Mab2Rec: Contextual Multi-Armed Bandits for Recommender Systems

Serdar Kadıoğlu [1, 2]

[1] AI Center of Excellence, Fidelity Investments, Boston, USA
[2] Department of Computer Science, Brown University, Providence, USA

Abstract

In this talk, we present a modular framework and showcase MAB2REC for building recommender systems from higher-order abstractions. The components of MAB2REC span multi-armed bandits, natural language processing, constraint-based sequential pattern mining, feature selection, performance and fairness evaluation. These components are embodied in industry-strength software and are contributed to the open-source community to enable other researchers and practitioners.

1 Introduction

The applications of personalization and recommenders are widespread in the industry, see, e.g., (Amatriain and Basilico 2016) for a comprehensive overview. Despite being so pervasive, building recommender *systems* still require complex machinery of data preparation, transformation, modeling, and production deployment. Each of these components requires significant domain expertise and serious engineering effort to deploy a robust system that can operate at scale.

In this talk, we share our multi-year effort on designing a modular framework and showcase MAB2REC[1] for building recommender systems based on higher-order abstractions. The components of MAB2REC span multi-armed bandits (Strong, Kleynhans, and Kadıoğlu 2019, 2021; Kilitçioglu and Kadıoğlu 2022), natural language processing (Kilitcioglu and Kadıoğlu 2021), constraint-based sequential pattern mining (Wang et al. 2022; Kadıoğlu et al. 2023; Wang and Kadıoğlu 2022; Ghosh et al. 2022), feature selection (Kadıoğlu, Kleynhans, and Wang 2021; Kleynhans, Wang, and Kadıoğlu 2021), performance and fairness evaluation (Michalský and Kadıoğlu 2021; Cheng, Kilitçioglu, and Kadıoğlu 2022; Thielbar et al. 2023). These components are embodied in industry-strength software and are contributed to the open-source community to enable other researchers and practitioners.

In the field, thanks to decade-long efforts, several specialized frameworks exist for recommender systems. Most prominent players of the technology industry contributed dedicated software such as Torch Recommenders from Meta (Naumov and Mudigere 2020), TensorFlow Recommenders from Google (Pasumarthi et al. 2019), Recommenders from Microsoft (Argyriou, González-Fierro, and Zhang 2020), and Merlin Recommenders from NVIDIA (Oldridge et al. 2020). These frameworks can be used to create models for large-scale recommendation systems with a strong focus on scalability, performance, and deployment.

Despite these efforts, two issues remain; one challenge for users of these systems and another challenge for companies who want to deploy them.

First, these are monolithic frameworks that are specialized in one and only one task, i.e., building recommenders. The effective use of these systems still requires expert skills. This poses a significant learning curve for data scientists and practitioners, and even with the best effort, the skill is not immediately transferable to another task, e.g., propensity modeling, natural language processing, text featurization, feature selection, pattern mining, and so on. Such in-depth knowledge might be desirable for recommendation experts, and at the same time, is a blocker for the broader community interested in personalization applications.

Second, these frameworks require serious engineering capabilities to meld the training and testing pipeline together in an effort for real-time performance at scale, plus hardware requirements for extremely data-hungry neural networks that expect millions of user and item interactions. This engineering requirement is a barrier for industry players that are not necessarily software-driven, unlike the high-tech companies that built these systems for their specific needs. Personalization is not a privileged application that should remain only accessible to technology companies, and most other businesses have opportunities to personalize touchpoints for their end-users. In fact, from the end-user's perspective, today, we would like any interaction to be personalized seamlessly for our needs regardless of the provider.

On the one hand, we have practitioners with a general background in machine learning but cannot commit to becoming recommenders system experts and prefer more transferable skills. On the other hand, we have companies that must meet their business demand for personalization but cannot commit to building and maintaining the large-scale infrastructure only to satisfy requirements of a single task.

[1] http://github.com/fidelity/mab2rec

To address these challenges, we advocate a modular approach that tightly integrates yet maintains the independence of individual components, thus satisfying two of the most critical aspects of industrial applications, *generality* and *specificity*. On the one hand, we ensure that each component remains self-contained and is ready to serve in other applications beyond recommender systems. On the other hand, when these components are combined, a unified theme emerges for recommender systems. We present the details of each component in the context of recommender systems and other applications.

We release each component as an open-source library, and most importantly, we release their integration under MAB2REC, an industry-strength open-source software for building bandit-based recommender systems. By bringing standalone components together, MAB2REC realizes a powerful and scalable toolchain to build and deploy business-relevant personalization applications.

Beyond the showcase of our toolchain, MAB2REC and its underlying components, we share our experience and best practices for user adoption, model training, performance evaluation, deployment, and governance within the enterprise and the broader community.

References

Amatriain, X.; and Basilico, J. 2016. Past, present, and future of recommender systems: An industry perspective. In *ACM RecSys*.

Argyriou, A.; González-Fierro, M.; and Zhang, L. 2020. Microsoft Recommenders: Best Practices for Production-Ready Recommendation Systems. In *Proceedings of the Web Conference*, 50–51.

Cheng, D.; Kilitçioglu, D.; and Kadıoğlu, S. 2022. Bias mitigation in recommender systems to improve diversity. In *CIKM*, CEUR.

Ghosh, S.; Yadav, S.; Wang, X.; Chakrabarty, B.; and Kadıoğlu, S. 2022. Dichotomic Pattern Mining Integrated with Constraint Reasoning for Digital Behaviour Analyses. *Frontiers in AI*.

Kadıoğlu, S.; Kleynhans, B.; and Wang, X. 2021. Optimized Item Selection to Boost Exploration for RecSys. In *CPAIOR*.

Kadıoğlu, S.; Wang, X.; Hosseininasab, A.; and van Hoeve, W.-J. 2023. Seq2Pat: Sequence-to-pattern generation to bridge mining with machine learning. *AI Magazine*.

Kilitcioglu, D.; and Kadıoğlu, S. 2021. Representing the Unification of Text Featurization using a Context-Free Grammar. *AAAI*.

Kilitçioglu, D.; and Kadıoğlu, S. 2022. Non-Deterministic Behavior of TS with Linear Payoffs and How to Avoid It. *TMLR*, 2022.

Kleynhans, B.; Wang, X.; and Kadıoğlu, S. 2021. Active Learning Meets Optimized Item Selection. In *DSO Workshop, IJCAI*.

Michalský, F.; and Kadıoğlu, S. 2021. Surrogate Ground Truth Generation to Enhance Binary Fairness Evaluation in Uplift Modeling. In *IEEE ICMLA*, 1654–1659.

Naumov, M.; and Mudigere, D. 2020. DLRM: An advanced, open source deep learning recommendation model.

Oldridge, E.; Perez, J.; Frederickson, B.; Koumchatzky, N.; Lee, M.; Wang, Z.; Wu, L.; Yu, F.; Zamora, R.; Yılmaz, O.; et al. 2020. Merlin: a gpu accelerated recommendation framework. *IRS*.

Pasumarthi, R. K.; Bruch, S.; Wang, X.; Li, C.; Bendersky, M.; Najork, M.; Pfeifer, J.; Golbandi, N.; Anil, R.; and Wolf, S. 2019. Tf-ranking: Scalable tensorflow library for learning-to-rank. In *KDD*.

Strong, E.; Kleynhans, B.; and Kadıoğlu, S. 2019. MABWiser: A Parallelizable Contextual MAB Library for Python. In *IEEE ICTAI*.

Strong, E.; Kleynhans, B.; and Kadıoğlu, S. 2021. MABWiser: Parallelizable Contextual Multi-armed Bandits. *IJAIT*, 30(4).

Thielbar, M.; Kadıoğlu, S.; Zhang, C.; Pack, R.; and Dannull, L. 2023. Surrogate Membership for Inferred Metrics in Fairness Evaluation. In Sellmann, M.; and Tierney, K., eds., *Learning and Intelligent Optimization*, 424–442. Cham: Springer International Publishing. ISBN 978-3-031-44505-7.

Wang, X.; Hosseininasab, A.; Pablo, C.; Serdar, K.; and van Hoeve, W.-J. 2022. Seq2Pat: Sequence-to-Pattern Generation for Constraint-based Sequential Pattern Mining. In *AAAI-IAAI*.

Wang, X.; and Kadıoğlu, S. 2022. Dichotomic Pattern Mining with Applications to Intent Prediction from Semi-Structured Clickstream Datasets. In *KDF-AAAI-22*.