ABSTRACT

In this work, we explore the application of Policy Gradient Reinforcement learning for ranking in search systems and recommender systems. With the rapid increase of data on the internet in recent years, it is essential that users receive the items/documents that are most relevant to their needs. As a result, not only must the search results or items be relevant, but the most relevant results should be at the top of the list. Ranking is the process of arranging a list of items or examples in a specific order based on their relevance, importance, or some other criteria. Ranking is a common task in various machine learning applications, and it is used in fields such as information retrieval, recommendation systems, natural language processing, and more. Ranking in search systems and recommender systems is the task of determining the order in which items, such as web pages, products, or content, are presented to users based on their relevance to the user's query, preferences, or behavior. This process is crucial for search engines, e-commerce websites, content streaming services, and many other applications where user engagement and satisfaction are important. This thesis addresses key challenges in ranking for search systems and recommender systems, such as scalability, utilization of shared information between documents and addressing high variance in existing RL models for ranking. We address these issues and propose models to improve the efficiency and effectiveness of ranking for search systems and recommendation systems.

In recent times, different machine learning approaches have been sucesfully applied for ranking in different domains. Reinforcement Learning provides a powerful framework for addressing many of the inherent challenges in ranking, particularly in terms of personalization, handling sparse feedback, and optimizing for long-term outcomes.

By continuously learning from user interactions and adapting rankings in real-time, RL-based models can significantly enhance ranking systems across various domains. Reinforcement learning (RL) is a type of machine learning paradigm where an agent learns to make decisions by interacting with an environment. RL has been successfully applied in various domains, including robotics, game playing, finance, healthcare, and autonomous systems. The field continues to evolve with ongoing research and the development of new techniques and applications. Deep RL combines RL algorithms with deep neural networks to handle complex and high-dimensional input data.

High variance in reinforcement learning (RL) models, especially in ranking tasks, poses challenges that hinder performance and stability. This variance stems from the stochastic nature of environments and the randomness in policy actions. In learning-to-rank systems, noisy or biased click data can lead to misleading gradient estimates, making the learning process erratic and inefficient [35]. This can lead to slow convergence due to the need for more samples to achieve reliable gradient estimates, which is computationally costly [5], and overfitting, where the model captures training data patterns that do not generalize well, degrading live performance [136]. Additionally, variance amplifies biases, such as position bias, where rankings disproportionately influence user interactions, potentially reinforcing suboptimal behaviors [131]. In this thesis, we utilized the Markov Decision Process framework for the ranking model in search and recommender systems. We present three Policy Gradient Reinforcement Learning based algorithms in our work. Firstly, we proposed a Deep Policy Gradient based algorithm for large scale ranking task for the search system. RL-based approaches have been applied for the ranking task effectively, but the existing Policy based methods suffer from noisy gradients and high variance, resulting in unstable learning. The natural policy gradient algorithms such as REINFORCE perform Monte Carlo based sampling, thus taking samples randomly, which leads to high variance. With action space becoming increasingly large, i.e., with a very large number of documents, traditional Reinforcement learn-

ing approaches lack the complex model required in the scenario to deal with a large number of items. We propose a Deep Reinforcement learning based method for the ranking task in this work to address these issues. By combining Deep learning with the Reinforcement Learning framework, our approach can learn a complex function as deep neural networks can provide significant function approximation. We utilized the actor-critic method in our model, where the critic network can effectively reduce variance by using different approaches like clipped delayed policy updates, clipped double q learning, etc. We employed the TD3 approach to train the RL agent with a listwise loss function, which executes delayed policy updates, resulting in lower variance value estimates. We conducted experiments on the different Letor datasets for various ranking metrics and showed that our method outperforms various state-of-the-art baselines.

Next, we propose a Multi-Agent Reinforcement learning for large scale datasets, extending our previous work to multi-agent settings utilizing the shared information within different items. The existing ranking methods for document search ranking overlook the correlation between the documents by ignoring the shared information between different documents. We utilize the multi-agent system setting to capture the correlation between the documents by sharing the information among multiple agents during learning. This is executed through a centralized critic frame- work that has access to global information. Also, variance and noise in RL-based approaches is aggravated in large scale datasets having millions of documents. We address these issues utilizing the Deep RL actor-critic framework that learns the value estimates directly from the samples. Deep policy gradient based approaches have proven to be of significance in scenarios/problems dealing with large action space(very large number of items) through powerful approximation capabilities of neural networks. We also conducted experiments on the two large-scale Microsoft Letor datasets for different ranking metrics and showed that our method outperforms various state-of-the-art baselines.

Lastly, we propose an algorithm for large scale recommender systems providing the

list of top items users might be interested in. Recommender systems use various algorithms to determine the top-ranked items for a user based on their preferences, historical interactions, and other relevant data. The goal is to present users with personalized recommendations that are likely to be of high interest. However, as the item space and user base grow in size, scalability remains a critical concern for recommender systems. Most of the existing policy gradient techniques in recommender systems suffer from high variance, which increases instability throughout the learning process. Policy Gradient methods, such as Proximal policy optimization, have proven to be effective in problems with large action spaces as they learn the optimal policy directly from the samples. In this work, we model collaborative filtering problem through Markov Decision Process(MDP) framework for training the RL agent. Proximal Policy Optimization approaches are today recognized amongst the most effective reinforcement learning techniques, delivering state-of-the-art performance and even outperforming Deep Q learning methods. We propose a switching hybrid recommender system in this work that utilizes two distinct recommender system approaches. We show that our method outperforms various baseline methods on the popular Movielens datasets for different evaluation metrics. In our work, we performed extensive experiments on various datasets for different ranking metrics to demonstrate the effectiveness of the proposed methods over the existing approaches. One limitation of our work is explicit data required for the training, i.e the labelled dataset used in the search system and recommender system problem in our work. This can be a barrier especially in environments where obtaining data is expensive or time-consuming. One way to address the above issue could be be incoprporating more implicit data for train ing such as click based data, browsing histories, search queries, etc.

In the future, our work can be extended to learn from implicit feedback mechanisms such as from the clicks for search and recommendations rather than the strict requirement of having the specified ground truth labels. The click based data is comparatively less expensive and more extensively available compared to explicit

feedback. Also, we can incorporate more diversified search results in the ranking task. Further, more recent advances in Reinforcement learning research, such as inverse reinforcement learning framework, adversarial RL, etc., can be used to model the ranking problem.