The Hong Kong University of Science and Technology

Dept of Information Systems, Business Statistics and Operations Management

OM Seminar

Near-Optimal Real-Time Personalization with Simple Transformers

by

Mr. Lin AN

Carnegie Mellon University

Date	:	**28 November 2025 (Friday)**
Time	:	**2:30pm – 3:45pm**
Venue	:	**Meeting room 4047, LSK Business Building**

Abstract:

Real-time personalization has advanced significantly in recent years, with platforms using machine learning models to predict user preferences from rich behavioral data. Traditional embedding-based models reduce real-time recommendation to nearest-neighbor search, which is extremely fast, but they struggle to capture complex user behaviors that matter for accuracy. Transformer-based models overcome these limitations by modeling sequential behavior, but their architectures make the downstream optimization problem challenging.

We focus on a specific class of transformers, simple transformers, which contain a single self-attention layer. We show that simple transformers can represent complex user preferences such as variety effects, complementarity and substitution effects, and irrational choice behaviors, while remaining far more tractable than deeper architectures. We then present an efficient real-time personalization algorithm under simple transformer models that achieves near-optimal performance with sub-linear runtime in the size of the item pool.

Finally, we discuss our collaboration with Glance, a lock-screen content platform serving over 400 million users in Asia. In an A/B test on 200,000 high-activity users, increasing variety in line with our model produced a 3.7% increase in valuable session count, a key retention metric.

Bio:

Lin An is a fifth-year PhD candidate in the Algorithms, Combinatorics, and Optimization program at Carnegie Mellon University's Tepper School of Business, co-advised by Andrew A. Li and Benjamin Moseley. His research centers on how to use AI to make better operational decisions, with a particular focus on combining optimization and modern AI-based models. His primary areas of application include recommendation systems, resource allocation, and inventory management.

Website: https://www.andrew.cmu.edu/user/linan/

All interested are welcome!

Enquiries: Dept of ISOM