Perspectives on the Evaluation of Recommender Systems (PERSPECTIVES)

EVA ZANGERLE, Universität Innsbruck, Austria
CHRISTINE BAUER, Utrecht University, The Netherlands
ALAN SAID, University of Gothenburg, Sweden

Evaluation is a cornerstone in the process of developing and deploying recommender systems. The PERSPECTIVES workshop brought together academia and industry to critically reflect on the evaluation of recommender systems. Particularly, the workshop aimed to shed light on the different, and maybe even diverging or contradictory perspectives on the evaluation of recommender systems. Papers reporting a reflection on problems regarding recommender systems evaluation and lessons learned were solicited. The workshop combined flash presentations of accepted papers, a keynote from industry, and an interactive part with discussions in break-out rooms as well as in the plenum. The workshop complemented the program of the main conference as it emphasized problems and lessons learned, fostered exchange integrating various perspectives on evaluation, and sought to move the recommender systems community forward as an outcome of the workshop.

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

Additional Key Words and Phrases: evaluation, methods, recommender systems

ACM Reference Format:
Eva Zangerle, Christine Bauer, and Alan Said. 2021. Perspectives on the Evaluation of Recommender Systems (PERSPECTIVES). In *Fifthteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3460231.3470929

1 INTRODUCTION

Evaluation is essential when conducting rigorous research in the area of recommender systems (RecSys). As for most systems, evaluation demands attention in each and every phase through the system's life cycle—in design and development as well as for continuous improvement while in operation. Thereby, the evaluation may assess the core performance of a system in its very sense or may embrace the entire context in which the system is used [3, 4, 7, 8].

This introduction pinpoints that the evaluation of RecSys may target a wide spectrum of different aspects being evaluated, and it also shows that the evaluation of a RecSys may span the evaluation of early ideas and approaches up to elaborate systems in operation. Naturally, we do (and have to) take various perspectives on the evaluation of RecSys. Thereby, the term "perspective" may, for instance, refer to various purposes of a RecSys [5], the various stakeholders affected by a RecSys [1, 2], or the potential risks that are ought to be minimized [6]. Further, we have to consider that various methodological approaches and experimental designs represent further different perspectives on evaluation. The perspective on the evaluation of RecSys may also be substantially characterized by the available resources. For instance, academia and industry have different resources at their disposal for evaluation activities. The access to resources will likely be different for students compared to established researchers equipped with large teams and budget.

Manuscript submitted to ACM

A simple glance at, e.g., the RecSys community's yearly RecSys Challenge (http://www.recsyschallenge.com) highlights the varied evaluation metrics, methods, and strategies important for the various companies and use cases involved in the challenge over the last few years. While this is simply an example of the varied evaluation perspectives important in different settings, studying the evaluations used in the papers published at the main RecSys conference shows an even more divergent set of recommendation goals and metrics used to identify how well they have been met.

Acknowledging that there are various perspectives on the evaluation of RecSys, we want to put into discussion *whether there is a "golden standard" for the evaluation of a RecSys, and—if so—if it is indeed "golden" in any sense.* We postulate that the many and varied perspectives are all valid and reasonable, and aim to reach out to the RecSys community to entice discussion on the topic.

The goal of the workshop is to capture the current state of evaluation, and gauge whether there is, or should be, a different target that RecSys evaluation should strive for. The workshop addresses the question "where should we go from here as a community?" and aims at coming up with concrete steps for action.

A critical interest of this workshop is to integrate the perspectives from both academia and industry. We have a particularly strong commitment to integrate researchers at the beginning of their careers, and want to equally integrate established researchers. It is our particular concern to *give a voice to the various perspectives involved.*

2 TOPICS OF INTEREST AND MATERIAL

The workshop solicited papers addressing topics such as those listed below. Going beyond papers, we sought to gather feedback from participants before the workshop with respect to pressing issues regarding the evaluation of recommender systems that should be addressed in the workshop. Hence, the topics discussed during the workshop went beyond the list of topics below.

Topics of interest include, but are not limited to, the following:

- Case studies of difficult, hard-to-evaluate scenarios
- Evaluations with contradicting results
- Showcasing (structural) problems in RecSys evaluation
- Integration of offline and online experiments
- Multi-Stakeholder evaluation
- Divergence between evaluation goals and what is actually captured by the evaluation
- Nontrivial and unexpected experiences from practitioners

We deliberately solicited papers reporting problems and (negative) experiences regarding RecSys evaluation, as reflection on unsuccessful, inadequate, or insufficient evaluations is a fruitful source for yet another perspective on RecSys evaluation that can spark discussions. Accordingly, submissions could also address the following themes: (a) "lessons learned" from the successful application of RecSys evaluation or from "post mortem" analyses describing specific evaluation strategies that failed to uncover decisive elements, (b) "overview papers" analyzing patterns of challenges or obstacles to evaluation, and (c) "solution papers" presenting solutions for specific evaluation scenarios. Additionally, (d) "visionary papers" discussing novel and future evaluation aspects were be considered as well.

The workshop materials can be found on the workshop website at https://perspectives-ws.github.io/. Accepted papers are published as open access workshop proceedings via ceur-ws.org[1]. Supplemental material (e.g., presentation

[1]https://ceur-ws.org

slides, posters, summaries of the discussions in the break-out rooms, etc.) are—on authors' approval—available on the workshop website.

ACKNOWLEDGMENTS

We would like to thank the authors, presenters, and reviewers for their valuable contributions to the workshop.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* (2020). https://doi.org/10.1007/s11257-019-09256-1

[2] Christine Bauer and Eva Zangerle. 2019. Leveraging Multi-Method Evaluation for Multi-Stakeholder Settings. In *Proceedings of the 1st Workshop on the Impact of Recommender Systems* (Copenhagen, Denmark, 19 September) *(ImpactRS '19)*, Oren Sar Shalom, Dietmar Jannach, and Ido Guy (Eds.). 3 pages. arXiv:2001.04348 arXiv:2001.04348.

[3] 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)*. ACM, New York, NY, USA, 15–22. https://doi.org/10.1145/2532508.2532512

[4] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transaction on Information Systems* 22, 1 (Jan. 2004), 5–53. https://doi.org/10.1145/963770.963772

[5] Dietmar Jannach and Gediminas Adomavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, MA, USA) *(RecSys '16)*. ACM, New York, NY, USA, 7–10. https://doi.org/10.1145/2959100.2959186

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

[7] Dietmar Jannach, Oren Sar Shalom, and Joseph A Konstan. 2019. Towards More Impactful Recommender Systems Research. In *Proceedings of the 1st Workshop on the Impact of Recommender Systems co-located with 13th ACM Conference on Recommender Systems, ImpactRS@RecSys 2019)* (Copenhagen, Denmark) *(CEUR Workshop Proceedings, Vol. 2462)*. CEUR-WS.org. http://ceur-ws.org/Vol-2462/short6.pdf

[8] Alan Said, Domonkos Tikk, Klara Stumpf, Yue Shi, Martha Larson, and Paolo Cremonesi. 2012. Recommender Systems Evaluation: A 3D Benchmark. In *Proceedings of the Workshop on Recommendation Utility Evaluation: Beyond RMSE* (Dublin, Ireland) *(RUE '12, Vol. 910)*. CEUR Workshop Proceedings, 21–23. http://ceur-ws.org/Vol-910/