, it also contributes to evaluation because the experiments demonstrate that performance measurements may heavily depend on statistical properties of the input data, which the authors discuss in detail. With a probabilistic analysis, Cañamares and Castells [21] address the question of whether popularity is an effective or misleading signal in recommendation. Their work illustrates the contradictions between the accuracy that would be measured in common biased offline experimental settings

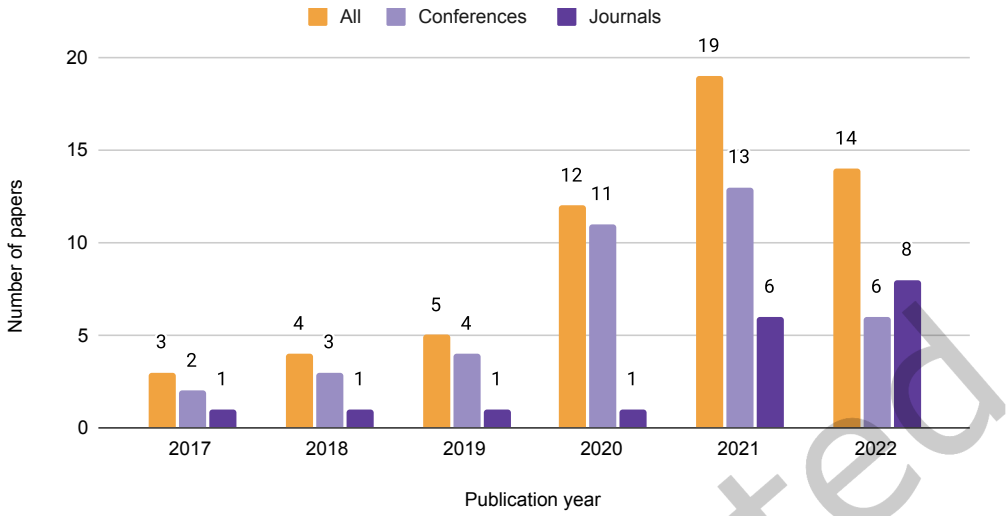

Fig. 3. Number of papers per year.

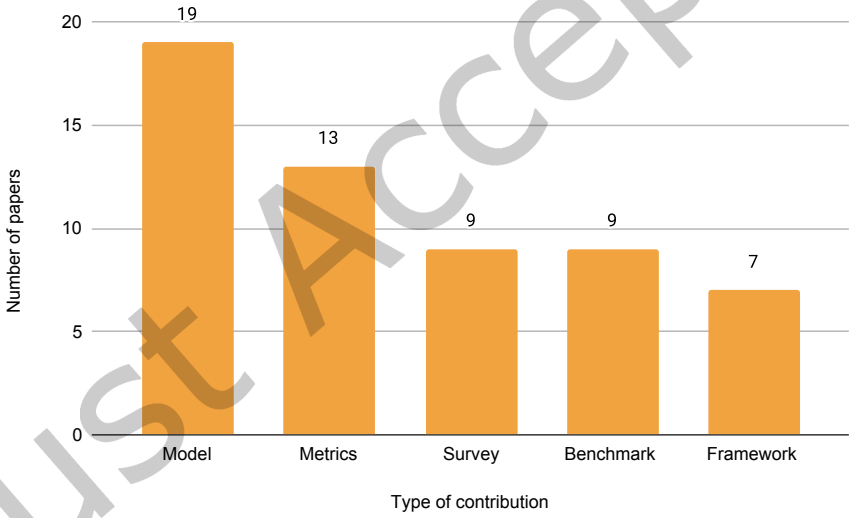

Fig. 4. Number of papers per type of contribution.

and the measured with unbiased observations. Cañamares and Castells [22] demonstrate the importance of item sampling in offline experiments. Based on a thorough literature review, Carraro and Bridge [23] propose a new sampling approach to debiasing offline experiments. A second line of model papers considers user-related aspects as an important ingredient of recommender systems. For example, Frumerman et al. [42] investigate the meaning of "rejected" recommendations in a more fine-grained manner. Symeonidis et al. [91] consider short-term intentions to inform models. Jin et al. [54] rely on a psychometric modeling method to study the key qualities of conversational recommender systems. In a large-scale user study, Chen et al. [25] investigate how serendipity

improves user satisfaction with recommendations; their results inform the modeling for recommendations. Ostendorff et al. [75] study users' preferences for link-based versus text-based recommendations using qualitative evaluation methods. Lu et al. [65] investigate whether and how annotations made by external assessors (thus, not the recommender system's users) are a viable source for preference labeling. Guo et al. [47] study order effects in recommendation sequences, which has implications for the design of recommender systems. Said and Bellogín [80] evaluate and model inconsistencies in user rating behavior to improve the performance of recommendation methods. These papers considering user-related aspects have in common that each work primarily studies phenomena to improve recommendation models and the discussion of the results also contributes to methodological issues regarding the evaluation of recommender systems.

Among papers focusing on metrics, one set of papers compares metrics (e.g., [70, 71, 77]), whereas some papers focus their analysis on a specific type of metrics; for instance, sampling metrics (e.g., [60, 63]) and folding metrics (e.g., [94]). In a similar spirit, Bellogín et al. [17] study biases in information retrieval metrics. Another line of metrics papers aims for harmonization of metrics (e.g., [2, 76]) or metric improvements (e.g., [64]). Balog and Radlinski [11] propose how to measure the quality of explanations in recommender systems. Saraswat et al. [84] propose combining both performance and user satisfaction metrics in offline evaluation, leading to improved correlation with desired business metrics. Finally, Diaz and Ferraro [37] makes a metrics analysis and discussion leading into the proposal of an altogether metric-free evaluation method.

Papers discussing infrastructural aspects of recommender systems can be categorized into two types of framework papers: Those that contribute with a recommendation toolkit and those proposing a conceptual framework. The presented toolkits are iRec [87], Elliot [10], LensKit [39], and librec-auto [88].[6] The framework by Bellogín and Said [19] provides guidelines for reproducibility; their paper also provides an in-depth analysis to support their guidelines. Eftimov et al. [38] propose a general framework that fuses different evaluation measures and aims at helping users to rank systems. Considering users' expectations and perceptions, Belavadi et al. [16] study the relationships between several user evaluation criteria.

Several papers provide an extensive critical evaluation across a (wide) set of approaches (Table 3). Dallmann et al. [35] study sampling strategies for sequential item recommendation. They compare 4 methods across 5 datasets and find that both sampling strategies—uniform random sampling and sampling by popularity—can produce inconsistent rankings compared with the full ranking of the models. Ferrari Dacrema et al. [41] and its extended version Ferrari Dacrema et al. [40] perform a reproducibility study, critically analyzing the performance of 12 neural recommendation approaches in comparison to well-tuned, established, non-neural baseline methods. Their work identifies several methodological issues and finds that 11 out of the 12 analyzed approaches are outperformed by far simpler, yet well-tuned, methods (e.g., nearest-neighbor or content-based approaches). In a similar vein, Latifi and Jannach [61] perform a reproducibility study where they benchmark Graph Neural Networks (GNN) against an effective session-based nearest neighbor method. Also, this work finds that the conceptually simpler method outperforms the GNN-based method. Anelli et al. [9] perform a reproducibility study, systematically comparing 10 collaborative filtering algorithms (including approaches based on nearest-neighbors, matrix factorization, linear models, and techniques based on deep learning). Different to Ferrari Dacrema et al. [40, 41], Anelli et al. [9] benchmark all algorithms using the very same datasets (MovieLens-1M [48], Amazon Digital Music [74], and epinions [92]) and the identical evaluation protocol. Based on their study on modest-sized datasets, they conclude—similar to other works—that the latest models are often not the best-performing ones. Kouki et al. [59] compare 14 models (8 baseline and 6 deep learning) for session-based recommendations using 8 different popular evaluation metrics. After an offline evaluation, they selected the 5 algorithms that performed the best and ran a second round of evaluation using human experts (user study). [90] provides benchmarks

[6]Note that the work by Sun et al. [90]—besides providing benchmarks across several datasets, recommendation approaches, and metrics—also proposes the toolkit daisyRec.

across several datasets, recommendation approaches, and metrics; beyond that, this work introduces the toolkit daisyRec. Zhu et al. [99] compare 24 models for click-through rate (CTR) prediction on multiple dataset settings. Their evaluation framework for CTR (including the benchmarking tools, evaluation protocols, and experimental settings) is publicly available. Latifi et al. [62] focus on sequential recommendation problems, for which they compare the Transformer-based BERT4Rec method [89] to nearest-neighbor methods, showing that the nearest-neighbor methods achieve comparable performance to BERT4Rec for the smaller datasets, whereas BERT4Rec outperforms the simple methods when the datasets are larger.

Table 3. Benchmark papers.

Papers	**Details**
Anelli et al. [9]	Reproducibility study. An in-depth, systematic, and reproducible comparison of 10 collaborative filtering algorithms (including approaches based on nearest-neighbors, matrix factorization, linear models, and techniques based on deep learning) using 3 datasets and the identical evaluation protocol. Provide a guide for future research with respect to baselines and systematic evaluation.
Dallmann et al. [35]	Study sampling strategies for sequential item recommendation. Compare 4 methods across 5 datasets and find that both, uniform random sampling and sampling by popularity, can produce inconsistent rankings compared with the full ranking of the models.
Ferrari Dacrema et al. [40, 41]	Reproducibility study. Critical analysis of the performance of 12 neural recommendation approaches with reproducible setup. Comparison against well-tuned, established, non-neural baseline methods. Identification of several methodological issues, including choice of baselines, propagation of weak baselines, and a lack of proper tuning of baselines.
Kouki et al. [59]	Compare 14 models (8 baseline and 6 deep learning) for session-based recommendations using 8 different popular evaluation metrics.
Latifi and Jannach [61]	Reproducibility study. Benchmark Graph Neural Networks against an effective session-based nearest neighbor method. The conceptually simpler method outperforms the GNN-based method both in terms of Hit Ratio and the MRR.
Latifi et al. [62]	Compare the Transformer-based BERT4Rec method [89] to nearest-neighbor methods for sequential recommendation problems across 4 datasets using exact and sampled metrics. The nearest-neighbor methods achieve comparable or better performance than BERT4Rec for the smaller datasets, whereas BERT4Rec outperforms the simple methods for the larger ones.
Sun et al. [90]	Benchmarks across several datasets, recommendation approaches, and metrics; in addition, it introduces the toolkit daisyRec.
Zhu et al. [99]	Open benchmarking for click-through rate prediction with a rigorous comparison of 24 existing models on multiple dataset settings in a reproducible manner. The evaluation framework for CTR (including the benchmarking tools, evaluation protocols, and experimental settings) are publicly available.

Table 4 provides an overview of survey papers on the evaluation of recommender systems. Some of the papers provide an extensive critical evaluation across a (wide) set of datasets and approaches on a specialized topic (e.g., [26, 40, 41, 59, 61]). Others provide a (systematic) review of the literature landscape on a specialized topic

(e.g., [4–6, 36, 52, 83, 98]). The framework by Zangerle and Bauer [96] is based on a survey of previous literature on the respective topic. Similarly, Zhao et al. [98] starts with a survey of literature on aspects related to offline evaluation for top-N recommendation, which builds the basis for their systematic comparison of a selected set of 12 algorithms across 8 datasets.

Table 4. Survey papers on the evaluation of recommender systems.

Papers	Details
Al Jurdi et al. [4]	Classification of natural noise management (NNM) techniques and analysis of their strengths and weaknesses. Comparative statistical analysis of the NNM mechanisms.
Alhijawi et al. [5]	Specifically address the objectives: relevance, diversity, novelty, coverage, and serendipity. Reviews the definitions and measures associated with these objectives. Classifies over 100 articles (published from 2015 to 2020) regarding objective-oriented evaluation measures and methodologies. Collect 43 objective-oriented evaluation measures.
Ali et al. [6]	Survey on the evaluation of scholarly recommender systems. Analysis suggests that there is a focus on offline experiments, whereby either simple/trivial baselines are used or no baselines at all.
Chin et al. [26]	Compare 45 datasets used for implicit feedback based top-k recommendation based on characteristics (similarities and differences) and usage patterns across papers. For 15 datasets, they evaluate and compare the performance of 5 different recommendation algorithms.
Dehghani Champiri et al. [36]	Focus on context-aware scholarly recommender systems. Classification evaluation methods and metrics on usage.
Jannach [52]	Provide an overview of evaluation aspects as reported in 127 papers on conversational recommender systems. Argue for a mixed methods approach, combining objective (computational) and subjective (perception-oriented) techniques for the evaluation of conversational recommenders, because these are complex multi-component applications, consisting of multiple machine learning models and a natural language user interface.
Sánchez and Bellogín [83]	Focus on point-of-interest recommender systems. Systematic review covering 10 years of research on that topic, categorizing the algorithms and evaluation methodologies used. The common problems are that both, the algorithms and the used datasets (statistics), are described in insufficient detail.
Zangerle and Bauer [96]	Introduce "Framework for EValuating Recommender systems", derived from the discourse on recommender systems evaluation. Categorization of the evaluation space of recommender systems evaluation. Emphasis on the required multi-facettedness of a comprehensive evaluation of a recommender system.
Zhao et al. [98]	Survey of 93 offline evaluation for top-N recommendation algorithms. Provide an overview of aspects related to evaluation metrics, dataset construction, and model optimization. In addition, this work presents a systematic comparison of 12 top-N recommendation algorithms (covering both traditional and neural-based algorithms) across 8 datasets.

3.3 Experiment Types

While many types of experiments can be performed, the results presented in this section rely on the established definitions of online, offline, and user study respectively.

As shown in Fig. 5, the vast majority of the papers (38) use offline experiments. Considerably fewer papers (12) report user studies. Comparably few (6) report on online experiments. Ten papers do not report any evaluation, these are mainly survey papers [4–6, 36, 52, 83], papers on metrics [6, 60, 94], and one paper contributing with a framework [88].

While most papers (39) employ one experiment type, there are 7 papers that combine two types, and one paper [59] combining all three types (Table 5). Interestingly, all papers using an online experiment, combine it with another experiment type; four papers using an online experiment [44, 73, 77, 91], also carry out offline experiments, one combines online experiments with user studies [16], and one paper combines all three experiment types [59]. Further two papers [42, 80] use offline experiments and user studies.

3.4 Datasets

Table 6 provides an overview of the datasets used in the papers. In total, our analysis contains 80 datasets. We distinguish between papers that use pre-collected, established datasets (65 datasets) and papers that propose a custom dataset (15 datasets, see the last row of Table 6). In a graphical overview, Fig. 6 presents the number of papers relying on each dataset. Note that in this chart, we have aggregated different versions of a dataset into a

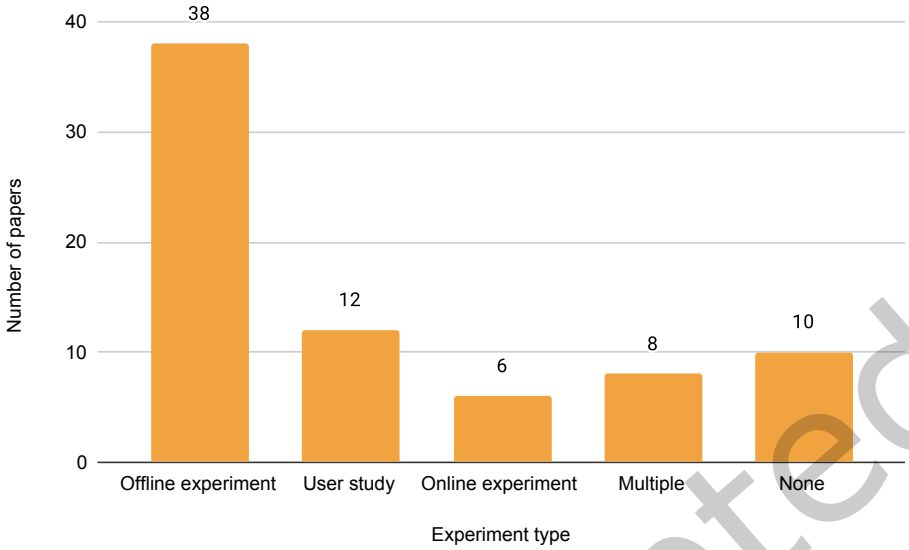

Fig. 5. Number of papers per experiment type.

Table 5. Papers using more than one experiment type.

Papers	Online experiment	Offline experiment	User study
Gilotte et al. [44]	x	x	
Narita et al. [73]	x	x	
Peska and Vojtas [77]	x	x	
Symeonidis et al. [91]	x	x	
Frumerman et al. [42]		x	x
Said and Bellogín [80]		x	x
Belavadi et al. [16]	x		x
Kouki et al. [59]	x	x	x

single dataset category (for instance, we combined the widely used MovieLens datasets MovieLens 100k, 1M, 10M, 20M, 25M, Latest, and HetRec).

Table 6 and Fig. 6 show that the dataset usage distribution for established (pre-collected) datasets is dominated by the MovieLens datasets. MovieLens datasets are used 32 times in the papers investigated, with MovieLens 1M being the most popular dataset (19 usages). Furthermore, the Amazon review datasets are used in 24 papers, followed by the LastFM dataset, appearing in the evaluation of 9 papers. We also observe that 43 and hence, 66.15% of the listed datasets are only used in a single paper. Further 8 datasets are used in two of the papers in our study and another 14 datasets are employed in three or more papers.

Generally, the majority of papers relied on existing, pre-collected datasets: out of 146 dataset usages, 15 were custom datasets. These findings are in line with a previous analysis of datasets being used for recommender systems evaluation [13], with a focus on the use of data pruning methods for the years 2017 and 2018. Generally, the high number of datasets employed at a low rate makes a direct comparison of recommendation approaches

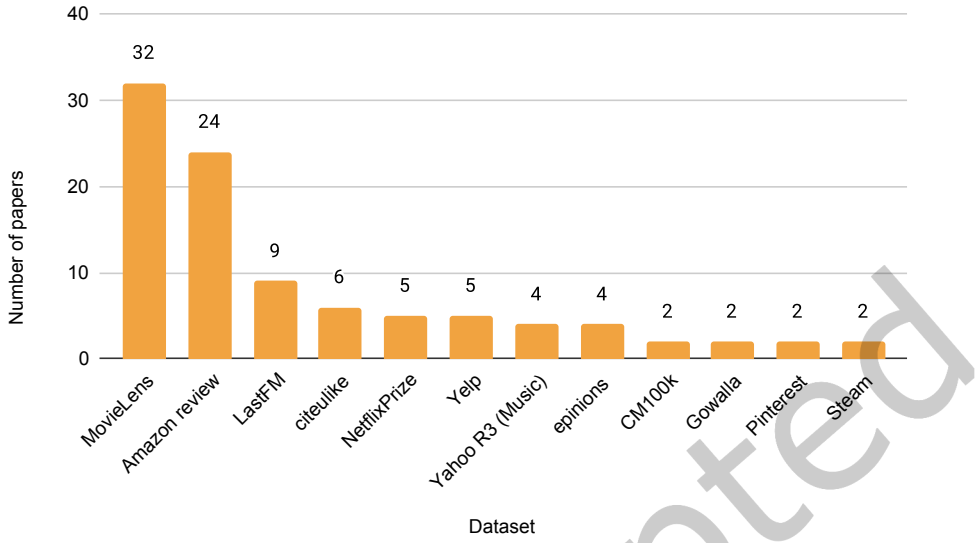

Fig. 6. Overview of datasets used in at least two papers, where different versions of a dataset are aggregated into a single dataset category for the Amazon review, MovieLens, and citeulike datasets.

hardly possible. Particularly, given the vastly different characteristics of these. In contrast, we also observe that established datasets like the MovieLens dataset family, are used frequently, allowing for a better comparison of approaches.

Table 6. Overview of datasets used in surveyed papers.

Datasets	Papers	# Papers
Amazon Beauty [74]	[26, 35, 62]	3
Amazon Book [74]	[95]	1
Amazon Digital Music [74]	[9, 26]	2
Amazon Electronics [74]	[26, 90, 98]	3
Amazon Home & Kitchen [74]	[64]	1
Amazon Instant Video [74]	[41]	1
Amazon Kindle Store [74]	[87]	1
Amazon Movies & TV [74]	[26, 40, 98]	3
Amazon Musical Instruments [74]	[26, 40]	2
Amazon Patio, Lawn & Garden [74]	[26]	1
Amazon Sports & Outdoors [74]	[64]	1
Amazon Toys & Games [74]	[26, 98]	2
Amazon Video Games [74]	[26, 35, 98]	3
Avazu[7]	[99]	1
BeerAdvocate [68]	[37]	1
Book crossing [100]	[90]	1
citeulike-a [93]	[40, 41, 95]	3
citeULike-t [93]	[26, 40, 64]	3
Clothing Fit [72]	[87]	1
CM100k [21]	[70, 71]	2
CoatShopping [86]	[23]	1
Criteo[8]	[99]	1

[7]https://www.kaggle.com/c/avazu-ctr-prediction
[8]https://www.kaggle.com/c/criteo-display-ad-challenge

Table 6. Overview of datasets used in surveyed papers.

Datasets	Papers	# Papers
epinions [92]	[9, 40, 64, 90]	4
Filmtrust [46]	[40]	1
Flixster[9]	[26]	1
Frappe [12]	[40]	1
Good Books [10]	[87]	1
Good Reads [11]	[87]	1
Gowalla [27]	[40, 61]	2
LastFM [24]	[17, 19, 20, 26, 40, 61, 87, 90, 98]	9
Library-Thing [97]	[37]	1
Million Playlist Dataset[12]	[38]	1
Million Post Corpus [85]	[16]	1
MovieLens 100k [48]	[26, 40, 41]	3
MovieLens 1M [48]	[9, 10, 17, 19, 20, 22, 35, 37, 40, 41, 60, 62, 63, 70, 71, 80, 87, 90, 98]	19
MovieLens 10M [48]	[26, 94]	2
MovieLens 20M [48]	[26, 35, 40, 62, 76]	5
MovieLens 25M [48]	[84]	1
MovieLens Latest [48]	[65]	1
MovieLens HetRec[13]	[40]	1
MoviePilot[14]	[80]	1
NetflixPrize[15]	[20, 40, 41, 87, 98]	5
Open Bandit [81]	[82]	1
Pinterest [43]	[40, 41]	2
Steam [56]	[35, 62]	2
Ta Feng Grocery Dataset[16]	[40]	1
Tradesy [49]	[95]	1
TREC Common Core 2017 [7][17]	[37]	1
TREC Common Core 2018 [18]	[37]	1
TREC Deep Learning Document Ranking 2019 [32]	[37]	1
TREC Deep Learning Document Ranking 2020 [32]	[37]	1
TREC Deep Learning Passage Ranking 2019 [32]	[37]	1
TREC Deep Learning Passage Ranking 2020 [33]	[37]	1
TREC Robust 2004[19]	[37]	1
TREC Web 2009 [28]	[37]	1
TREC Web 2010[20]	[37]	1
TREC Web 2011[21]	[37]	1
TREC Web 2012 [29]	[37]	1
TREC Web 2013[22]	[37]	1
TREC Web 2014 [30]	[37]	1
Webscope R3 [67]	[23]	1
Yelp[23]	[19, 40, 80, 90, 98]	5
Yahoo R3 (Music)[24]	[22, 70, 71, 87]	4

[9]https://sites.google.com/view/mohsenjamali/home
[10]https://github.com/zygmuntz/goodbooks-10k
[11]https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks
[12]https://research.atspotify.com/datasets/
[13]https://grouplens.org/datasets/hetrec-2011/
[14]http://www.moviepilot.de/
[15]https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data
[16]https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset
[17]https://github.com/trec-core/2017
[18]https://github.com/trec-core/2018
[19]https://trec.nist.gov/data/t13_robust.html
[20]https://trec.nist.gov/data/web10.html
[21]https://trec.nist.gov/data/web2011.html
[22]https://github.com/trec-web/trec-web-2013
[23]https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset
[24]https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=3

Table 6. Overview of datasets used in surveyed papers.

Datasets	Papers	# Papers
Yahoo R4[25]	[26]	1
Xing [1]	[42]	1
Custom	[2, 11, 21, 25, 31, 44, 47, 54, 58, 59, 69, 73, 75, 77, 91]	15

A further aspect to consider regarding the comparability of approaches is dataset pre-processing. Typical pre-processing steps include removing users, items, or sessions with a low number of interactions or converting explicit ratings to binary relevance values. As Ferrari Dacrema et al. [40] note in their survey on the reproducibility of deep learning recommendation approaches, it is important that all pre-processing steps are clearly stated in the paper and that the removal of data is justified and motivated. Also, pre-processing should be included in the code published. Inspecting the papers of our survey, we find that eight papers mention that they convert explicit rating data to a binary relevance score or song play counts to explicit ratings [17, 23, 26, 37, 38, 62, 64, 90]. Furthermore, users, items or sessions with fewer and/or more interactions than a given threshold are removed in twelve papers [9, 22, 26, 35, 42, 61, 62, 64, 77, 90, 91, 98]. Zhao et al. [98] refer to this pre-processing step as n-core filtering. They perform a study on three aspects in the context of evaluating recommender systems: evaluation metrics, dataset construction, and model optimization. For dataset construction, they find that 44% of the papers in their study do not provide any information about pre-processing, and 34% of the papers apply n-core filtering with n set to 5 or 10. Sun et al. [90] also study the impact of different thresholds for filtering users and items. Here it is important to note that, for instance, the MovieLens datasets are already pre-processed to some extent as they only include users with more than or equal to 20 interactions.

In the following, we focus our analysis on datasets that have been used at least three times in the surveyed papers. Table 7 provides an overview of these twelve datasets, where we list the domain, the feedback type (hence, whether the dataset features explicit or implicit data; in the case of explicit ratings, we also add the rating scale), the size of the dataset captured by the number of interactions, and the type of side information contained. Notably, five out of the twelve most popular datasets stem from the movie or music domain. In terms of the type of ratings contained, the citeulike and LastFM datasets provide implicit feedback (0 or 1), while the other datasets provide explicit ratings on a scale from 0 (or 1) to 5 stars. Interestingly, when inspecting the size of the datasets, the most popular datasets appear to be relatively small, with the most popular dataset (MovieLens 1M) holding 1, 000, 000 interactions.

Another interesting aspect when investigating the choice of datasets for the evaluation of recommender systems is the number of different datasets used by individual papers. Evaluating a recommender system on diverse datasets is critical to gaining insights into the generalizability and robustness of the recommender system proposed. When inspecting the number of different datasets used in the experiments, we find that 26 papers (45.61% of all papers contained in the study) rely on a single dataset, five papers (8.77%) rely on two datasets, seven papers (12.28%) use 3 datasets and another ten papers (17.54%) use four or more datasets. Out of these, three papers used more than ten different datasets: In extensive experiments, Ferrari Dacrema et al. [41] benchmark deep learning-based recommender systems against a set of relatively simple baselines. Diaz and Ferraro [37] showcase a metric-free evaluation method for recommendation and retrieval based on a set of 16 datasets. Chin et al. [26] conduct an empirical study on the impact of datasets on the evaluation outcome and resulting conclusions. Their study shows a different distribution of dataset popularity among recommender systems evaluation than we observe in the analysis at hand. However, we conjecture that this is due to the diverse inclusion criteria of the studies. For instance, Chin et al.'s study is restricted to implicit feedback top-k recommendation tasks. Notably,

[25]https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=4

Table 7. Details of datasets used in at least three papers. We list the domain of the dataset, the type of feedback, number of interactions contained, and side information provided.

Datasets	Domains	Feedback	Interactions	Side Information
Amazon Electronics, Products, Video Games [74]	Products	[1,5]	20,994,353 (E), 371,345 (B), 2,565,349 (V)	product information (e.g., description, color, product images, technical details), timestamp
citeulike-a, citeulike-t	Scientific Papers	{0,1}	204,987 (a), 134,860 (t)	tags, bag-of-words, and raw text for each article, citations between articles
epinions [92]	Products	[0,5]	922,267	explicit trust relationships among users, timestamps
LastFM [24]	Music	{0,1}	19,150,868	artist, song name, timestamp
MovieLens (100k, 1M, 20M) [48]	Movies	[0,5]	100,000 (100k)–20,000,000 (20M)	movie metadata (e.g., title, genre), user metadata (e.g., age, gender), rating timestamp
NetflixPrize[26]	Movies	[1,5]	100,000,000	movie metadata (title, release year), rating date
Yelp [27]	Business	[0,5]	6,990,280	business metadata (address, category, etc.), user metadata (user name, user stats (no. of reviews, user votes, etc.)), rating timestamp

our analysis also contains nine papers (15.79%) that did not use any dataset. The reason here is that most of these papers are surveys [4–6, 36, 52, 83, 96]. Furthermore, Ekstrand [39] describes the Python LensKit software framework and Sonboli et al. [88] describe the librec-auto toolkit.

Our analysis contains 13 versions of the Amazon review datasets, seven different versions (or subsets) of the MovieLens dataset, and two versions of the citeulike dataset. Considering the usage of different versions of the same dataset, we find that five papers use different versions of the same aggregated dataset. In their survey on dataset usage, Chin et al. [26] use eight versions of the Amazon reviews dataset and three versions of the MovieLens dataset (out of a total of 15 individual datasets used). In their reproducibility study, Ferrari Dacrema et al. [40] used four versions of the MovieLens datasets, both versions of the citeulike datasets, and two versions of the Amazon reviews dataset (out of 17 individual datasets used). In their prior reproducibility study, Ferrari Dacrema et al. [41] used two versions of the MovieLens dataset.

We further investigate which datasets are jointly used in evaluations. For this analysis, analyze the sets of datasets co-used in the papers (note that the co-usage of individual datasets is already presented in Table 6). We employed a frequent itemset approach (i.e., the Apriori algorithm [3]) and present the results in Table 8. This table shows the set of datasets employed together and the number of papers that co-use these datasets. The most frequently combined datasets are LastFM and MovieLens 1M (appearing in seven papers). The MovieLens 1M dataset appears in pairs with the NetflixPrize and the Yelp datasets in five papers. In the list of sets of datasets that appear in four papers, we find not only pairs but also triples of datasets that are jointly used for evaluation in three papers. Unsurprisingly, the MovieLens datasets and other popular datasets are dominant. This aspect has also been raised by Chin et al. [26] and our results are in line with these previous findings.

Table 8. Combinations of datasets (pairs and triples) frequently co-occurring in experiments. We list all sets of datasets that co-occur in at least 3 papers (ML = MovieLens).

Dataset Combinations	# Papers
{LastFM, ML 1M}	7
{ML 1M, NetflixPrize}, {ML 1M, Yelp}	5
{ML 1M, Yahoo R3}, {LastFM, Yelp}, {LastFM, NetflixPrize}, {LastFM, ML 1M, NetflixPrize}, {LastFM, ML 1M, Yelp}	4
{Amazon Movies & TV, LastFM}, {Amazon Electronics, LastFM}, {Amazon Beauty, ML 20M}, {epinions, ML 1M}, {ML 100k, ML 20M}, {ML 100k, ML 1M}, {ML 1M, ML20M}	3

Inspecting the papers that use custom datasets, we observe that the majority of these papers feature (or create) a custom dataset for three distinctive reasons. One reason is user surveys [2, 25] and user studies

Table 9. Overview of the metrics used in surveyed papers.

Metrics	Abbr.	Papers	#
Area Under Curve	AUC	[25, 35, 38, 60, 77]	5
Average Coverage of Long Tail	ACLT	[9]	1
Average Percentage of Long Tail	APLT	[9]	1
Average Precision	AP	[37, 60, 64, 95]	4
Average Recommendation Popularity	ARP	[9]	1
Binary Preference-based measure	bpref	[17]	1
Clickthrough rate	CTR	[77, 84, 91, 99]	4
Conversion rate	CVR	[31]	1
Coverage (item)	Coverage	[38, 59, 98]	3
Coverage (user)		[87]	1
Discounted Cumulative Gain	DCG	[95]	1
Expected Free Discovery	EFD	[9]	1
Expected Popularity Complement	EPC	[9, 87]	2
Expected Profile Distance	EPD	[87]	1
F-measure	F1	[9]	1
Fallout		[71]	1
Gini		[9, 87]	2
Hit Rate	HR	[35, 38, 40, 59, 61, 62, 90, 98]	8
Hits		[87]	1
Intra-list Diversity	ILD	[87]	1
Inferred Average Precision	InfAP	[17]	1
Item Coverage	IC	[9]	1
Jaccard coefficient		[65]	1
Logistic Loss	Logloss	[99]	1
Mean Absolute Error	MAE	[95]	1
Mean Average Precision	MAP	[9, 23, 37, 40, 59, 77, 90, 98]	8
Mean Reciprocal Rank	MRR	[9, 40, 59, 61, 62, 77, 90, 98]	8
Mean Squared Error	MSE	[58, 73]	2
normalized Discounted Cumulative Gain	nDCG	[9, 17, 19–23, 26, 35, 37, 40, 41, 59, 60, 62, 64, 76, 77, 90, 98]	20
Novelty		[98]	1
Overlap		[65]	1
Pearson Correlation Coefficient	PCC	[65]	1
Popularity		[59]	1
Popularity-based Ranking-based Equal Opportunity	PREO	[9]	1
Popularity-based Ranking-based Statistical Parity	PRSP	[9]	1
Precision	P	[9, 17, 19–23, 38, 40–42, 44, 59, 64, 65, 70, 71, 77, 87, 90, 91, 98]	22
Recall	R	[9, 19, 22, 23, 26, 37, 40, 41, 59, 60, 63, 65, 77, 87, 90, 95, 98]	17
Reciprocal Rank	RR	[37, 64]	2
Root Mean Squared Error	RMSE	[65, 69, 73, 80, 94]	5
Custom		[2, 11, 25, 37–39, 54, 75, 81, 94]	12
Total number of metrics: 40			

being conducted [11, 47, 54, 75], where the result of the user study itself is presented as a novel dataset. For instance, Chen et al. [25] perform a user study to get a deeper understanding of the impact of serendipity on user satisfaction on a popular mobile e-commerce platform in China. A further reason for using custom datasets is the recent trend towards counterfactual (off-policy) learning, which requires an unbiased, missing-at-random dataset [22, 31, 44, 58, 73]. Furthermore, several papers perform evaluations based on proprietary data provided by a private sector business entity [44, 59, 69, 73, 77, 91].

3.5 Metrics

The reviewed literature features an extensive range of datasets, as depicted in Section 3.4. This variety is also mirrored in the selection of evaluation metrics. We divide the metrics into two categories: conventional metrics widely utilized in the field, and specific metrics proposed for the unique problem addressed within a certain paper. We refer to these as custom metrics (see the final row of Table 9). A visual representation of the most frequently used metrics—those employed in at least two papers within our surveyed literature—is provided in Fig. 7.

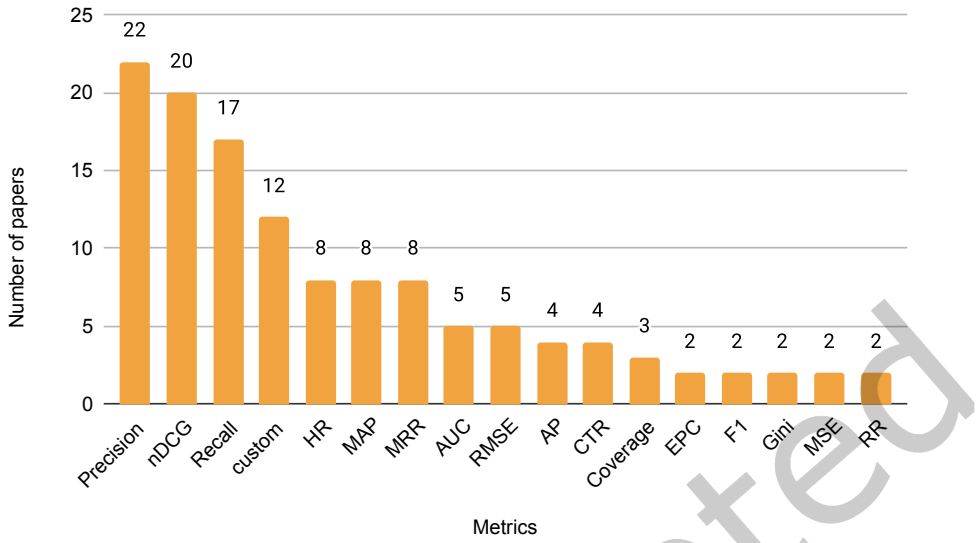

Fig. 7. Overview of metrics used in at least two papers (NB: Coverage refers to item coverage).

Traditionally, recommender systems research has relied on a standard set of metrics, including Precision, Recall, and normalized Discounted Cumulative Gain (nDCG) [18, 45]. These metrics have gained significant popularity in the examined literature. However, our analysis also uncovers the existence of a diverse array of less prevalent metrics, as illustrated in Table 9. In essence, a selected group of metrics is featured prominently: Precision is employed in 22 out of the 57 reviewed papers (approximately 36%), nDCG in 20 papers (around 35%), and Recall in 17 papers (nearly 30%). These findings resonate with the notion that ranking and relevance metrics align more closely with actual user preferences than a minimized rating prediction error does [34, 45]. Yet, metrics associated with rating prediction, such as RMSE, MAE, and MSE, still figure prominently in a considerable portion of the reviewed literature, appearing in a total of seven papers (about 12%). While a vast majority of papers do not employ rating prediction metrics, the fact that more than one in ten papers uses them contradicts the general consensus in the recommender systems research field, which holds that rating prediction is an inadequate surrogate for actual user preference [8].

Fig. 7 portrays the disparity in popularity among various metrics. Precision, nDCG, and Recall are roughly twice as favored as any of the other top metrics. These three metrics epitomize the core characteristics of recommender and information retrieval systems, notably relevance and ranking.

Furthermore, it is worth mentioning that out of the total 40 metrics employed in the reviewed papers, 23 metrics (approximately 58%) are each applied in just a single paper. Some of these uniquely applied metrics are specific to individual papers that utilize an extensive range of metrics. For example, Silva et al. [87] introduce metrics such as user-coverage, EPC, EPD, Gini, and Hits, while Anelli et al. [9] introduce various non-accuracy metrics like Average Coverage of Long Tail, Average Percentage of Long Tail, Expected Free Discovery, and Popularity-based Ranking-based Equal Opportunity, among others. Moreover, five metrics appear in only two papers each, and a single metric is utilized in three papers. The variation in metric usage complicates the comparison and benchmarking across different papers, as emphasized in the discussion on dataset usage (see Section 3.4).

Similarly, we scrutinize the number of metrics utilized per paper. It is crucial to emphasize that the quantity of metrics employed does not necessarily reflect the quality or completeness of a paper or recommender system. Nonetheless, the use of multiple metrics can yield insights into different facets of a system. When analyzing our

Table 10. The categories of value the metrics express.

Categories	Metrics
Relevance	AP, AUC, F1, fallout, Hits, HR, InfAP, Logloss, MAP, P, R
Success Rate	CTR, CVR
Rating Prediction Accuracy	bpref, MAE, MSE, RMSE
Ranking	DCG, nDCG, MRR, RR
Non-accuracy	ACLT, APLT, Coverage, EFD, EPC, EPD, Gini, IC, ILD, Jaccard, Overlap, PCC, Popularity, PREO, PRSP

data, we discover that 18 papers (32%) use only a single metric, and surprisingly, ten papers (18%) do not use any metrics whatsoever. Although the majority of papers that abstain from using metrics are categorized as literature reviews (refer to Table 4), there are exceptions. Furthermore, nine papers (16%) apply two metrics, while 5 papers (9%) employ three metrics. In total, 42 papers (74%) utilize three or fewer metrics. With this understanding, we now probe into the variety of metrics. In Table 10, we present a classification of evaluation metrics into overarching categories that correspond to specific recommendation tasks, like ranking, rating prediction, and relevance. Despite the absence of a universally accepted classification of metrics in the recommender systems research field, our categorization resonates with the general application scenarios of recommendations and the desired attributes of a recommender system.

Table 11. Combinations of metrics used frequently in the surveyed papers. Tuples with asterisks contain metrics from at least two of the categories in Table 10, excluding custom metrics. (NB: Coverage in refers to item coverage)

Metric combinations	# Papers
{nDCG, P}*	14
{nDCG, R}*	13
{P, R}	12
{nDCG, P, R}*	10
{nDCG, MAP}*, {R, MAP}, {nDCG, R, MAP}*	8
{nDCG, P, MAP}*, {P, MAP}, {nDCG, P, R, MAP}*, {nDCG, MRR}, {P, R, MAP}	7
{nDCG, MAP, MRR, R}*, {MRR, P, MAP, R}*, {nDCG, MRR, MAP}*, {MRR, MAP, R}*, {MRR, P, MAP}*, {nDCG, P, MRR, MAP}*, {MRR, P}*, {MRR, R}*, {MRR, MAP}*, {nDCG, HR}*, {nDCG, P, MRR, R}*, {MRR, HR}*, {MRR, P, R}*, {nDCG, P, MRR}*, {nDCG, MRR, R}*, {nDCG, MAP, MRR, P, R}*	6
{P, HR}, {nDCG, HR, MRR}*	5
{nDCG, P, HR, MAP}*, {P, HR, R, MAP}, {nDCG, HR, R}*, {nDCG, HR, R, MAP}*, {MRR, P, HR, R}*, {nDCG, P, HR, MRR}*, {nDCG, HR, MRR, R}*, {nDCG, P, HR, R}*, {MRR, MAP, HR, R}*, {MAP, MRR, P, HR, R}*, {nDCG, MRR, P, HR, MAP}*, {nDCG, MAP, MRR, HR, R}*, {nDCG, MAP, P, HR, R}*, {nDCG, MRR, P, HR, R}*, {nDCG, HR, MRR, MAP}*, {nDCG, MRR, P, HR, R, MAP}*, {MRR, P, HR, MAP}*, {nDCG, P, HR}*, {P, HR, R}, {MRR, HR, R}*, {MRR, P, HR}*, {nDCG, HR, MAP}*, {HR, R, MAP}, {P, HR, MAP}, {MRR, HR, MAP}*, {HR, R}, {HR, MAP}	4
{Coverage, HR}*, {P, AUC}, {AUC, R}, {nDCG, AUC}*, {P, Coverage, HR}*, {P, Coverage}*, {nDCG, AUC, R}*, {nDCG, AP}*, {AP, R}	3

In the context of metrics, it is interesting to explore the combinations of metric types, that is, the characteristics being measured in tandem. Given that recommendations apply across diverse contexts, the extensive array of metrics used mirrors the various goals pursued by recommendation applications and the stakeholders involved. By concentrating on metrics adopted in three or more papers, we examine the employed combinations in the surveyed literature (refer to Table 11). A key observation from this table is that the majority of combinations encompass ranking and relevance metrics, while combinations incorporating other metric types are less prevalent. This observation contrasts with current discussions in the recommender systems community, with the only beyond-accuracy metric appearing in the table being item coverage. This indicates that beyond-accuracy metrics are seldom used in combination with other metrics, including other beyond-accuracy metrics such as novelty, fairness, or any of the metrics in the bottom row of Table 10. A similar comment can be made regarding the utilization of success rate metrics.

Additionally, in agreement with the discourse within the recommender systems community, particularly regarding rating prediction, it is worth mentioning that no rating prediction error metrics are present in this table. This could signal a decrease in the overall usage of these metrics. Even when acknowledging that some papers use these metrics (as noted above), they do so without merging them with the more widely accepted evaluation tools and metrics.

4 DISCUSSION

With this survey paper, we aim to provide an analysis of a snapshot of research on the evaluation of recommender systems. We gain insights into the type of experiments the community performs when researching on evaluation aspects, the data it focuses on, and the metrics that are seen as important.

First, we find that, within research on evaluation aspects of recommender systems, there is a strong focus on offline experiments, a result that is in line with what has been shown in earlier overviews of recommender systems research in general, e.g., [6, 53]. We observe that several papers combine two types of experiments; this is seen as contributing to getting a more comprehensive picture than when using one experiment type only (see, e.g., Zangerle and Bauer [96]). However, with 8 of 57 papers that employ such a multi-method approach, the number of papers taking this approach is low.[28] Interestingly, when investigating the use of online experiments, we find that online experiments are predominantly combined with another experiment—typically with an offline experiment. Overall, this indicates that the landscape of research on the evaluation of recommender systems is a narrow one, with a strong focus on offline experiments, at least in published literature. As our review concentrates on research that specifically focuses on the evaluation of recommender systems, it does not allow for drawing conclusions concerning evaluation practices of the recommender systems research at large. Still, suppose that the broader landscape of recommender systems research embraces the full spectrum of experiment types (i.e., online experiments, user studies, offline experiments), then research on the evaluation of recommender systems needs to reflect the broad spectrum too. In case the broader landscape of recommender systems research has a strong focus on offline evaluations (as, for instance, shown in Jannach [52] and Jannach and Bauer [53]), the community is encouraged to embrace the wider spectrum in their evaluation efforts. For the specialized topic of conversational recommender systems, Jannach [52] provides a good rationale for why it is essential to involve humans in the evaluation process of such systems (thus, encouraging to use user studies and online experiments). With their FEVR framework, Zangerle and Bauer [96] provide guidance concerning the multifaceted aspects that need to be considered in a comprehensive evaluation (thus, encouraging to use the full spectrum of experiment types). In the realm of research that specifically focuses on the evaluation of recommender systems, it appears worthwhile to embrace the full spectrum and possibly demonstrate how the results of different experiment types

[28] Note that 10 papers in our sample (for instance, several survey papers) do not use any experiment type.

may diverge or complement each other. In this regard, we want to point to Kouki et al. [59], which is the only work covered by our survey that embraces all three experiment types.

Second, we observe a popularity gap in the use of datasets. On the one hand, the same few (and relatively old) datasets (i.e., MovieLens, Amazon review dataset) are used extensively; on the other hand, as many as 50% of the datasets (32) are used in only one single paper each. While the use of the same (or similar) datasets across multiple papers can increase comparability and benchmarking, in many cases it is disputable whether those few datasets indeed represent the best choice. First, older datasets are typically significantly smaller than newer, or current, datasets. This, in turn, raises questions regarding generalizability and applicability in the current landscape but also points to a lack of validation concerning the scalability of the evaluated recommendation models and approaches to larger datasets. Second, we have to be aware that older datasets may not be good proxies of the user behavior and preferences of today's users. As a result, good performance results with outdated datasets may not work sufficiently well in current practice. Third, with the focus on MovieLens and Amazon reviews, it is difficult to assess whether, and how, the evaluation results generalize to other domains. Yet, while the newly-created datasets may better reflect these issues, these do not allow for comparison because of their one-time use. Against this background, we encourage the community to use more recent datasets and—where feasible—demonstrate generalizability by including datasets from multiple domains. To facilitate reproducibility, researchers are strongly encouraged to make datasets publicly available.

Third, when analyzing the employed performance metrics, we observe a similar picture as for dataset usage: only a few metrics are widely used, i.e., Precision, nDCG, and Recall. There are a number of metrics that are, comparatively, rarely used in experiments validating the performance of recommendation approaches. Interestingly, next to Precision, nDCG, and Recall, a large number of papers (22) introduce specific custom metrics. These custom metrics make it difficult, if not impossible, to compare recommendation quality across, and even within, papers. The observation of the (still) high popularity of error metrics (used in 8 papers, 13%) goes against the general consensus in the recommender systems research field that these are poor proxies to assess recommender performance related to actual user preferences. Further, our review indicates that beyond-accuracy metrics are rarely used in research on the evaluation of recommender systems, which is not aligned with the discourse in the recommender systems field that evaluation concerning beyond-accuracy qualities are crucial. We note that our review surveys papers that focus on the evaluation of recommender systems; thus, while the consideration of beyond-accuracy metrics is also essential for papers with a focus on evaluation, this observation does not allow to draw conclusions about the use of beyond-accuracy metrics in recommender systems research practice in general. However, other surveys that cover evaluation practice in recommender systems show a similar picture: for instance, the recent review by Alhijawi et al. [5], drawing a sample from works published from 2015 to 2020, found that the main objective of all reviewed papers was to generate relevant recommendations, whereas other objectives did not get the same attention as relevance (only 21.3% of the reviewed works considered diversity, 6.1% coverage, 3.4% serendipity, and 6.1% novelty) and, in the recent survey on offline evaluation for top-N recommendation algorithms by Zhao et al. [98], only two of 93 papers (2.15%) used beyond-accuracy metrics. In short, the community is encouraged to use appropriate metrics and, particularly, include beyond-accuracy metrics in their evaluation efforts, as both are essential for both, research on the evaluation of recommender systems and also for research on recommender systems at large.

Our literature review comes with certain limitations. In our search strategy, we relied on the paper keywords provided by the authors. This may have caused relevant papers contributing to evaluation being excluded from our datasets because these were not tagged with the keywords used in our query. For example, we observe that some papers do not put the evaluation of recommender systems at the core of the investigation, but—in addition—also contribute to evaluation. For instance, the core contribution of Cañamares and Castells [20] is a recommendation model. In addition, their work demonstrates that the performance measurements may heavily depend on the statistical properties of the input data sample, which is a significant contribution to evaluation

and is also discussed accordingly in the paper. Other papers with a core contribution outside the evaluation field might not use the keyword "evaluation" and our query might have missed those. However, a query using only the keywords "recommender systems" or "recommendation systems" to an enormous number of papers (1, 698 hits as of 19 July 2023) for the time frame 2017-2022, which was not reasonable to process manually for this review. Moreover, we note that our review provides a snapshot of research on the evaluation of recommender systems in the limited time frame of 2017–2022. Accordingly, this review does not allow for deriving conclusions about how the evaluation practices have evolved over (longer) time. Given the observations in our snapshot—namely, that offline experiments are the dominant experiment type; that long-established but small datasets are commonly used; and that novel metrics have been shown to be of little value to assess the performance of recommender systems—, we conjecture that the advancements in these regards are limited overall. A longitudinal analysis would be a worthwhile research path to follow to gain a deeper insight into the developments made in the field of recommender systems evaluation. A further limitation is that we restricted our literature search to the ACM Digital Library. While we searched the extended collection of this library, which includes the essential conferences and journals where recommender systems research is typically published, we may have missed relevant papers published outside the typical venues, especially those outside of the general research space related to "computing". As the recommender systems field is increasingly viewed as an interdisciplinary research field, papers may be dispersed across a much wider scale of venues.

5 CONCLUSIONS

To gain insight into recent research focused on the evaluation of recommender systems, we conducted a systematic literature review. Our analysis covered papers published from 2017 to 2022, providing a thorough understanding of the current state of research on the evaluation of recommender systems within the research and practitioner communities. Throughout our review, we identified and discussed strengths and weaknesses in the field of recommender systems evaluation research. We observed notable strengths that demonstrate the continuous evolution and refinement of evaluation practices. These strengths are exemplified by the ongoing development of metrics, experiment types, and datasets that better accommodate the diverse use cases and requirements of recommender systems.

However, our analysis also brought to light certain weaknesses that require attention and improvement. One significant weakness is the persistent focus on recommendation problems that are deemed suboptimal proxies for user preferences, such as rating prediction. Additionally, the utilization of small and outdated datasets remains a challenge that hampers the overall advancement of recommender systems. To drive further progress and development in the realm of recommender systems, it is imperative for the research community to embrace the identified strengths and move away from outdated perspectives that contribute to the weaknesses. Achieving this objective is a collaborative effort that necessitates the collective expertise and participation of the entire recommender systems research community.

ACKNOWLEDGMENTS

This research was funded in whole, or in part, by the Austrian Science Fund (FWF): P33526. This research was funded in whole or in part by Vinnova.

REFERENCES

[1] Fabian Abel, Yashar Deldjoo, Mehdi Elahi, and Daniel Kohlsdorf. 2017. Recsys challenge 2017: Offline and online evaluation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys 2017)*. Association for Computing Machinery, New York, NY, USA, 372–373.

[2] Adekunle Oluseyi Afolabi and Pekka Toivanen. 2020. Harmonization and Categorization of Metrics and Criteria for Evaluation of Recommender Systems in Healthcare From Dual Perspectives. *International Journal of E-Health and Medical Communications (IJEHMC)*

11, 1 (2020), 69–92.
[3] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference of Very Large Data Bases (VLDB 1994, Vol. 1215)*. Santiago, Chile, 487–499.
[4] Wissam Al Jurdi, Jacques Bou Abdo, Jacques Demerjian, and Abdallah Makhoul. 2021. Critique on Natural Noise in Recommender Systems. *ACM Trans. Knowl. Discov. Data* 15, 5, Article 75 (may 2021), 30 pages. https://doi.org/10.1145/3447780
[5] Bushra Alhijawi, Arafat Awajan, and Salam Fraihat. 2022. Survey on the Objectives of Recommender Systems: Measures, Solutions, Evaluation Methodology, and New Perspectives. *Comput. Surveys* 55, 5, Article 93 (dec 2022), 38 pages. https://doi.org/10.1145/3527449
[6] Zafar Ali, Irfan Ullah, Amin Khan, Asim Ullah Jan, and Khan Muhammad. 2021. An Overview and Evaluation of Citation Recommendation Models. *Scientometrics* 126, 5 (may 2021), 4083–4119. https://doi.org/10.1007/s11192-021-03909-y
[7] James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen M Voorhees. 2017. TREC 2017 Common Core Track Overview. In *TREC*. 14 pages. https://trec.nist.gov/pubs/trec26/papers/Overview-CC.pdf
[8] Xavier Amatriain and Justin Basilico. 2016. Past, Present, and Future of Recommender Systems: An Industry Perspective. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 211–214. https://doi.org/10.1145/2959100.2959144
[9] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-N Recommendation Algorithms: A Quest for the State-of-the-Art. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 121–131. https://doi.org/10.1145/3503252.3531292
[10] Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2405–2414. https://doi.org/10.1145/3404835.3463245
[11] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3397271.3401032
[12] Linas Baltrunas, Karen Church, Alexandros Karatzoglou, and Nuria Oliver. 2015. Frappe: Understanding the Usage and Perception of Mobile App Recommendations In-The-Wild. *CoRR* abs/1505.03014 (2015). arXiv:1505.03014 http://arxiv.org/abs/1505.03014
[13] Jöran Beel and Victor Brunel. 2019. Data Pruning in Recommender Systems Research: Best-Practice or Malpractice?. In *Proceedings of ACM RecSys 2019 Late-Breaking Results co-located with the 13th ACM Conference on Recommender Systems, RecSys 2019 Late-Breaking Results, Copenhagen, Denmark, September 16-20, 2019 (CEUR Workshop Proceedings, Vol. 2431)*, Marko Tkalcic and Sole Pera (Eds.). CEUR-WS.org, Aachen, Germany, 26–30. http://ceur-ws.org/Vol-2431/paper6.pdf
[14] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2015. Research Paper Recommender Systems: A Literature Survey. *International Journal on Digital Libraries* 17, 4 (2015), 305–338. https://doi.org/10.1007/s00799-015-0156-0
[15] Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. 2013. Research Paper Recommender System Evaluation: A Quantitative Literature Survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation* (Hong Kong, China) *(RepSys '13)*. Association for Computing Machinery, New York, NY, USA, 15–22. https://doi.org/10.1145/2532508.2532512
[16] Poornima Belavadi, Laura Burbach, Stefan Ahlers, Martina Ziefle, and André Calero Valdez. 2021. Expectation, Perception, and Accuracy in News Recommender Systems: Understanding the Relationships of User Evaluation Criteria Using Direct Feedback. In *HCI International 2021 - Late Breaking Papers: Design and User Experience*, Constantine Stephanidis, Marcelo M. Soares, Elizabeth Rosenzweig, Aaron Marcus, Sakae Yamamoto, Hirohiko Mori, Pei-Luen Patrick Rau, Gabriele Meiselwitz, Xiaowen Fang, and Abbas Moallem (Eds.). Springer International Publishing, Cham, Germany, 179–197. https://doi.org/10.1007/978-3-030-90238-4_14
[17] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2017. Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal* 20, 6 (2017), 606–634.
[18] Alejandro Bellogín and Alan Said. 2018. Offline and Online Evaluation of Recommendations. In *Collaborative Recommendations*, Shlomo Berkovsky, Iván Cantador, and Domonkos Tikk (Eds.). World Scientific, Chapter Chapter 9, 295–328. https://doi.org/10.1142/9789813275355_0009
[19] Alejandro Bellogín and Alan Said. 2021. Improving accountability in recommender systems research through reproducibility. *User Modeling and User-Adapted Interaction* 31, 5 (2021), 941–977. https://doi.org/10.1007/s11257-021-09302-x
[20] Rocío Cañamares and Pablo Castells. 2017. A Probabilistic Reformulation of Memory-Based Collaborative Filtering: Implications on Popularity Biases. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 215–224. https://doi.org/10.1145/3077136.3080836
[21] Rocío Cañamares and Pablo Castells. 2018. Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor,

MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 415–424. https://doi.org/10.1145/3209978.3210014

[22] Rocío Cañamares and Pablo Castells. 2020. On Target Item Sampling in Offline Recommender System Evaluation. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/3383313.3412259

[23] Diego Carraro and Derek Bridge. 2022. A Sampling Approach to Debiasing the Offline Evaluation of Recommender Systems. *J. Intell. Inf. Syst.* 58, 2 (apr 2022), 311–336. https://doi.org/10.1007/s10844-021-00651-y

[24] Òscar Celma. 2010. *Music Recommendation and Discovery in the Long Tail*. Springer, Berlin, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-13287-2

[25] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *The World Wide Web Conference* (San Francisco, CA, USA) *(TheWebConf '19)*. Association for Computing Machinery, New York, NY, USA, 240–250. https://doi.org/10.1145/3308558.3313469

[26] Jin Yao Chin, Yile Chen, and Gao Cong. 2022. The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets?. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 141–149. https://doi.org/10.1145/3488560.3498519

[27] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (San Diego, CA, USA) *(KDD '11)*. Association for Computing Machinery, New York, NY, USA, 1082–1090. https://doi.org/10.1145/2020408.2020579

[28] Charles L Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the trec 2009 web track*. Technical Report. WATERLOO UNIV (ONTARIO).

[29] Charles L Clarke, Nick Craswell, and Ellen M Voorhees. 2012. *Overview of the TREC 2012 web track*. Technical Report. NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD.

[30] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. *TREC 2014 web track overview*. Technical Report. MICHIGAN UNIV ANN ARBOR.

[31] Randell Cotta, Mingyang Hu, Dan Jiang, and Peizhou Liao. 2019. Off-Policy Evaluation of Probabilistic Identity Data in Lookalike Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 483–491. https://doi.org/10.1145/3289600.3291033

[32] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[33] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees, and Ian Soboroff. 2021. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2369–2375. https://doi.org/10.1145/3404835.3463249

[34] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of Recommender Algorithms on Top-n Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 39–46. https://doi.org/10.1145/1864708.1864721

[35] Alexander Dallmann, Daniel Zoller, and Andreas Hotho. 2021. A Case Study on Sampling Strategies for Evaluating Neural Sequential Item Recommendation Models. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 505–514. https://doi.org/10.1145/3460231.3475943

[36] Zohreh Dehghani Champiri, Adeleh Asemi, and Salim Siti Salwah Binti. 2019. Meta-Analysis of Evaluation Methods and Metrics Used in Context-Aware Scholarly Recommender Systems. *Knowl. Inf. Syst.* 61, 2 (nov 2019), 1147–1178. https://doi.org/10.1007/s10115-018-1324-5

[37] Fernando Diaz and Andres Ferraro. 2022. Offline Retrieval Evaluation Without Evaluation Metrics. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 599–609. https://doi.org/10.1145/3477495.3532033

[38] Tome Eftimov, Bibek Paudel, Gorjan Popovski, and Dragi Kocev. 2021. A Framework for Evaluating Personalized Ranking Systems by Fusing Different Evaluation Measures. *Big Data Research* 25 (2021), 100211. https://doi.org/10.1016/j.bdr.2021.100211

[39] Michael D. Ekstrand. 2020. LensKit for Python: Next-Generation Software for Recommender Systems Experiments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2999–3006. https://doi.org/10.1145/3340531.3412778

[40] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *ACM Trans. Inf. Syst.* 39, 2, Article 20 (jan 2021), 49 pages. https://doi.org/10.1145/3434185

[41] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 101–109. https://doi.org/10.1145/3298689.3347058

[42] Shir Frumerman, Guy Shani, Bracha Shapira, and Oren Sar Shalom. 2019. Are All Rejected Recommendations Equally Bad? Towards Analysing Rejected Recommendations. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*

(Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 157–165. https://doi.org/10.1145/3320435.3320448

[43] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015)*. IEEE Press, 4274–4282. https://doi.org/10.1109/ICCV.2015.486

[44] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 198–206. https://doi.org/10.1145/3159652.3159687

[45] Asela Gunawardana, Guy Shani, and Sivan Yogev. 2022. Evaluating Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, USA, 547–601. https://doi.org/10.1007/978-1-0716-2197-4_15

[46] G. Guo, J. Zhang, and N. Yorke-Smith. 2013. A Novel Bayesian Similarity Measure for Recommender Systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence* (Beijing, China) *(IJCAI '13)*. AAAI Press, 2619–2625.

[47] Xunhua Guo, Lingli Wang, Mingyue Zhang, and Guoqing Chen. 2022. First Things First? Order Effects in Online Product Recommender Systems. *ACM Transations on Computer-Human Interacttion* 30, Article 15 (aug 2022), 35 pages. https://doi.org/10.1145/3557886

[48] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (Dec. 2015), 1–19.

[49] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (2016), 144–150. https://doi.org/10.1609/aaai.v30i1.9973

[50] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* 22, 1 (jan 2004), 5–53. https://doi.org/10.1145/963770.963772

[51] Ngozi Ihemelandu and Michael D. Ekstrand. 2021. Statistical Inference: The Missing Piece of RecSys Experiment Reliability Discourse. In *Proceedings of the Perspectives on the Evaluation of Recommender Systems, Workshop 2021 co-located with the 15th ACM Conference on Recommender Systems (RecSys 2021), Amsterdam, The Netherlands, September 25, 2021 (CEUR Workshop Proceedings, Vol. 2955)*, Eva Zangerle, Christine Bauer, and Alan Said (Eds.). CEUR-WS.org, Aachen, Germany, 10 pages. https://ceur-ws.org/Vol-2955/paper9.pdf

[52] Dietmar Jannach. 2023. Evaluating conversational recommender systems: A landscape of research. *Artificial Intelligence Review* 56, 3 (2023), 2365–2400. https://doi.org/10.1007/s10462-022-10229-x

[53] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara Fallacy: Towards more Impactful Recommender Systems Research. *AI Magazine* 41, 4 (Dec. 2020), 79–95. https://doi.org/10.1609/aimag.v41i4.5312

[54] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key Qualities of Conversational Recommender Systems: From Users' Perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI '21)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/3472307.3484164

[55] Thorsten Joachims, Ben London, Yi Su, Adith Swaminathan, and Lequn Wang. 2021. Recommendations as Treatments. *AI Magazine* 42, 3 (2021), 19–30. https://doi.org/10.1609/aaai.12014

[56] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining*. IEEE, 197–206. https://doi.org/10.1109/ICDM.2018.00035

[57] Barbara Kitchenham, Stuart Charters, David Budgen, Pearl Brereton, Mark Turner, Steve Linkman, Magne Jørgensen, Emilia Mendes, and Giuseppe Visaggio. 2007. *Guidelines for performing systematic literature reviews in software engineering*. EBSE Technical Report EBSE-2007-01, version 2.3. Keele University and University of Durham.

[58] Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, Nobuyuki Shimizu, and Yasuo Yamamoto. 2022. Doubly Robust Off-Policy Evaluation for Ranking Policies under the Cascade Behavior Model. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 487–497. https://doi.org/10.1145/3488560.3498380

[59] Pigi Kouki, Ilias Fountalis, Nikolaos Vasiloglou, Xiquan Cui, Edo Liberty, and Khalifeh Al Jadda. 2020. From the Lab to Production: A Case Study of Session-Based Recommendations in the Home-Improvement Domain. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 140–149. https://doi.org/10.1145/3383313.3412235

[60] Walid Krichene and Steffen Rendle. 2020. On Sampled Metrics for Item Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1748–1757. https://doi.org/10.1145/3394486.3403226

[61] Sara Latifi and Dietmar Jannach. 2022. Streaming Session-Based Recommendation: When Graph Neural Networks Meet the Neighborhood. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 420–426. https://doi.org/10.1145/3523227.3548485

[62] Sara Latifi, Dietmar Jannach, and Andrés Ferraro. 2022. Sequential recommendation: A study on transformers, nearest neighbors and sampled metrics. *Information Sciences* 609 (2022), 660–678. https://doi.org/10.1016/j.ins.2022.07.079

[63] Dong Li, Ruoming Jin, Jing Gao, and Zhi Liu. 2020. On Sampling Top-K Recommendation Evaluation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 2114–2124. https://doi.org/10.1145/3394486.3403262

[64] Roger Zhe Li, Julián Urbano, and Alan Hanjalic. 2021. New Insights into Metric Optimization for Ranking-Based Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 932–941. https://doi.org/10.1145/3404835.3462973

[65] Hongyu Lu, Weizhi Ma, Min Zhang, Maarten de Rijke, Yiqun Liu, and Shaoping Ma. 2021. Standing in Your Shoes: External Assessments for Personalized Recommender Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1523–1533. https://doi.org/10.1145/3404835.3462916

[66] Malte Ludewig, Noemi Mauro, Sara Latifi, and Dietmar Jannach. 2019. Performance Comparison of Neural and Non-Neural Approaches to Session-Based Recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 462–466. https://doi.org/10.1145/3298689.3347041

[67] Benjamin M. Marlin, Richard S. Zemel, Sam Roweis, and Malcolm Slaney. 2007. Collaborative Filtering and the Missing at Random Assumption. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* (Vancouver, BC, Canada) *(UAI'07)*. AUAI Press, Arlington, Virginia, USA, 267–275.

[68] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining (ICDM 2012)*. IEEE Press, 1020–1025.

[69] James McInerney, Brian Brost, Praveen Chandar, Rishabh Mehrotra, and Benjamin Carterette. 2020. Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1779–1788. https://doi.org/10.1145/3394486.3403229

[70] Elisa Mena-Maldonado, Rocío Cañamares, Pablo Castells, Yongli Ren, and Mark Sanderson. 2020. Agreement and Disagreement between True and False-Positive Metrics in Recommender Systems Evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 841–850. https://doi.org/10.1145/3397271.3401096

[71] Elisa Mena-Maldonado, Rocío Cañamares, Pablo Castells, Yongli Ren, and Mark Sanderson. 2021. Popularity Bias in False-Positive Metrics for Recommender Systems Evaluation. *ACM Trans. Inf. Syst.* 39, 3, Article 36 (may 2021), 43 pages. https://doi.org/10.1145/3452740

[72] Rishabh Misra, Mengting Wan, and Julian McAuley. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 422–426.

[73] Yusuke Narita, Shota Yasui, and Kohei Yata. 2021. Debiased Off-Policy Evaluation for Recommendation Systems. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 372–379. https://doi.org/10.1145/3460231.3474231

[74] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong) *(EMNLP-IJCNLP 2019)*. Association for Computational Linguistics, 188–197. https://doi.org/10.18653/v1/D19-1018

[75] Malte Ostendorff, Corinna Breitinger, and Bela Gipp. 2021. A Qualitative Evaluation of User Preference for Link-Based vs. Text-Based Recommendations of Wikipedia Articles. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, Proceedings (ICADL 2021)*, Hao-Ren Ke, Chei Sian Lee, and Kazunari Sugiyama (Eds.). Springer International Publishing, Cham, Germany, 63–79. https://doi.org/10.1007/978-3-030-91669-5_6

[76] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 75–84. https://doi.org/10.1145/3460231.3474234

[77] Ladislav Peska and Peter Vojtas. 2020. Off-Line vs. On-Line Evaluation of Recommender Systems in Small E-Commerce. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media* (Virtual Event, USA) *(HT '20)*. Association for Computing Machinery, New York, NY, USA, 291–300. https://doi.org/10.1145/3372923.3404781

[78] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating Recommender Systems from the User's Perspective: Survey of the State of the Art. *User Modeling and User-Adapted Interaction* 22, 4–5 (oct 2012), 317–355. https://doi.org/10.1007/s11257-011-9115-7

[79] Steffen Rendle, Li Zhang, and Yehuda Koren. 2019. On the Difficulty of Evaluating Baselines: A Study on Recommender Systems. https://doi.org/10.48550/ARXIV.1905.01395

[80] Alan Said and Alejandro Bellogín. 2018. Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction* 28, 2 (2018), 97–125. https://doi.org/10.1007/s11257-018-9202-0

[81] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2020. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. https://doi.org/10.48550/ARXIV.2008.07146

[82] Yuta Saito, Takuma Udagawa, Haruka Kiyohara, Kazuki Mogi, Yusuke Narita, and Kei Tateno. 2021. Evaluating the Robustness of Off-Policy Evaluation. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 114–123. https://doi.org/10.1145/3460231.3474245

[83] Pablo Sánchez and Alejandro Bellogín. 2022. Point-of-Interest Recommender Systems Based on Location-Based Social Networks: A Survey from an Experimental Perspective. *Comput. Surveys* 54, 11s, Article 223 (sep 2022), 37 pages. https://doi.org/10.1145/3510409

[84] Prabhat Kumar Saraswat, Samuel William, and Eswar Reddy. 2021. A Hybrid Approach for Offline A/B Evaluation for Item Ranking Algorithms in Recommendation Systems. In *The First International Conference on AI-ML-Systems* (Bangalore, India) *(AIMLSystems 2021)*. Association for Computing Machinery, New York, NY, USA, Article 21, 6 pages. https://doi.org/10.1145/3486001.3486241

[85] Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*. Association for Computing Machinery, Tokyo, Japan, 1241–1244. https://doi.org/10.1145/3077136.3080711

[86] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *International conference on machine learning*. PMLR, 1670–1679.

[87] Thiago Silva, Nícollas Silva, Heitor Werneck, Carlos Mito, Adriano C.M. Pereira, and Leonardo Rocha. 2022. IRec: An Interactive Recommendation Framework. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3165–3175. https://doi.org/10.1145/3477495.3531754

[88] Nasim Sonboli, Masoud Mansoury, Ziyue Guo, Shreyas Kadekodi, Weiwen Liu, Zijun Liu, Andrew Schwartz, and Robin Burke. 2021. Librec-Auto: A Tool for Recommender Systems Experimentation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 4584–4593. https://doi.org/10.1145/3459637.3482006

[89] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1441–1450. https://doi.org/10.1145/3357384.3357895

[90] Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. 2020. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 23–32. https://doi.org/10.1145/3383313.3412489

[91] Panagiotis Symeonidis, Andrea Janes, Dmitry Chaltsev, Philip Giuliani, Daniel Morandini, Andreas Unterhuber, Ludovik Coba, and Markus Zanker. 2020. Recommending the Video to Watch Next: An Offline and Online Evaluation at YOUTV.De. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 299–308. https://doi.org/10.1145/3383313.3412257

[92] Jiliang Tang, Huiji Gao, and Huan Liu. 2012. MTrust: Discerning Multi-Faceted Trust in a Connected World. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2124295.2124309

[93] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1235–1244. https://doi.org/10.1145/2783258.2783273

[94] Doris Xin, Nicolas Mayoraz, Hubert Pham, Karthik Lakshmanan, and John R. Anderson. 2017. Folding: Why Good Models Sometimes Make Spurious Recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 201–209. https://doi.org/10.1145/3109859.3109911

[95] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased Offline Recommender Evaluation for Missing-Not-at-Random Implicit Feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 279–287. https://doi.org/10.1145/3240323.3240355

[96] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *Comput. Surveys* 55, 8, Article 170 (2022), 38 pages. https://doi.org/10.1145/3556536

[97] Tong Zhao, Julian McAuley, and Irwin King. 2015. Improving latent factor models via personalized feature projection for one class recommendation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 821–830. https://doi.org/10.1145/2806416.2806511

[98] Wayne Xin Zhao, Zihan Lin, Zhichao Feng, Pengfei Wang, and Ji-Rong Wen. 2022. A Revisiting Study of Appropriate Offline Evaluation for Top-N Recommendation Algorithms. *ACM Trans. Inf. Syst.* 41, 2, Article 32 (dec 2022), 41 pages. https://doi.org/10.1145/3545796

[99] Jieming Zhu, Jinyang Liu, Shuai Yang, Qi Zhang, and Xiuqiang He. 2021. Open Benchmarking for Click-Through Rate Prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 2759–2769. https://doi.org/10.1145/3459637.3482486

[100] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) *(WWW '05)*. Association for Computing Machinery, New York, NY, USA, 22–32. https://doi.org/10.1145/1060745.1060754