Adapting to Dynamic User Preferences in Recommendation Systems via Deep Reinforcement Learning

TU Delft Delft University of Technology

Author
Luca Pantea <P.L.Pantea@student.tudelft.nl>

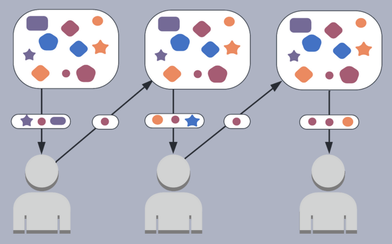
Supervisors
Dr. Frans Oliehoek <F.A.Oliehoek@tudelft.nl>
Dr. Aleksander Czechowski <A.T.Czechowski@tudelft.nl>
Davide Mambelli <D.Mambelli@tudelft.nl>
Oussama Azizi <O.Azizi@tudelft.nl>

1. Problem Statement

The repeated interplay between users and recommender systems often creates feedback loops that result in recommendations increasingly tailored to the user's preferences.

This can result in:

- Influence on user behaviour
- Adversity in learning the user's preferences over extended horizons.



2. Research Objective

1. Benchmark two Deep Reinforcement Learning: Generic *FullSlateQ [Sunehag et al. 2015]* and *SlateQ [Ie et al. 2019b]* under preference dynamics.
2. Observe if the algorithms are capable to optimize for long-term user engagement and cumulative satisfaction.

3. Our Contribution

- Model recommendation task as a *slate-MDP*.
- Construct a simulated environment with various degrees of preferential dynamics [*Top right plot*].
- Provide an in-depth analysis of the performance of *FullSlateQ* and *SlateQ* over extended horizons.

References

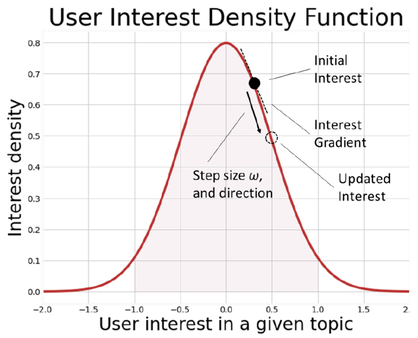
Eugene Ie, Vihan Jain, JingWang, Sanmit Narvekar, Ritesh Agarwal, RuiWu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, Jim McFadden, Tushar Chandra, and Craig Boutilier. 2019b. Reinforcement Learning for Slate-based Recommender Systems: A Tractable Decomposition and Practical Methodology. CoRR abs/1905.12767 (2019) arXiv:1905.12767 http://arxiv.org/abs/1905.12767

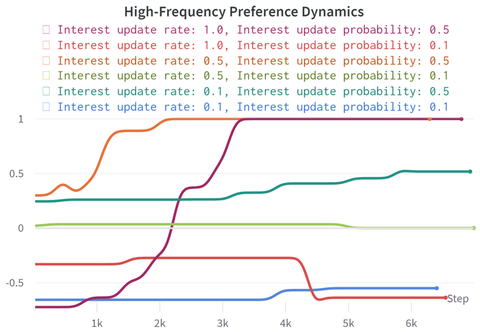
Peter Sunehag, Richard Evans, Gabriel Dulac-Arnold, Yori Zwols, Daniel Visentin, and Ben Coppin. 2015. Deep Reinforcement Learning with Attention for Slate Markov Decision Processes with High-Dimensional States and Actions. CoRR abs/1512.01124 (2015). arXiv:1512.01124 http://arxiv.org/abs/1512.01124

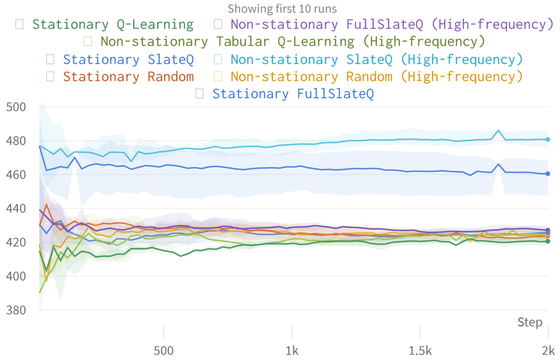
Eugene Ie, Chih-Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019a. RecSim: A Configurable Simulation Platform for Recommender Systems. CoRR abs/1909.04847 (2019). arXiv:1909.04847 http://arxiv.org/abs/1909.04847

4. Methodology

- We make use of *RecSim* [Ie et al. 2019a] for creating stylized user, document and user response models.
- We assume a *multinomial proportional choice model* for defining the user choice behaviour.
- We introduce user engagement and satisfaction latent user attributes to measure the effectiveness of long-term value optimization.
- We adopt two models of user preference dynamics:
    1. Function-based interest evolution. [*Left plot*]
    2. Session termination.







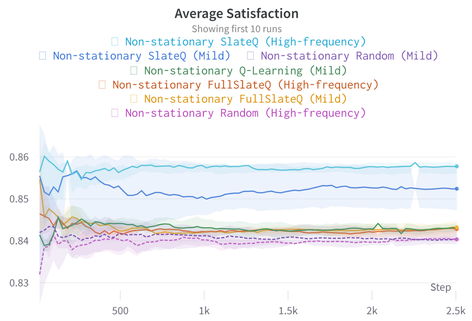
5. Results

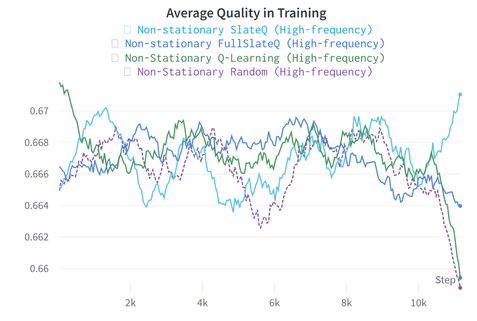
Research Objective 1:

- *SlateQ* outperforms *FullSlateQ* in both stationary and non-stationary environments. On average, 10.57% better than FullSlateQ in dynamic environemnts.
- Remarkably, *SlateQ* performs better under high-frequency preference dynamics than in stationary environments [*Left plot*].
- With slate size 3, *FullSlateQ* takes approximately 6X the training time of *SlateQ*.

Research Objective 2:

- Both *FullSlateQ* and *SlateQ* behave *myopically*, thus failing to learn a higher quality recommendation policy, which further maximizes long-term user engagement. [*Right plot*]
- Satisfaction stays rather high for both algorithms, yet *FullSlateQ* performs similarly to the random baseline. [*Left plot*]





6. Conclusion

- *SlateQ* offers notable improvements (10.57%) in user engagement compared to *FullSlateQ* in dynamic environments.
- *SlateQ* renders RL tractable with slates, through decomposing slate Q-values into Q-values for individual items, and thus practical in commercial applications.
- Both *SlateQ* and *FullSlateQ* fail to make a suitable tradeoff between guiding the user's preferences towards higher-quality documents at the expense of temporarily diminishing the user budget.