Cornell Bowers C·IS
College of Computing and Information Science

Off-Policy Learning for Diversity-aware Candidate Retrieval in Two-stage Decisions

Rayhan Khanna (Advised by Professor Thorsten Joachims and Haruka Kiyohara)
Cornell University

Introduction / Key Takeaways

What are two-stage decision systems?

- Used in real-world applications like:
  - **Search engines**: first select relevant results, then rank them
  - **Recommendation systems**: retrieve a few candidate items, then decide what to show
  - **Retrieval-augmented generation (RAG)**: first pull documents, then generate a response (question -> document search -> response)
- Allows us to **narrow items to a top-k list** then **select or rank** the best → efficient pipeline

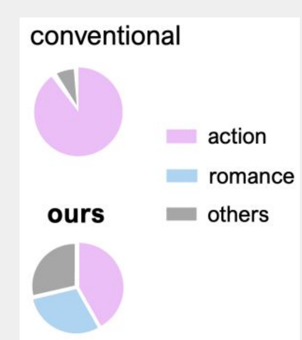
The problem:

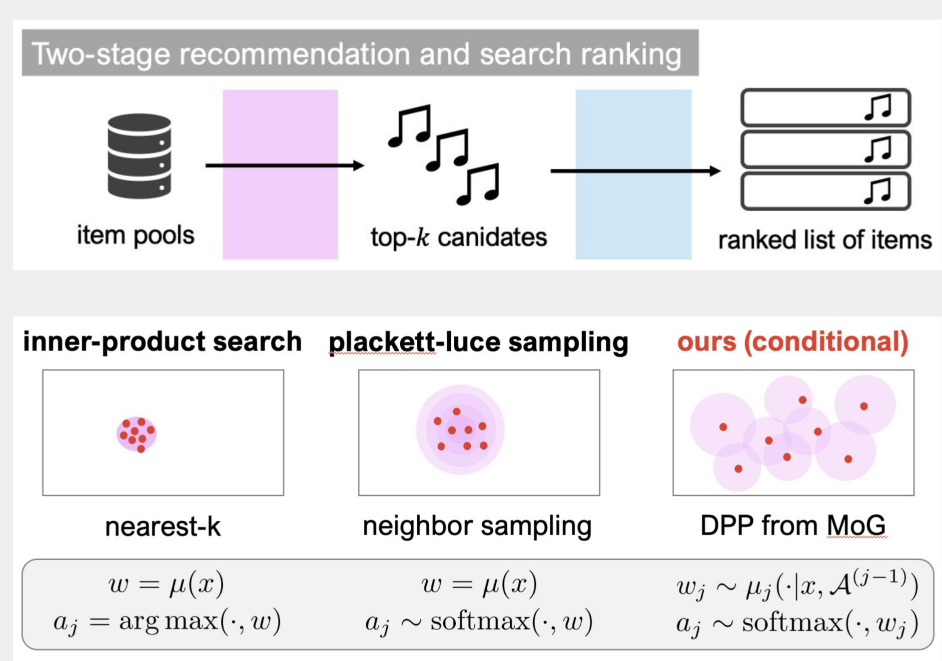
- Most current systems:
  - Use **simplified models** (like collaborative filtering) to **retrieve candidates**
  - Assume users have **one preference** (like only action movies)
  - **Optimize** using (logged) **historical data** (clicks, purchases, etc.), which is **biased** and **sparse**
- This leads to **low diversity** in what's **retrieved** (similar recommended items)

Methods

Sampling Diverse Candidates

- We use a **two-stage sampling process**:
  - **First**, sample a diverse set of user preference vectors
  - **Then**, retrieve items based on each preference, ensuring variety across topics/types





Conceptual comparison between the proposed method and conventional approaches and the resulting proportions of categories in candidates. While the baselines represent a single preference per context, our sampling process simulates a more complex, Determinantal Point Process (DPP) sampling from a mixture-of-Gaussian (MoG) distribution. This helps model users' multiple and distributional interests such as preferring action movies for 45% of time and romance movies for 30% of time.

Kernelized Sampling

- Because there's bias/sparsity in historical/logged data, we use **Kernel Importance Sampling (IS)** to overcome this
  - It shares reward signal across similar items, reducing variance and improving performance

$$\widehat{V}_{\mathrm{KIS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{K(y, y_i; x_i, \tau)}{\bar{\pi}_0(\psi(y_i) \mid x_i)} r_i$$

$$\widehat{V}_{\mathrm{IS}} = \frac{1}{n}\sum_{i} \frac{\pi_{\theta_2}(a_i \mid x_i, A_i^k)}{\pi_0(a_i \mid x_i)} r_i$$

Modeling Multiple User Preferences

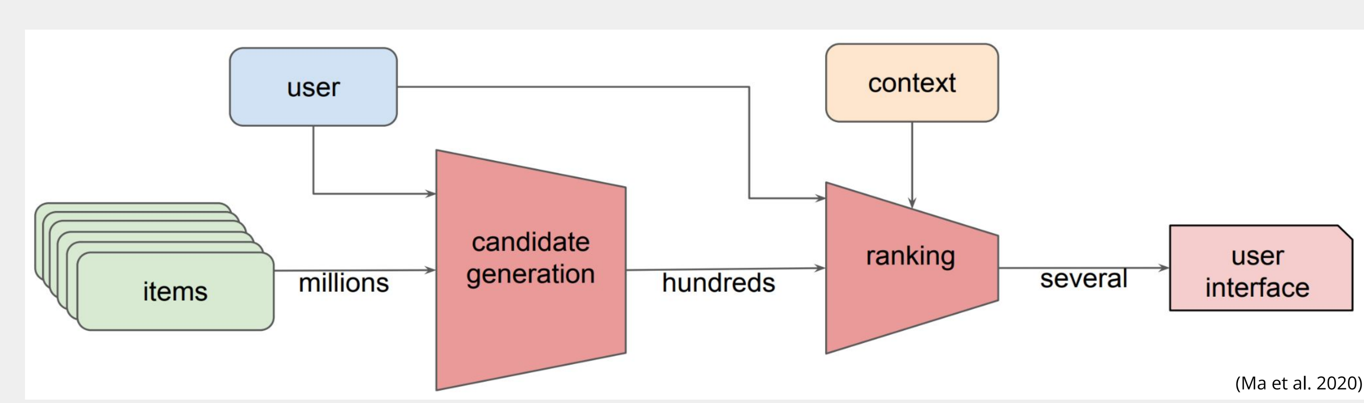
- Instead of assuming a user has one interest (ex. sports), we model a **multi-modal distribution**, allowing for mixed interests
- This enables retrieval of diverse items in the candidate set

Synthetic Experiment Setup

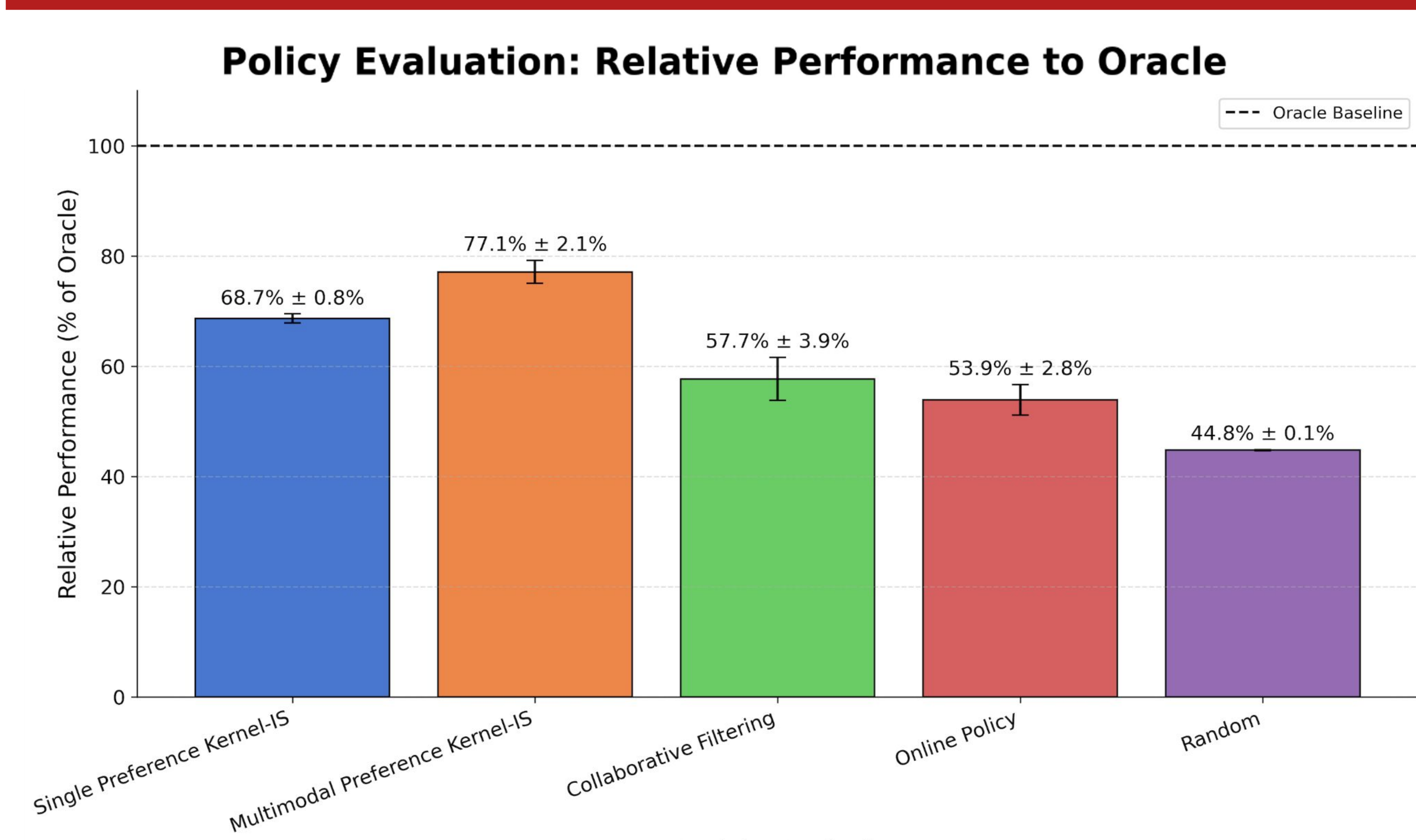
- **Setup**:
  - Synthetic bandit with **1,000 users** and **10,000 items**
  - **Two-stage setup**: get top-10 candidates → rank top-5
  - Logged feedback generated from user-item model
- **Evaluation:**
  - Off-policy learning over **five seeds**
  - Simulated online evaluation via ground-truth reward

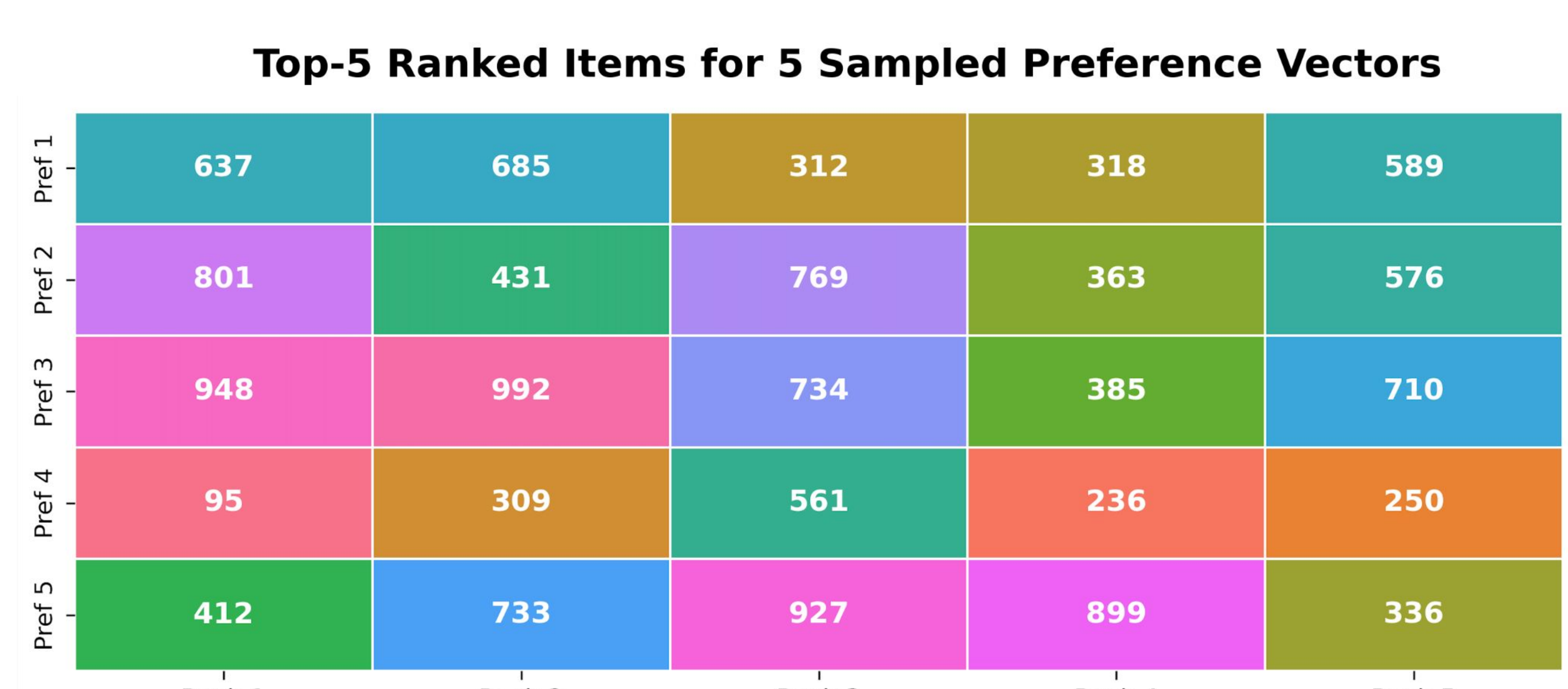
Project Goals

- We aim to design a **data-efficient off-policy learning framework** that:
  1. Models multiple user preferences (multimodal)
  2. Selects diverse candidate sets tailored to varied interests
  3. Optimizes for user engagement signals (ex. view time)
  4. Learns from logged feedback through Kernel-IS, avoiding risky live tests



Results





Future Work

- **Document summarization:**
  - Document selection = 1st stage
  - Summary generation = 2nd stage
  - Logged LLM/human summaries + BERTScore = bandit feedback to train policies

References

Kiyohara, H., Khanna, R., & Joachims, T. (2025). Off-policy learning for diversity-aware candidate retrieval in two-stage decisions. In *ICML 2025 Workshop on Scaling Up Intervention Models*.

Ma, J., Zhao, Z., Yi, X., Yang, J., Chen, M., Tang, J., Hong, L., & Chi, E. H. (2020). Off-policy learning in two-stage recommender systems. In *Proceedings of The Web Conference 2020 (WWW '20)* (pp. 463–473). ACM. https://doi.org/10.1145/3366423.3380130