Chapter 6

Discussions

We further analyze and elaborate on the major contributions of this thesis in this section. The fundamental goal of this work was to analyze and improve upon the existing Reinforcement Learning based approaches for ranking tasks in large scale datasets for search and recommender systems. In recent years Reinforcement learning has been succesfully applied to various domain such as marketing, healthcare, advertisment, news recommendations, network routing, load distribution, autonomous vehicles and so on. The basic concept of reinforcement learning is easy to grasp: utilize feedback to reinforce and thus strengthen favourable outcomes. Reinforcement learning, as opposed to making major changes infrequently, makes frequent incremental updates. There are numerous advantages to this, including constant improvement of results and faster identification of new potential outcomes. Effective decision making entails adapting past experiences and pertinent contextual knowledge to a new scenario. The primary paradigm in deep reinforcement learning is for an agent to develop gradual knowledge that aids decision making into its network weights through gradient descent optimization.

- In the 1st work, we addressed the issue of ranking in large scale datasets with policy gradient based approaches. With the multifold increase of data over the internet in recent years, it has become imperative that users should be re-

turned the items/documents that are most relevant to their needs. Therefore, the search results or items have to be not only relevant but also most relevant results should be at the top of the list. Further, with the increasing size of item space and users, scalability remains a key issue. Recently, reinforcement learning methods have been applied effectively to information retrieval tasks, however traditional Reinforcement learning algorithms suffer from lack of complexity required when the state space becomes extremely large with millions of items. Further, noisy gradients and increase in variance is another in policy gradient RL algorithms with increasing item space.

We thus, propose an MDP based approach utilizing Policy gradient algorithms and Deep Reinforcement learning techniques which can be effective for very large item space and addressing the issues in existing RL based algorithms. The powerful function approximation abilities of the Deep NN structure mitigates the need to have traditional RL tabular calculations. Further, we used the policy gradient framework and thus, the agent need not calculate the value funnction for all the actions in a state, which can easily become intractable in problems with large action space such as ranking. We propose an algorithm DRLRANK, based on the state-of-the-art algorithm twin-delayed DDPG (TD3) [37], to improve on these limitations described above by training the RL agent using techniques such as clipped double-q learning, delayed policy updates, and so on. It smooths the target policy using noise regularisation to reduce variance. Furthermore, TD3 employs clipped double Q learning, in which the smallest value of the two critic networks is used to underestimate Q values. This, together with the delayed update, leads in a more stable approximation with reduced bias.

The following summarizes our work's contribution in brief:

We present a Deep Reinforcement Learning-based method for learning to rank problems. We address the challenges associated with other Reinforcement

Learning approaches, such as high variance and noisy gradients, with the state-of-the-art TD3-based Learning to Rank algorithm, which employs techniques such as delayed policy updates, clipped double q learning, and so on. We merge the Reinforcement learning paradigm with deep learning; as deep neural networks may provide significant function approximation, our model can learn a complex function. Furthermore, we adopted a policy-based technique for large-scale ranking since policy gradient algorithms have been successfully applied to problems with large action spaces (a large number of items) as they do not rely on computing a value for each item as in value-based methods. We performed extensive experiments on the Letor datasets for different evaluation measures such as NDCG and MAP. Experimental results demonstrate our proposed approach is able to achieve better performance as compared to various state-of-the-art baselines.

- In our next work we proposed a Muti Agent RL based approach for the large scale ranking task extending our initial work in the multi agent settings. While existing LTR methods utilize the document attributes for the ranking task, they overlooks the correlation between the documents by overlooking shared information between them. In multi agent setting, multiple learning agents each having it's own policy focus on to solve a problem in a shared environment. We utilize the Multi-Agent framework to capture the correlation between documents by sharing information among multiple agents, with the centralised critic framework having access to this global information. The query document pair information is used by most existing Learning to Rank algorithms, but the shared information between the documents is ignored. In this work we propose a ranking model for large scale with Multi-Agent Reinforcement Learning framework to extract and utilize the correlation between documents. The agents share their observations about the environment in an attempt to learn a coordinated model. Further, sharing the global state across agents increases correlation, which is necessary for learning a coordi-

nated strategy. Consequently, by sharing information about the documents across the different agents, we presume that different agents would learn the link between the documents and the coordination of their policies. We propose a Multi-Agent RL algorithm for Learning to Rank based on the Multi-Agent DDPG(MADDPG) algorithm. The multi-agent design offers significant benefits such as coordinated activities and stability via a centralized critic. In this work we utilized the Centralized Training with Decentralized Execution (CTDE) framework for multi agent systems which has been effective for challenges such as scalability and partial observability. The main idea behind CTDE framework is to supply more extensive state information to agents during learning so that their individual experiences can be aggregated through centralization of training, allowing efficient training for the agents. This approach enables agents to develop optimal coordination rules among themselves without developing a full decision model of the environment. o the best of our knowledge, this is the first Multi-Agent Deep reinforcement learning based approach applied for document retrieval in LTR. In contrast to using a single-agent framework, learning of numerous agents in a shared environment has proven to be more beneficial due to the extra knowledge gained from estimating the policies of other agents. We utilize CTDE framework for the learning where multiple agents can share the parameters of a single value network and a replay buffer, thereby reducing the algorithm's policy search space. We also conducted experiments on the two large scale Microsoft Letor datasets, MSLR-Web 10k and MSLR-Web 30k, and showed that our method outperforms various state-of-the-art baselines.

- Next, we proposed a Proximal Policy Optimization based Hybrid Recommender system for large scale datasets. The first goal was to effectively apply and improve upon the existing work for Reinforcement Learning based recommender systems. Modern recommender system incorporates millions of item Access to the large labeled dataset has been a huge bottleneck for super-

vised learning methods. Deep reinforcement learning based approaches have been significantly effective for different real life scenarios owing to the powerful approximation capabilities of the deep neural network architecture. The value-function approach has been the dominant approach conventionally for RL, in which all function approximation effort is directed towards estimating a value function, with the action-selection policy represented implicitly as the "greedy" policy with respect to the estimated values. The value-function technique has been successful in many applications, but it has some drawbacks. For instance, it is aimed towards determining deterministic policies, although the optimal policy is oftentimes stochastic, i.e., selecting actions with different probabilities.

However, one key issue with algorithms like REINFORCE is that their high variance in gradient estimations leads to slower convergence. Several approaches to reducing this variance have been proposed, but none of them address the underlying reason, however continuous sampling from a probabilistic strategy introduces noise into the gradient estimate at each time-step. The fundamental issue is the existence of large state-spaces, which makes the use of traditional tabular methods impractical.

Our contributions are summarized below:

We propose a RL-based hybrid recommender system for a large action space, i.e., a very large set of items. We formulated the recommender system problem as a Markov Decision Process propose a RL-based hybrid recommender system for a large action space, i.e., a very large set of items, training our agent with the state-of-the-art Proximal Policy Optimization method. As discussed above and in more detail in Section 5.1, the PPO-based algorithm reduces the high variance and increases stability. We also addressed the cold start issue in our approach with a switching Hybrid recommender configuration. We classified users as cold or hot on the basis of ratings they provided. Fi-

nally, we show the effectiveness of the proposed approach on the two popular datasets : Movielens 1m and Movielens-100k, for different evaluation metrics. We demonstrated through experiments the improvements for different evaluation metrics such as Precision and Recall. In our approach we can see the performance of agent stabilizes in the reward distribution after enough episodes, which we may attribute to the Policy gradient approach obtaining stability in convergence due to more sample data associated with the dataset. Further, with the hybridized setting we also addressed the cold start issue when the user does not have enough ratings associated to be provided with optimal recommendations.