Supplementary Material of A Reinforcement Learning Approach for Personalized Diversity in Feeds Recommendation

Li He[1], Kangqi Luo[2], Zhuoye Ding[2], Hang Shao[3], and Bing Bai[1]

[1] Tsinghua University, 1632226712@qq.com, baibing12321@163.com
[2] JD.com, luokangqi@jd.com, dingzhuoye@jd.com
[3] Zhejiang future technology institute (Jiaxing), shaohang@zfti.org.cn

Appendix for Greedy-based Re-ranking. In our recommender system, k is equal to 10 and $n >> k$. Our used re-ranking module outputs each item of the final list in order, and a greedy-based search method [1] is applied to select the best candidate item at each step. At the first step, we directly select the one with the highest relevance score. Afterwards, taking the i-th step as an example, each candidate item is scored by jointly considering its three factors: 1) $Score_{rel}$, the point-wise relevance score which ranges from 0 to 1; 2) $Score_{nov}$, the novelty score with respect to the previous $i-1$ output items; 3) $Score_{penalty}$, a penalty score indicating whether the candidate item together with the previous $i-1$ output items conflict with a pre-defined diversity rule. Specifically, the score $Score_{nov}$ is defined as the Kullback-Leibler (KL) divergence [4] between the product category distribution of the candidate item and the product category distribution of the previous $i-1$ output items. In terms of $Score_{penalty}$, the corresponding diversity rule is defined at the category level, for example, for any N consecutive products, at most M products belong to the same category. If the candidate item together with the previous $i-1$ output items conflict with the pre-defined diversity rule, the $Score_{penalty}$ is set as 1. Otherwise, the penalty is set as 0. This kind of diversity rule serves as a soft constraint, which strongly penalizes the candidate item staying too close to other items in the same category, thus avoiding bad user experience. Each candidate item at the i-th step is assigned with a weighted re-ranking score according to the following scoring function:

$$Score_{rerank} = \alpha * Score_{rel} + (1-\alpha) * Score_{nov} - Score_{penalty}, \qquad (1)$$

where α is a hyper-parameter. Finally, the item which has the highest weighted re-ranking score $Score_{rerank}$ is selected as the best candidate item for the i-th step. Similar processes are applied for other steps.

Appendix for MDP Formulation of Personalized Diversity Process. We describe our elements design more specifically.

State. We use a tuple (u, req_t, c_t, e_t) to represent the state s_t, where user information u contains the user's basic profile such as age, gender, average browsing depth in history, and the frequency of visits in a long period. This is used to

describe the user's long-term preference. The term req_t denotes contextual information of the current request, including the time of the request (e.g. in which hour), whether it is a weekday or a weekend, and the time interval since the last click, add-to-cart, or purchase behavior. The click sequence c_t keeps track of the click behaviors in a short period (*e.g.*, within 1 hour), and may contain behaviors of previous sessions. For products in c_t and e_t, we take their product profiles into account, such as categories and brands. We use click sequence c_t and exposure sequence e_t to reflect the user's short-term intents and interests.

Action. To achieve personalized diversity, a set of candidate diversity rules need to be designed in advance as the action space. Since users are easily aware of whether the recommended items belong to the same category or not, we follow a greedy-based approach and design candidate rules at the category level. Based on the definition provided in the section of greedy-based re-ranking, a diversity rule can be defined as for any N consecutive products, at most M products belong to the same category. A series of diversity rules can be generated by adjusting the parameters N and M. For the design of the action space $\mathcal{A}$, we follow two intuitions. First, the candidate diversity rules should be capable to satisfy the diversity demands of users with different long-term preferences and short-term interests. Second, the current fixed diversity rule should be one of the candidate diversity rule. Hence, a well-designed action space should consist of three parts: 1) the current fixed diversity rule as the middle diversity degree, 2) several diversity rules with higher diversity degrees by increasing N or decreasing M, and 3) the same number of diversity rules with lower diversity degrees by decreasing N or increasing M. This approach has twofold advantages. First, the candidate diversity rule set becomes balanced, so it avoids sharp performance degradation in deploying the personalized diversity re-ranking model online. Second, it makes the comparison between the fixed diversity rule and the personalized diversity policy model more interpretable.

Appendix for Off-Policy Policy Gradient. With the behavior policy been provided, we can obtain an approximation of the policy gradient $\nabla_\theta J(\pi_\theta)$ based on the logged data collected by policy β. The intuition of off-policy policy gradient estimation is to circumvent the distribution mismatch between the target policy π_θ and the behavior policy β. Using importance weighting strategy [3, 5], the approximation of the off-policy REINFORCE gradient can be derived.

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{s_t \sim d_t^\pi(\cdot), a_t \sim \pi_\theta(\cdot|s_t)} \left[R(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right] \\
&= \sum_{s_t, a_t} d_t^\pi(s_t) \pi_\theta(a_t|s_t) R(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \\
&= \sum_{s_t, a_t} d_t^\beta(s_t) \beta(a_t|s_t) \frac{d_t^\pi(s_t) \pi_\theta(a_t|s_t)}{d_t^\beta(s_t) \beta(a_t|s_t)} R(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \\
&= \mathbb{E}_{s_t \sim d_t^\beta(\cdot), a_t \sim \beta(\cdot|s_t)} \left[\frac{d_t^\pi(s_t)}{d_t^\beta(s_t)} \frac{\pi_\theta(a_t|s_t)}{\beta(a_t|s_t)} R(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right].
\end{aligned}
$$

Since the transition probability $p(s_{t+1}|s_t, a_t)$ is not modelled in our task, it is difficult to estimate the state distribution $d_t(\cdot)$. Therefore, we follow the work of

Chen et al. [2] and simply ignore the term $d_t^{\pi}(s_t)/d_t^{\beta}(s_t)$, resulting in a slightly biased estimation of policy gradient under the behavior policy β.

$$\nabla_\theta J(\pi_\theta) \approx \mathbb{E}_{s_t \sim d^\beta(\cdot), a_t \sim \beta(\cdot|s_t)} \left[\frac{\pi_\theta(a_t|s_t)}{\beta(a_t|s_t)} R(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]. \quad (2)$$

We can use the above gradient to perform off-policy learning.

We share an attempt on the reward design, discuss some limitations, and give several future works of this personalized diversity re-ranking model.

An Attempt on Reward Design. In the above experiments, we design the number of products that a user browses in each user request as an immediate reward for optimizing the user browsing depth. In the online experiment, we observe there is an insignificant slight decline in UCTR. In fact, there is a contrary relationship between click and browse to some extent. The increase of diversity may bring more products that do not exactly match the user's explicit interests, so the number of clicks may be reduced. For those recommender systems that do not want to have a little bit loss of click, we made an attempt on the reward design. We collect another 16 consecutive days of user interaction data, where the immediate reward is defined as a weighted linear combination of the number of exposures and the number of clicks in each user request. We set the weights for the click reward and the exposure reward as 0:1, 2:1, and 4:1, respectively. Here we also apply the metric *ExposureGain* for offline model selection. Not surprisingly, the offline results show that the greater the weight of the click reward, the smaller the value of *ExposureGain*. We select the best model under each weights and observe their online performances. For user browsing depth, the model of weight 0:1 has the highest improvement, and the model of weight 2:1 outperforms that of 4:1. On the other hand, the reverse applies for UCTR and UNOC. This attempt may be able to give some inspirations to those recommendation systems which have different tendencies to click and browse.

Limitations and Future Works. As described in the main paper, the pointwise LTR model and the personalized diversity policy network are conducted concurrently. We cannot obtain the information of top-ranked products before we choose an diversity rule. Thus, the representation of state only contain the user information, contextual information, previous click and exposure sequences. If the personalized diversity policy model has the access to the products, we can incorporate them into the state presentation, which is likely to improve the policy performance. Besides, the design of action space can also be further improved. Although we have designed several candidate diversity rule sets with different numbers of actions and different highest (lowest) diversity degrees, which is also limited. Hence, the number of actions and the diversity degrees can be further tuned for better performance. Currently, all the candidate diversity rules are defined at the category level. In fact, the diversity rule can be considered from other perspectives, not just in the sense of product categories. For example, with more and more kinds of materials appearing in the feed streaming recommendation, such as video and live, we can carry on the personalized diversity in

the sense of material types. These limitations are also potentialities for further researches.

In summary, we believe this paper provides useful insights and experiences to help people on optimizing long-term user engagement while ensuring instant metrics, and on developing personalized relevance-diversity trade-off in feeds recommendation. Moreover, although we follow the greedy based re-ranking in consideration of compatibility with the existing components, our model can be further incorporated into list-wise re-ranking approaches by providing high-quality candidate recommendation results.

References

1. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. p. 335–336. Association for Computing Machinery (1998)
2. Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., Chi, E.H.: Top-k off-policy correction for a REINFORCE recommender system. In: WSDM. pp. 456–464. ACM (2019)
3. Munos, R., Stepleton, T., Harutyunyan, A., Bellemare, M.G.: Safe and efficient off-policy reinforcement learning. arXiv preprint arXiv:1606.02647 (2016)
4. Nowozin, S., Cseke, B., Tomioka, R.: f-gan: Training generative neural samplers using variational divergence minimization. In: NeurIPS. pp. 271–279 (2016)
5. Precup, D., Sutton, R.S., Dasgupta, S.: Off-policy temporal-difference learning with function approximation. In: ICML. pp. 417–424 (2001)