This thesis seeks to address the challenge of making adaptive accuracy-cost balancing inside systems for large-scale machine learning-based recommendation services. We show that it is important to make performance tradeoff decisions at a per-query basis instead of a predefined policy for all queries. We show that we can achieve a better tradeoff between accuracy and cost by leveraging lightweight machine learning models to make more adaptive decision-making inside systems infrastructure.

Large-scale recommendation services have two computation-heavy components with strict accuracy and latency targets: scoring (typically achieved by complex machine learning models) and candidate retrieval (typically achieved by approximate nearest neighbor search). We first introduce a caching system for scoring component in recommendation systems (in particular search advertising systems). Inside the cache, we leverage lightweight machine learning models to make adaptive cache refresh decisions, which provides a better balance between recommendation accuracy and computation cost. This leads to a better net profit in the search advertising context. We then present the learned adaptive termination for approximate nearest neighbor search inside the candidate retrieval component. We leverage lightweight machine learning models to decide how much to search for each query, which provides a better balance between the search accuracy and latency (computation cost).

In the proposed work, we seek opportunities to further improve the learned adaptive termination for approximate nearest neighbor search. We plan to fully automate the training and tuning of the lightweight machine learning models, introduce additional features and strategies, and explore the possibility of neural network models.

Thesis Committee:
David G. Andersen, Chair
Michael Kaminsky
Rashmi K. Vinayak
Yuxiong He, Microsoft AI and Research

Thesis Summary:
https://drive.google.com/open?id=1mGf3cPNjdV4ZGTYQMYLQWqJDP-YVIBAi