On Challenges of Evaluating Recommender Systems in an Offline Setting

Aixin Sun
axsun@ntu.edu.sg
Nanyang Technological University
Singapore

ABSTRACT

In the past 20 years, the area of Recommender Systems (RecSys) has gained significant attention from both academia and industry. We are not in short of research papers on various RecSys models or online systems from industry players. However, in terms of model evaluation in offline settings, many researchers simply follow the commonly adopted experiment setup, and have not zoomed into the unique characteristics of the RecSys problem. In this tutorial, I will briefly review the commonly adopted evaluations in RecSys then discuss the challenges of evaluating recommender systems in an offline setting. The main emphasis is the consideration of global timeline in the evaluation, particularly when a dataset covers user-item interactions that have been collected from a long time period.

CCS CONCEPTS

• **Information systems → Recommender systems**; *Evaluation of retrieval results.*

KEYWORDS

recommender systems, evaluation, global timeline.

ACM Reference Format:
Aixin Sun. 2023. On Challenges of Evaluating Recommender Systems in an Offline Setting. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3604915.3609495

1 MOTIVATION

Recent years have witnessed the significant development in Recommender Systems (RecSys), evidenced by the ever increasing number of publications on this topic as well as the number of applications that are supported by various recommenders. Although we are exciting about new models and new applications in this area, the evaluation of recommenders has often been overlooked. Researchers typically follow existing settings in their experiments with an aim of a fair comparison with earlier publications, without zooming into the practical setting of RecSys.

Recently, a few papers report counter-intuitive observations made from experiments, both in offline and online settings. For example, it is observed that users who have many interactions with a recommendation system receive poorer recommendations [8], and "using only the more recent parts of a dataset can drastically improve the performance of a recommendation system" [15]. These counter-intuitive observations motivate us to relook at the evaluation of recommendation models, in an offline setting, and the key challenges.

2 TUTORIAL OUTLINE

This tutorial is prepared for 90 minutes, targeting on the students who are familiar with recommender systems in general. Hence, the tutorial can be considered as an intermediate to advanced level tutorial, with a specific focus on RecSys evaluation in an offline setting, from an accuracy perspective.

The tutorial will be organized in three parts. The first part is on the review of commonly used RecSys evaluations. The content for this part will be mainly based on two recent survey papers on RecSys evaluation [1, 16]. Different evaluation objectives and measures will be covered, including those measures that are used in industry like Click-through Rate (CTR), Conversion Rate (CVR), and Gross Merchandise Value (GMV).

The second part is on the revisit of the evaluation in an offline setting, particularly the observation of the global timeline. The key issue here is not what measures/metrics to use, but how these measures are computed from a dataset. We will start with the ill-defined popularity model. In essence, popularity is often considered as the simplest recommendation baseline and is widely used for comparison purpose in evaluation. In reality, popularity has a strong temporal perspective. However, in many evaluations reported in academic research, the temporal perspective has become transparent due to various challenges, like data sparsity. We will use real examples to illustrate how popularity works in reality and how popularity is defined and evaluated in research papers. From the popularity evaluation, we extend the discussion to data leakage and its impact on RecSys evaluation results [10, 13, 17]. As models are often developed to achieve better measures, if the evaluation is not conducted correctly, there might be an impact on the effectiveness of these models in reality. Following the data leakage, we will further discuss another potential issue of ignoring timeline in evaluation, the simplification of user preference modeling.

In the last part of the tutorial, there will be a summary of the criticism on RecSys, with the key focus from the evaluation perspective. Although there many large-scale empirical evaluations [5, 12, 13, 17, 18], there remain questions on reproducibility, and technical and theoretical flaws [4, 6]. We will also cover a bit on the challenges in evaluating RecSys from different perspectives in offline settings [2, 3, 14].

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3609495

This tutorial is concluded with a fresh look at RecSys evaluation on how to conduct more meaningful evaluations by considering the global timeline [11]. Here are the topics in an itemized view:

Part I

- Introduction (10 min)
  - Recommender system basics
  - Applications powered by RecSys
- Commonly used RecSys evaluation metrics (20 min)
  - Commonly used metrics in academic research
  - Metrics used for different applications in online settings *e.g.*, e-commence, advisement, video, music, and news recommendations.

Part II

- Challenges in computing the offline metrics (40 min)
  - How RecSys works in practice with Popularity as an example
  - Data partition schemes in RecSys experiments using offline datasets
  - Data leakage due to not maintaining global timeline
  - The impact on understanding the RecSys research problem

Part III

- Criticism on RecSys from evaluation perspective (10 min)
  - The counter-intuitive observations
  - The common pitfalls in evaluating RecSys
- More practical evaluations (10 min)
  - The meaning of fair comparison
  - The observation of global timeline

3 TUTORIAL PRESENTER

Dr. Aixin Sun is an Associate Professor and Associate Chair (Academic) at the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU), Singapore. He received B.A.Sc (1st class honours) and Ph.D. both in Computer Engineering from NTU Singapore in 2001 and 2004 respectively. Dr. Sun is an associate editor of ACM Transactions on Information Systems (TOIS), Neurocomputing, an editorial board member of Journal of the Association for Information Science and Technology (JASIST), and Information Retrieval Journal. He has served as the Doctoral Consortium co-chair for WSDM2023, demonstration track co-chair for SIGIR2020, ICDM2018, CIKM2017, PC co-chair for AIRS2019, and general chair for ADMA2017. He has also served as Area Chair, Senior PC member or PC member for many conferences including SIGIR, WWW, WSDM, EMNLP, AAAI, and IJCAI. Dr. Sun has co-authored a few research papers related to RecSys evaluation [7–9, 11]

REFERENCES

[1] Bushra Alhijawi, Arafat Awajan, and Salam Fraihat. 2022. Survey on the Objectives of Recommender Systems: Measures, Solutions, Evaluation Methodology, and New Perspectives. *ACM Comput. Surv.* 55, 5, Article 93 (2022). https://doi.org/10.1145/3527449

[2] Pablo Castells and Alistair Moffat. 2022. Offline recommender system evaluation: Challenges and new directions. *AI Magazine* 43, 2 (2022), 225–238. https://doi.org/10.1002/aaai.12051

[3] Hung-Hsuan Chen, Chu-An Chung, Hsin-Chien Huang, and Wen Tsui. 2017. Common Pitfalls in Training and Evaluating Recommender Systems. *SIGKDD Explorations* 19, 1 (2017), 37–45.

[4] Paolo Cremonesi and Dietmar Jannach. 2021. Progress in Recommender Systems Research: Crisis? What Crisis? *AI Magazine* 42, 3 (Nov. 2021), 43–54. https://doi.org/10.1609/aimag.v42i3.18145

[5] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *RecSys*. ACM, 101–109.

[6] 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 (2021). https://doi.org/10.1145/3434185

[7] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2020. A Re-visit of the Popularity Baseline in Recommender Systems. In *SIGIR*. ACM, 1749–1752.

[8] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2022. Do Loyal Users Enjoy Better Recommendations? Understanding Recommender Accuracy from a Time Perspective. In *ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR)* (Madrid, Spain). ACM, 92–97. https://doi.org/10.1145/3539813.3545124

[9] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2023. A Critical Study on Data Leakage in Recommender System Offline Evaluation. *ACM Trans. Inf. Syst.* 41, 3 (2023), 75:1–75:27. https://doi.org/10.1145/3569930

[10] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2020. Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In *RecSys*. ACM, 681–686.

[11] Aixin Sun. 2023. Take a Fresh Look at Recommender Systems from an Evaluation Standpoint. In *SIGIR*. ACM, 2629–2638. https://doi.org/10.1145/3539618.3591931

[12] Zhu Sun, Hui Fang, Jie Yang, Xinghua Qu, Hongyang Liu, Di Yu, Yew-Soon Ong, and Jie Zhang. 2023. DaisyRec 2.0: Benchmarking Recommendation for Rigorous Evaluation. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 7 (2023), 8206–8226. https://doi.org/10.1109/TPAMI.2022.3231891

[13] 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 *RecSys*. ACM, 23–32. https://doi.org/10.1145/3383313.3412489

[14] Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. 2021. Quality Metrics in Recommender Systems: Do We Calculate Metrics Consistently?. In *RecSys* (Amsterdam, Netherlands). ACM, 708–713. https://doi.org/10.1145/3460231.3478848

[15] Robin Verachtert, Lien Michiels, and Bart Goethals. 2022. Are We Forgetting Something? Correctly Evaluate a Recommender System With an Optimal Training Window. In *Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES) at RecSys22*. Seattle, WA, USA.

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

[17] 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

[18] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. 2022. BARS: Towards Open Benchmarking for Recommender Systems. In *SIGIR*. ACM, 2912–2923. https://doi.org/10.1145/3477495.3531723