Adoption of recommendation systems: Observations, Trends and Leveling the Playing Field

Jaidev Shah
Microsoft AI, USA
jaidevshah@microsoft.com

Miguel González-Fierro
Microsoft, Spain
miguel.gonzalezfierro@microsoft.com

Abstract

1 Author Bio

Jaidev Shah is an Applied Scientist at Microsoft AI. On the Bing webpage recommendations team, Jaidev was a major contributor in leveraging LLMs to bring significant recommendation quality gains to the web-scale system. Jaidev's work has been included in top recommendations and search conferences such as Recsys and CIKM. At these conferences, he has engaged with dozens of industry practitioners, from startups to companies with established systems operating at the cutting edge, to discuss and understand real-world recommendation system challenges across applications and the solutions deployed to address them. Jaidev is now part of the core ranking team at Bing Search, working on problems such as knowledge distillation, leveraging LLMs for search relevance, and learning to rank. Jaidev holds a Bachelor's and Master's degree in Computer Science from Columbia University.

Miguel González-Fierro is a Principal Data Science Manager at Microsoft and a core maintainer of the Recommenders repository, a well-adopted open source project, originally developed by Microsoft, and now under the Linux Foundation of AI & Data (LF AI&Data). He has worked with hundreds of Microsoft customers to help them build and deploy recommendation systems. Miguel also works with internal teams to deploy recommendation systems in production across various Microsoft products. Miguel holds a PhD in robotics from University Carlos III of Madrid, in collaboration with King's College London and graduated from MIT Sloan School of Management.

2 Introduction

The Recommenders team at Microsoft, together with other contributors, maintains the popular Recommenders GitHub repository[1], a well-known resource for best practices and algorithm implementations for different use-cases in recommendation systems. Recommenders provides modular utilities for model creation, data manipulation, and evaluation alongside algorithms like NeuralCF, SAR, Neural Recommendation with Multi-Head Self-Attention (NRMS), ALS, Wide&Deep, Bayesian Pairwise Ranking (BPR), Graph Networks and Deep Knowledge-Aware Network (DKN) on different environments including CPU, GPU, and Spark [1]. Microsoft's Recommenders team has worked with hundreds of customers to help them design and deploy recommendation systems as well as to evaluate whether the investment is worth it for their use-cases. Through driving discussion, we hope to share insights from the team's experience working with hundreds of customers across various industries, primarily discussing the key observations and bottlenecks to industry adoption of recommendation systems. We will also share our view on potential solutions and insights from industry for addressing some of the key technical challenges for new businesses developing recommendation systems, such as cold-start and data sparsity.

[1] https://github.com/recommenders-team/recommenders

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

3 Observations and Trends

The Recommenders team has worked with businesses across the spectrum in size, spanning across several industries including:

- Retail Corporations and E-commerce Companies.
- Media Companies including publishers and game studios.
- Telecom Providers.
- Sports leagues for their digital surfaces.
- Advertising and Marketing Firms.
- Financial Services Firms.
- Healthcare Companies.
- Automotive Companies.

We outline our key observations below:

(1) Out of the hundreds of Microsoft clients the Recommenders team has spoken to, 90-95% of those with an existing recommendation system in production have very simple architectures.
(2) There's a rising trend in customer requests for conversational recommendation systems across industries.
(3) Business executives, who are often the key decision-makers for investing in a recommendation system, want personalization on their solutions but generally lack a deep understanding of how these systems work or how to assess their potential value. This poses a major challenge for quick adoption.
(4) Decision-makers often do not have a clear understanding of the type and level of investment required to achieve a strong return on investment (ROI) from building and deploying a recommendation system for their specific use case. One of our most important learnings is that as recommendation practitioners, we must speak the same language as key business decision makers and set the right expectations for adoption to be successful.

(5) Business decision-makers often want to see fast outcomes to justify their investment, making addressing the cold-start challenge critical.
(6) Most systems that customers choose to adopt are batch recommendation systems, which generate recommendations periodically offline rather than in real-time. Batch recommendations served through Azure CosmosDB is effective and it is very uncommon for Microsoft clients from the aforementioned industries to deploy sophisticated systems like real-time or hybrid recommendation models.
(7) There is an emphasis on ethical AI and transparency in recommendation systems. Customers want to know how companies use their data and the flexibility to opt out.

4 Personalized Recommendations with Limited Data

4.1 LLMs for cold-start: leveling the playing field

For addressing the cold-start (and data sparsity) in real-world recommendation systems, LLMs have been shown to be an excellent approach to infer user preferences and help initialize cold-start items [6, 5].

Another approach to achieve personalization is to leverage LLMs to generate offline high quality user profile summaries from a user's historical engagements (clicks, views, orders) and inferred preferences [8]. Depending on how the summarization is engineered, these summaries can capture both long-term and short-term interests and serve as strong features for ranking models as well as for dense retrieval by using the LLM embedding of the user profile.

As LLMs become cheaper and faster for inference, online reranking is becoming increasingly feasible. LLMs, given their world knowledge, can provide strong ranking performance given the item text features in the zero and few-shot setting, in the absence of engagement data. With sparse engagement data for a newly launched recommendation system, a LLM ranker can offer well-reasoned and highly personalized recommendations given user interests, represented as:

(1) A user profile, summarized from previous engagements
(2) Embeddings from models that learn from engagement signals. For example, the user embedding can be learned using Collaborative Filtering from user-item engagement signals, projected to the token embedding space and fed to the ranking prompt of the LLM ranker [3].

Experiments at Bing have shown that as the recommendation system matures, Supervised Finetuning (SFT) and/or preference alignment with Direct Preference Optimization (DPO) on collected engagement logs are critical for adapting an LLM ranker to the user behavior specific to the recommendation system to achieve gains on top of a traditional cross-encoder model.

However, companies often may not want to bear the cost of using LLMs for online ranking. At Bing, we've found knowledge distillation to be a promising approach for achieving efficiency whilst preserving gains: leveraging a finetuned LLM as a teacher model and distilling to a compact student cross-encoder model.

4.2 Engagement data from other surfaces

A common observation is that some clients already have a search system installed on their surface. Search logs can be used effectively to alleviate the cold-start problem and build a performance recommendation system. For example, a useful approach that we leverage at Bing Webpages Recommendations is to collect co-occurrence clicks of webpages in the same Bing search browsing session. When two items co-occur across multiple user search browsing sessions, they are likely to be good recommendation candidates for each other. For our recommmendation system at Bing webpage recommendations, we use the webpage co-occurrence click count as a retrieval path and also training a Collaborative Filtering model using the co-occurrence click matrix. Furthermore, clicked items from search logs can be linked through semantically similar or normalized queries to produce item-item engagement data. For user personalization, the per-user engagements from search sessions can be used to initialize a user profile.

4.3 Meta-learning

Recently, meta-learning-based approaches have gained traction, deployed for use cases from retrieval to ranking and click-through rate (CTR) prediction. At KDD 2024, LiMAML [7] from LinkedIn was presented as an scalable method to achieve task personalization using meta-learning. This approach used the last few item interactions for each user to offline update, on a daily recurrence, user meta-embeddings. These meta-embeddings were served as a feature for the online ranking model, resulting in significant online improvements (in CTR and related engagement metrics, as well as Weekly Active Users) across both user segments: new LinkedIn users with minimal interactions and for regular users.

4.4 LLM-enabled Dense Retrieval

Leveraging smaller LLMs, such as LLama-7b, offers a promising approach to achieving robust retrieval across a large item corpus, particularly in the scenarios where engagement data is sparse or unavailable.

RepLLama [4] uses Llama-2-7b as the backbone model for two-tower dense retrieval, using the representation of the end-of-sequence token as the representative embedding. The model is then optimized with Info-NCE loss. RepLLaMA exhibits strong retrieval performance in the zero-shot setting, underscoring thier effectiveness for nascent systems.

Another approach we proposed borrows from synthetic query generation in the search domain [2]. LLMs can be used through an offline pipeline to generate and store synthetic topic and content tags from the candidate content. This enables keyword-based retrieval using traditional methods such as BM25.

5 Discussion Topics

At Introspectives, we would like to drive discussion around the following topics:

(1) What are main bottlenecks for successful adoption of recommendation systems by businesses and corporations?
(2) How can we better communicate and align a recommendation system to the goals of a business?

(3) What are some key technical and non-technical challenges for launching a recommendation system on a new business surface?
(4) Are LLMs leveling the playing field? Will LLMs drive a new wave of adoption?
(5) What new recommendation experiences can LLMs enable that were previously too challenging for businesses starting out? (e.g. Conversational Recommenders)
(6) What are some other areas of research that alleviate the data sparsity challenge and drive successful business outcomes?

CCS Concepts

• **Information systems → recommendation systems**.

Keywords

LLMs, Online and Offline Evaluation, Recommendation Control

ACM Reference Format:
Jaidev Shah and Miguel González-Fierro. 2024. Adoption of recommendation systems: Observations, Trends and Leveling the Playing Field . In *Proceedings of Conference on recommendation systems (RecSys '24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

References

[1] Andreas Argyriou, Miguel González-Fierro, and Le Zhang. "Microsoft Recommenders: Best Practices for Production-Ready Recommendation Systems". In: *Companion Proceedings of the Web Conference 2020.* WWW '20. Taipei, Taiwan: Association for Computing Machinery, 2020, pp. 50–51. ISBN: 9781450370240. DOI: 10.1145/3366424.3382692. URL: https://doi.org/10.1145/3366424.3382692.

[2] Akshay Jagatap, Srujana Merugu, and Prakash Mandayam Comar. "Improving search for new product categories via synthetic query generation strategies". In: *The Web Conference 2024.* 2024. URL: https://www.amazon.science/publications/improving-search-for-new-product-categories-via-synthetic-query-generation-strategies.

[3] Sein Kim et al. "Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '24. Barcelona, Spain: Association for Computing Machinery, 2024, pp. 1395–1406. ISBN: 9798400704901. DOI: 10.1145/3637528.3671931. URL: https://doi.org/10.1145/3637528.3671931.

[4] Xueguang Ma et al. *Fine-Tuning LLaMA for Multi-Stage Text Retrieval.* 2023. arXiv: 2310.08319 [cs.IR]. URL: https://arxiv.org/abs/2310.08319.

[5] Scott Sanner et al. "Large language models are competitive near cold-start recommenders for language-and item-based preferences". In: *Proceedings of the 17th ACM conference on recommender systems.* 2023, pp. 890–896.

[6] Jianling Wang et al. *Large Language Models as Data Augmenters for Cold-Start Item Recommendation.* 2024. arXiv: 2402.11724 [cs.IR]. URL: https://arxiv.org/abs/2402.11724.

[7] Ruofan Wang et al. *LiMAML: Personalization of Deep Recommender Models via Meta Learning.* 2024. arXiv: 2403.00803 [cs.IR]. URL: https://arxiv.org/abs/2403.00803.

[8] Fan Yang et al. "Palr: Personalization aware llms for recommendation". In: *arXiv preprint arXiv:2305.07622* (2023).