Introduction to the Special Issue on Perspectives on Recommender Systems Evaluation

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

Evaluation plays a vital role in recommender systems—in research and practice—whether for confirming algorithmic concepts or assessing the operational validity of designs and applications. It may span the evaluation of early ideas and approaches up to elaborate implementations of systems integrated into everyday product settings; it may target a wide spectrum of different factors being evaluated. In this special issue, we explore recommender systems evaluation—theory and practice—while considering a diverse set of perspectives. These include recommender systems purposes, stakeholders, methodological approaches, and consequences. The collection of articles in this special issue offers insightful analyses of current recommender system evaluation practices, acknowledging their limitations, and setting out future research directions. As recommender systems evolve, the need for adequate evaluation methods and approaches increases. This special issue sheds light on areas undergoing development or requiring added attention from the research and practitioner communities in recommender systems. The compilation serves as a call to the recommender systems research community, motivating continued research and exploration of evaluation metrics, methods, and strategies.

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

Additional Key Words and Phrases: Evaluation, recommender systems, reproducibility, datasets, metrics

ACM Reference Format:
Christine Bauer, Alan Said, and Eva Zangerle. 2024. Introduction to the Special Issue on Perspectives on Recommender Systems Evaluation. *ACM Trans. Recomm. Syst.* 2, 1, Article 1 (March 2024), 5 pages. https://doi.org/10.1145/3648398

1 INTRODUCTION

Recommender systems simplify our choices, reshape our interactions, and aid us in discovering relevant items. Evaluation is central to recommender system research and practice, determining the systems' reliability, effectiveness, business value, and user satisfaction. Given this significant role, considerate evaluation is a prerequisite for progress. Effectively, recommender systems research revolves around their evaluation. Thereby, evaluation encompasses assessing various aspects, from

This research was funded in whole, or in part, by both the Austrian Science Fund (FWF; P33526) and Vinnova. This work also received support from the EXDIGIT (Excellence in Digital Sciences and Interdisciplinary Technology) project, funded by Land Salzburg under grant number 20204-WISS/263/6-6022.
Authors' addresses: C. Bauer, Paris Lodron University Salzburg, Salzburg, 5020, Jakob-Haringer-Strasse 1, Austria; e-mail: christine.bauer@plus.ac.at; A. Said, University of Gothenburg, Forskningsgången 6, 41756 Gothenburg, Sweden; e-mail: alansaid@acm.org; E. Zangerle, University of Innsbruck, Technikerstr. 21A, 6020 Innsbruck, Austria; e-mail: eva.zangerle@uibk.ac.at.

ACM 2770-6699/2024/03-ART1
https://doi.org/10.1145/3648398

initial ideas to fully operational systems. This special issue—*Perspectives on Recommender Systems Evaluation*—aims to present the current state of the art and emerging trends in recommender systems evaluation. The number of articles and the scope of topics highlight the importance of taking diverse viewpoints into account when evaluating recommender systems.

Further, the articles in this special issue collectively address the following question: "Why are diverse evaluation approaches important for recommender systems?" They do so by showcasing the adaptability of recommender systems to various applications and domains, catering to different purposes and stakeholders, and pointing out why a uniform evaluation methodology or set of metrics is insufficient. Recommender systems' quality differs depending on the use case and the goals set for the specific application; evaluation should thus reflect these varied expectations.

The consequences of the evaluation perspective chosen for recommender systems are significant, affecting user satisfaction, business outcomes, retention, computational performance, and a multitude of technical objectives and stakeholder-related subjective values. Simply put, a one-size-fits-all approach is insufficient, failing to capture the varied values recommender systems are expected to bring. Instead, we need diverse evaluation perspectives to maximize the potential of recommender systems.

This special issue presents articles examining, developing, and analyzing various perspectives of recommender systems evaluation, exploring the methodologies, and presenting challenges across the various application domains of recommender systems. Altogether, the articles in this special issue give refreshing new perspectives to the evaluation landscape, paving the way for new directions concerning evaluation efforts in the recommender systems research community. Overall, this special issue captures the big picture of the current challenges and trends in recommender systems evaluation, identifying and describing many strengths and weaknesses in current evaluation practices in recommender systems research. The articles in this issue also show that specific weaknesses require particular attention and improvement. Beyond this, the articles contribute novel ideas and guides for specific improvements.

While the articles in this special issue give perspective on where we need to head as a research community, it needs collaborative effort and the collective expertise of the entire recommender systems research community to drive significant progress in the field. Thus, this special issue serves as a call to the recommender systems research community, motivating continued research and exploration of evaluation metrics, methods, and strategies.

2 PERSPECTIVES ON RECOMMENDER SYSTEMS EVALUATION

This special issue contains a selection of 10 conference papers on topics revolving around the evaluation of recommender systems. In the following, we provide a brief overview of the accepted papers.

Taking the user perspective, Jin et al. [7] introduce CRS-Que, a user-centered framework designed for evaluating conversational recommender systems from the user perspective. Extending the well-established ResQue framework [13], CRS-Que enables the evaluation of conversational qualities, including adaptability and understanding. To empirically validate this novel framework, two user studies are conducted, examining music exploration and mobile phone purchasing scenarios. Similarly focusing on the user and the effect of recommendations, Porcaro et al. [12] present a longitudinal user study on the impact of diversity in music recommendation. In particular, this study investigates the impact of exposure to unfamiliar genres via recommendations and the effects on listener attitudes such as openness or willingness to discovery. Among further noteworthy findings, the study shows that diversified recommendations have the potential to increase the users' willingness to explore new genres.

From a rather technical perspective, the work by Michiels et al. [10] introduces a test framework for assessing the correctness of implemented recommender systems algorithms (including RecPack Tests, an open source Python package). Drawing inspiration from software testing methodologies, the framework provides a range of black box and white box tests across all levels of abstraction, including unit, integration, and system testing. Furthermore, Daniil et al. [4] present a comprehensive reproducibility study on the propagation of popularity bias and particularly investigate the differences in results found by previous studies on popularity bias [1, 8, 11]. Beyond mere reproducibility, this work analyzes the properties and implementation of the recommender systems and the evaluation thereof. The findings show that discrepancies in results can be attributed to factors such as the data used, employed algorithms, methods for dividing users into groups, and the employed evaluation strategies.

Focusing on the evaluation process, the work by Ekstrand et al. [5] advocates for attending to the distributions of evaluation metrics across, for example, user groups or item providers and going beyond measuring pointwise effectiveness, the current evaluation practice. Furthermore, the article presents a variety of tools for performing distributional evaluations and analyzing the results thereof. Four further articles of this special issue focus on evaluation methods: Li et al. [9] study the impact of performing item sampling-based evaluations (computing a ranking for a target item and a set of random further items) compared to global evaluation (ranking the items of a hold-out set). This work shows that sampling-based recall@K can be mapped to the global recall@K. In addition, notably, the authors propose an adaptive sampling approach that dynamically samples negative items for given users. Al Jurdi et al. [2] propose an evaluation framework that enables the detection of data samples (automatically determined subgroups of users) on which a recommendation algorithm performs poorly. With this approach, the authors address the challenge that the overall metric results across all users may remain unchanged (or even improve) while harming the experience for coherent groups of users. Rahdari et al. [14] contribute with a click model for 2D carousel-type interfaces. Such interfaces are commonly used in practice. Still, prior linear models cannot be used to evaluate such interactive recommendations in data-driven offline evaluations. The proposed click model simulates how users interact with recommendations in labeled carousel interfaces, allowing a low-cost and more accessible alternative to the online empirical method of assessing the quality of a carousel-type recommender system interface. Ferraro et al. [6] argue that recommendations for cultural content (e.g., music, movies, and literature) should align with principles of cultural citizenship, recognizing the social role of such content. The article introduces a commonality metric designed to gauge the extent to which a recommender system "contributes to the strengthening of cultural citizenship by systematically promoting diversity of source and content within a given type of cultural content."

Finally, Bauer et al. [3] present a systematic literature study on the current state of recommender systems evaluation practices. Spanning a period of 6 years (2017–2022) and analyzing 57 papers, the study specifically addresses experiment types, datasets, and metrics employed in the evaluation process. The results show that offline studies are predominantly used, and only a few datasets and evaluation metrics are widely used. Conversely, a variety of datasets and metrics is used in only a few papers.

3 MOVING ON

Current evaluation practices in the recommender systems research community tend to focus on offline evaluation, using a wide scale of datasets, but still extensively using MovieLens datasets [3]. Many contributions to the evaluation landscape propose evaluation models or introduce (or adapt) evaluation metrics to increase validity and have more informative results. However, at the same time, this also impacts comparability across papers [3].

Table 1. Three Extrapolated, Not Mutually Exclusive, Perspectives and the Five Contribution Types

Perspective	Contribution	Datasets	Articles in This Special Issue
End user	Metrics, model	MovieLens, Amazon review, LastFM, citeulike	[6, 7, 12, 14]
Service	Benchmark, metrics, framework		[6, 10]
Profit		NetflixPrize, Yahoo R3, Yelp	
Nonprofit		Amazon review, MovieLens, epinions	
Community	Survey, framework, benchmark	MovieLens, Amazon review, LastFM, citeulike, epinions	[2–5, 5, 7, 9, 10, 14]

With this backdrop, it is interesting to consider the wider research and practitioner perspectives used in recommender systems evaluation research. Considering the analysis of the evaluation landscape performed in the work of Bauer et al. [3], we can extrapolate the evaluation perspectives of the analyzed body of literature (57 papers). Bauer et al. [3] classify the type of contribution as one of five types: "*Benchmark*—Providing an ... evaluation across a ... set of datasets," "*Framework*—Introducing a framework for evaluation ... ," "*Metrics*—Analyzing ... metrics of evaluation," "*Model*—Introducing a ... model," and "*Survey*—A literature survey." One possible extrapolation of a wider set of perspectives would be "*End user*—Aligning with the perspectives of the end users," "*Service*—Aligning with the perspectives of the service (and service provider)," and "*Community*—Aligning with the perspectives of the research and practitioner communities in recommender systems." A further perspective is to split the service perspective into two separate perspectives: one where there is an inherent profit perspective from a for-profit business and one where the service does not involve a financial perspective. Table 1 shows a suggested classification of the contribution types into the three perspectives. This classification is one interpretation of the types of perspectives research contributions may have; unquestionably, other types of perspectives and classifications could be made too.

Using this classification as a background for the articles in this special issue, we can extend the classification to the works published in the context of this special issue (column *Articles in this Special Issue* in Table 1).

Analogously, we can extrapolate the perspectives of datasets commonly used in recommender systems evaluation research. Considering the most common datasets identified in the work of Bauer et al. [3] (i.e., MovieLens, Amazon, LastFM, citeulike, NetflixPrize, Yelp, Yahoo, and epinions), we can classify the perspectives of the datasets. In this context, we refer to the perspective from which the dataset is spun. For instance, the MovieLens dataset comes from a service without underlying financial mechanisms—it is the result of a long-spanning research project/infrastructure hosted by the University of Minnesota's GroupLens Lab; however, the Netflix Prize dataset was created and released by a for-profit organization relying on the quality of their recommendation to increase the profit margin and customer satisfaction. Datasets such as Amazon Review and citeulike represent a nonprofit perspective, as they have been crawled and released by third parties (researchers) by means of, for example, scraping websites of these services. While the services themselves have underlying profit perspectives, these were not instrumental in creating these datasets. The overarching service perspective conveys the fact that the data, nevertheless, is a result of the recommender systems that these services run—that is, recommender systems employed by the services directly influence how the data was constructed.

While many possible perspectives can be taken with regard to data, algorithms, and metrics, it remains important to remember that recommender systems provide decision support for their users and, in doing so, must keep track of the end user perspective independent of remaining parameters.

ACKNOWLEDGMENTS

We thank all authors who submitted their work to this special issue and all reviewers who devoted a substantial amount of time and effort to providing constructive and insightful feedback to the authors.

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *CoRR abs/1907.13286* (2019). http://arxiv.org/abs/1907.13286

[2] Wissam Al Jurdi, Jacques Bou Abdo, Jacques Demerjian, and Abdallah Makhoul. 2024. Group validation in recommender systems: Framework for multi-layer performance evaluation. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3640820

[3] Christine Bauer, Eva Zangerle, and Alan Said. 2024. Exploring the landscape of recommender systems evaluation: Practices and perspectives. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3629170

[4] Savvina Daniil, Mirjam Cuper, Cynthia C. S. Liem, Jacco van Ossenbruggen, and Laura Hollink. 2024. Reproducing popularity bias in recommendation: The effect of evaluation strategies. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3637066

[5] Michael D. Ekstrand, Ben Carterette, and Fernando Diaz. 2024. Distributionally-informed recommender system evaluation. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3613455

[6] Andreas Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. 2024. Measuring commonality in recommendation of cultural content: Recommender systems to enhance cultural citizenship. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3643138

[7] Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. 2024. CRS-Que: A user-centric evaluation framework for conversational recommender systems. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3631534

[8] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval*. Lecture Notes in Computer Science, Vol. 12036. Springer, 35–42. https://doi.org/10.1007/978-3-030-45442-5_5

[9] Dong Li, Ruoming Jin, Zhenming Liu, Bin Ren, Jing Gao, and Zhi Liu. 2024. On item-sampling evaluation for recommender system. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3629171

[10] Lien Michiels, Robin Verachtert, Andres Ferraro, Kim Falk, and Bart Goethals. 2024. A framework and toolkit for testing the correctness of recommendation algorithms. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3591109

[11] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Mahdi Dehghan. 2022. The unfairness of popularity bias in book recommendation. In *Advances in Bias and Fairness in Information Retrieval*. Springer, 69–81. https://doi.org/10.1007/978-3-031-09316-6_7

[12] Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2024. Assessing the impact of music recommendation diversity on listeners: A longitudinal study. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3608487

[13] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys'11)*. ACM, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962

[14] Behnam Rahdari, Peter Brusilovsky, and Branislav Kveton. 2024. Towards simulation-based evaluation of recommender systems with carousel interfaces. *ACM Transactions on Recommender Systems* 2, 1 (2024). https://doi.org/10.1145/3643709

Received 5 February 2024; accepted 12 February 2024