



# Reinforcement Re-ranking with 2D Grid-based Recommendation Panels

Sirui Chen  
School of Information  
Renmin University of China  
Beijing, China  
chensr16@gmail.com

Zhiyu Li  
Alibaba Group  
Hangzhou, China  
tuanyu.lzy@taobao.com

Xiao Zhang  
Gaoling School of Artificial  
Intelligence  
Renmin University of China  
Beijing, China  
zhangx89@ruc.edu.cn

Yuan Wang  
Alibaba Group  
Hangzhou, China  
wy175696@taobao.com

Xu Chen  
Gaoling School of Artificial  
Intelligence  
Renmin University of China  
Beijing, China  
successcx@gmail.com

Quan Lin  
Alibaba Group  
Hangzhou, China  
tieyi.lq@taobao.com

Jun Xu\*  
Gaoling School of Artificial  
Intelligence  
Renmin University of China  
Beijing, China  
junxu@ruc.edu.cn

## ABSTRACT

Modern recommender systems usually present items as a streaming, one-dimensional ranking list. Recently there is a trend in e-commerce that the recommended items are organized grid-based panels with two dimensions where users can view the items in both vertical and horizontal directions. Presenting items in grid-based result panels poses new challenges to recommender systems because existing models are all designed to output sequential lists while the slots in a grid-based panel have no explicit order. Directly converting the item rankings into grids (e.g., pre-defining an order on the slots) overlooks the user-specific behavioral patterns on grid-based panels and inevitably hurts the user experiences. To address this issue, we propose a novel Markov decision process (MDP) to place the items in 2D grid-based result panels at the final re-ranking stage of the recommender systems. The model, referred to as Panel-MDP, takes an initial item ranking from the early stages as the input. Then, it defines *the MDP discrete time steps as the ranks in the initial ranking list, and the actions as the prediction of the user-item preference and the selection of the slots*. At each time step, Panel-MDP sequentially executes two sub-actions: first deciding whether the current item in the initial ranking list is preferred by

the user; then selecting a slot for placing the item if preferred, or skipping the item otherwise. The process is continued until all of the panel slots are filled. The reinforcement learning algorithm of PPO is employed to implement and learn the parameters in the Panel-MDP. Simulation and experiments on a dataset collected from a widely-used e-commerce app demonstrated the superiority of Panel-MDP in terms of recommending 2D grid-based result panels.

## CCS CONCEPTS

• Information systems → Recommender systems; • Computing methodologies → Reinforcement learning.

## KEYWORDS

re-ranking, reinforcement learning, recommender system

### ACM Reference Format:

Sirui Chen, Xiao Zhang, Xu Chen, Zhiyu Li, Yuan Wang, Quan Lin, and Jun Xu. 2023. Reinforcement Re-ranking with 2D Grid-based Recommendation Panels. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '23)*, November 26–28, 2023, Beijing, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3624918.3625311>

## 1 INTRODUCTION

In recent years, there is a trend in e-commerce recommendation apps that the results can be presented in 2D grid panels so that the users can browse results vertically and horizontally. Studies have shown that comparing with the traditional 1D ranking list, 2D grid panels are more attractive since the users can make broader and deeper explorations [8].

Unlike 1D ranking list where the users' top-down browsing manner makes it possible to directly allocate the items from the top according to their ranked preference scores, in 2D panels, the user

\*Corresponding author. Work partially done at Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

SIGIR-AP '23, November 26–28, 2023, Beijing, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0408-6/23/11...\$15.00

<https://doi.org/10.1145/3624918.3625311>

feedback is gained after the user browses the whole page. This makes that the correspondences between the slots and the preference scores are not immediate, which is usually overlooked in traditional recommender systems. One may consider straightforward heuristics, such as expanding the panels from the top-left to the bottom-right, to degenerate the 2D re-ranking problem into a conventional 1D re-ranking task. However, the user interaction patterns on grid panels are different from that of the ranking list [18]: they browse the result panel in both horizontal and vertical directions; their attentions are biased to the middle positions, and their attentions do not decrease monotonically and dramatically with the ranks. Directly mapping the slots in a 2D grid-based panel to the ranks in a 1D list deviates from the users' activity patterns and inevitably hurts the user experiences.

To address the issue, this paper proposes a novel Markov decision process (MDP)-based re-ranking model for final-stage recommendation, called Panel-MDP. Specifically, Panel-MDP formulates the process of pruning and placing the items from an initial ranking list to a grid-based result panel as sequential decision making. To overcome the difficulty that the slots in a grid-based panel have no explicit order, unlike existing MDP-based re-ranking models [1, 19, 21], Panel-MDP defines its time steps as the ranks in the initial list. Its actions are defined as compositions of two parts, respectively correspond to the two factors in the final re-ranking stage: first judging whether to expose the current item and then selecting the position to place the item if it needs to be exposed.

With the above configuration, Panel-MDP is enabled to generate grid-based results naturally. At each time step  $t = 0, 1, \dots$ , Panel-MDP processes the  $(t + 1)$ -th item in the initial ranking list: according to the current state and guided by the policy, it first selects an action  $a = (a^{\text{exp}}, a^{\text{pos}})$ , where  $a^{\text{exp}}$  denotes the selection of the item, and  $a^{\text{pos}}$  denotes the position to be placed. The item is placed to the slot indexed by  $a^{\text{pos}}$  if  $a^{\text{exp}} = 1$ , otherwise the item is discarded. Then, the MDP state is updated for processing the next item in the initial ranking. The process will continue until the result panel is full or all items in the initial ranking have been processed. To learn the model parameters, the reinforcement learning algorithm of Proximal Policy Optimization (PPO) is employed, which takes the (simulated) user feedback on the grid-based result panels as the training signals.

Panel-MDP is tailored for 2D grid-based result panels, and offers several advantages: (1) the novel configuration of time steps and actions enables it to generate grid-based result panels naturally. Moreover, it has the potential to generate more sophisticated result layouts; (2) it enables the model to capture the special user activity patterns on grid-based result panels during the model learning.

In summary, the main contributions of this paper are presented as follows: (1) we highlight a new re-ranking task where the items are allocated into two-dimensional grid-based panels; (2) we propose a novel MDP-based final-stage ranking model, called Panel-MDP, to tackle the task; (3) we demonstrated the effectiveness of Panel-MDP against the state-of-art baselines on a real-world dataset collected from a popular e-commerce app.

## 2 PRELIMINARIES

Suppose that a user  $u \in \mathcal{U}$  accesses a recommender system and the early-stage of the recommendation pipeline retrieves a set of  $K$  items and organizes them as an initial ranking list:  $\mathbf{L} = [i_1, \dots, i_K] \in \mathcal{I}^K$  where  $\mathcal{I}$  is the set of all items, and  $i_k \in \mathcal{I}$  is the  $k$ -th item in  $\mathbf{L}$ . Items are typically represented as real-valued embeddings, denoted by  $\mathbf{e}(i) \in \mathbb{R}^d$  where  $d$  is the dimension of the item embeddings.

In the final-stage recommendation, (a subset of) these  $K$  items are placed to a grid-based panel of size  $M$  rows and  $N$  columns, denoted as  $\mathbf{P} \in \mathcal{I}^{M \times N}$ , where each position  $P_{m,n} = i_k$  means item  $i_k$  is placed to the  $(m, n)$ -th position in the panel. Note that usually  $M \times N \leq K$ . After the user  $u$  has viewed a recommended panel, the system records the user feedback (e.g., 1 for purchase and 0 otherwise) which can be considered as the labels or rewards for MDP. The goal of the final-stage recommendation is to output a grid-based panel  $\mathbf{P}$  based on the input ranking list  $\mathbf{L}$ .

One straightforward approach to placing the items in grids is pre-defining a total order by using some heuristics (e.g., ordering the positions in a panel from left to right, and then top to down). Therefore, the task de-generates to traditional re-ranking in recommendation and any re-ranking models can be directly applied. Though preliminary successes have been achieved [1], the previous studies have also shown that user behaviors on 2D grid-based panels are very different from those on traditional sequential list-based rankings [18]. First of all, since users can browse items vertically and horizontally, they will allocate more attention to the middle part of the panel, called middle bias; Second, user studies also showed that the users' attention decay slowly by row and they often display row-skipping behavior; Third, we found that on product apps, the users tend to click one item on a result panel though they may have interest on multiple items. Note that though some of the above conclusions were drawn based on image search, we also observe similar phenomena in e-commerce recommendation.

The analysis indicates that directly applying the traditional re-ranking models to generate grid-based result panels is not an optimal solution. It is necessary to design a new model tailored for the task.

## 3 OUR APPROACH: PANEL-MDP

### 3.1 Configuration of Panel-MDP

Existing MDP-based re-ranking models usually define the time steps to the ranks in the output result list, and the actions to the input items. This configuration, however, can only generate ranking lists because the time steps naturally form a sequence.

Different from existing MDP-based ranking models, Panel-MDP defines its *time-steps as the ranks in the input list  $\mathbf{L}$ , and the actions are defined as the user-item preference predictions and the slots in the panel  $\mathbf{P}$* . In this way, Panel-MDP naturally generates a grid-based result panel, through processing the items  $\mathbf{L}$  in a sequential manner.

Specifically, Panel-MDP's time step, state, action, state transition, reward function and discount factor are defined as follows:

**Time step**  $t \in \{0, \dots, K - 1\}$ : We define the time step  $t$  on the input ranking list  $\mathbf{L} = [i_1 \dots, i_K]$ . Therefore, for  $\mathbf{L}$  with  $K$  items, the Panel-MDP will run at most  $K$  steps, each processing the item ranked at  $(t + 1)$ -th position in  $\mathbf{L}$ , denoted  $i_{t+1}$ .

**Action**  $a = (a^{\text{exp}}, a^{\text{pos}}) \in \mathcal{A} \subseteq \mathcal{A}^{\text{exp}} \times \mathcal{A}^{\text{pos}}$ , where the sub-action  $a^{\text{exp}} \in \mathcal{A}^{\text{exp}} = \{0, 1\}$  indicates whether the current item is selected to expose: 1 means expose the item and 0 otherwise. The sub-action  $a^{\text{pos}} \in \mathcal{A}^{\text{pos}} \subseteq \{(m, n) | m = 1, \dots, M; n = 1, \dots, N\}$  indicates the index of the slot in the grid-based result panel.

Specifically, at the  $t$ -th time step, the agent of Panel-MDP processes the item  $i_{t+1}$  in  $\mathbf{L}$  and selects an action  $a_t = (a_t^{\text{exp}}, a_t^{\text{pos}} = (m, n))$ . The agent will discard item  $i_{t+1}$  if  $a_t^{\text{exp}} = 0$ , without considering the values of  $a_t^{\text{pos}}$ . Otherwise (i.e.,  $a_t^{\text{exp}} = 1$ ), item  $i_{t+1}$  will be placed to the  $(m, n)$ -th slot in panel matrix  $\mathbf{P}$ , i.e.,  $P_{a_t^{\text{pos}}} = P_{m,n} \leftarrow i_{t+1}$ . Note that one slot at the result panel can be placed at most one item. The slot index will be removed from  $\mathcal{A}^{\text{pos}}$  if it has been used. Formally, let's denote the action space at the  $t$ -th step as  $\mathcal{A}_t = \mathcal{A}_t^{\text{exp}} \times \mathcal{A}_t^{\text{pos}}$ . After choosing an action  $a_t = (a_t^{\text{exp}}, a_t^{\text{pos}})$  and moving to the time step  $(t + 1)$ , the action space becomes  $\mathcal{A}_{t+1} = \mathcal{A}_{t+1}^{\text{exp}} \times \mathcal{A}_{t+1}^{\text{pos}}$  where

$$\mathcal{A}_{t+1}^{\text{pos}} = \begin{cases} \mathcal{A}_t^{\text{pos}} & a_t^{\text{exp}} = 0, \\ \mathcal{A}_t^{\text{pos}} \setminus \{a_t^{\text{pos}}\} & a_t^{\text{exp}} = 1, \end{cases} \quad (1)$$

and at the initial time-step  $t = 0$ ,  $\mathcal{A}_0^{\text{pos}} = \{(m, n) | m = 1, \dots, M; n = 1, \dots, N\}$ .  $\mathcal{A}_t^{\text{exp}} = \{0, 1\}$  for all time-steps  $t$ .

**State**  $s \in \mathcal{S}$ : The state of the Panel-MDP at the  $t$ -th time step can be described as a tuple  $s_t = [\mathbf{L}_t, t, \mathbf{H}_t]$  where  $\mathbf{L}_t = [i_t, \dots, i_K]$  denotes the candidate items which have not been seen yet in the input list, and  $\mathbf{H}_t = [a_0, \dots, a_{t-1}]$  is the sequence of actions already selected until step  $t$ . At the time step  $t = 0$ , the state is initialized as  $s_0 = [\mathbf{L}, 0, \mathbf{H}_0 = \emptyset]$ , where  $\emptyset$  is the empty list.

**State transition**  $T$ : After issuing the action  $a_t$  at state  $s_t$ , the system will transit to a new state. The state transition function  $T: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  is defined as:

$$\begin{aligned} s_{t+1} &= T(s_t = [\mathbf{L}_t = [i_t, \dots, i_K], t, \mathbf{H}_t = [a_0, \dots, a_{t-1}], a_t]) \\ &= [\mathbf{L}_{t+1} = [i_{t+1}, \dots, i_K], t + 1, \mathbf{H}_{t+1} = [a_0, \dots, a_{t-1}, a_t]], \end{aligned}$$

which simply discards the current item  $i_t$  from  $\mathbf{L}_t$ , increases the time step by one, and append the chosen actions  $a_t$  at the  $t$ -th step to the action history  $\mathbf{H}_t$ .

**Rewards**  $r_t = R(s_t, a_t)$ : After presenting a whole result panel to a user, the user may give feedback on the recommended items, i.e., purchase or not. Therefore, we can define the reward based on that feedback. For example,  $r_t = R(s_t, a_t = (1, (m, n))) = 1$  if the user purchases the item at the  $(m, n)$ -th position, and 0 otherwise.

**Discounted Factor**  $\gamma \in (0, 1]$  discounts rewards at a constant rate per time step.

## 3.2 Model Structure and Learning

We present an implementation of Panel-MDP policy with a famous Actor-Critic algorithm of Proximal Policy Optimization (PPO) [13].

**3.2.1 Model Structure.** The implementation of our Panel-MDP framework with PPO comprises three components: a state encoder to capture the representation of a given state; a policy network to make an action based on the state, and an advantage network to measure the advantage of action in the current state.

In general, the state representation can be described as a concatenation of the component embeddings in the state:

$$\mathbf{e}^s = \text{concat}(\mathbf{e}^L, \mathbf{e}^t, \mathbf{e}^H),$$

where operator ‘concat’ concatenates all the inputted vectors,  $\mathbf{e}^t$  is the embedding of the time step  $t$  which is automatically learned in an end2end manner;  $\mathbf{e}^L$  encodes the candidate item list  $\mathbf{L}_t = [i_t, \dots, i_K]$ , and is defined as the mean pooling of the output vectors by a multi-layer Perceptron (MLP) layer:  $\mathbf{e}^L = \text{Mean}(\text{MLP}_1([\mathbf{e}_{i_t}^i, \dots, \mathbf{e}_{i_K}^i]))$ , where  $\text{Mean}(\cdot)$  denotes the mean pooling of all the input vectors, and  $\mathbf{e}_{i_t}^i$  is the embedding of the  $t$ -th item in the input list  $\mathbf{L}$ . As for  $\mathbf{e}^H$ , it encodes the previous action sequence  $\mathbf{H}_t = [a_1, \dots, a_{t-1}]$ . It is defined as the final output vector of a GRU after the whole sequence has been scanned:  $\mathbf{e}^H = \text{GRU}([\mathbf{e}_{a_1}^a, \dots, \mathbf{e}_{a_{t-1}}^a])$ , where  $\mathbf{e}^a$  are action embeddings which can also be automatically learned in an end2end manner.

For policy network, the policy  $\pi$  for the first sub-action  $a^{\text{exp}}$  is:

$$\pi(a^{\text{exp}} = 1 | s) = \sigma(\text{MLP}_3(\mathbf{h})), \quad (2)$$

where  $\sigma$  is the sigmoid function, and  $\mathbf{h} = \text{MLP}_2(\mathbf{e}^s)$  is transformed from the state embedding  $\mathbf{e}^s$ .

As for the policy  $\beta$  for the second sub-action  $a^{\text{pos}}$ , we introduce the slot position embeddings and  $\forall p \in \mathcal{A}_t^{\text{pos}}$ :

$$\beta(a^{\text{pos}} = p | s) = \frac{\exp(\mathbf{h}^T \mathbf{e}^p)}{\sum_{p' \in \mathcal{A}_t^{\text{pos}}} \exp(\mathbf{h}^T \mathbf{e}^{p'})}, \quad (3)$$

where  $\mathbf{e}^p$  (and  $\mathbf{e}^{p'}$ ) is the embedding of slot  $p$  (and  $p'$ ).

Considering the high cost of directly estimating the advantages, we employ the GAE method [12] to estimate the two advantages, namely  $A(s, a^{\text{exp}})$  and  $A(s, a^{\text{pos}})$ , on the basis of the value function  $V(s)$ :  $V(s) = \text{MLP}_4(\mathbf{e}^s)$ , where  $\text{MLP}_4$  outputs a scalar value.

**3.2.2 Model Learning.** To learn the model parameters (including those in the four MLPs, GRU, time-step embeddings, position embeddings, and action embeddings), a batch of  $B$  trajectories  $\mathcal{D} = \{\tau_b\}_{b=1}^B$  are collected, where each  $\tau_b$  is generated by the following procedure: for  $t = 1, \dots, K$ , given a state  $s_t$ , the agent will sample an action  $a_t = (a_t^{\text{exp}}, a_t^{\text{pos}})$  from the two stochastic policies:  $a_t^{\text{exp}} \sim \pi(a^{\text{exp}} | s)$  and  $a_t^{\text{pos}} \sim \beta(a^{\text{pos}} | s)$ . After executing  $a_t$ , the agent receives a reward  $r_t$  and updates the state to  $s_{t+1}$ . The procedure is repeated until the  $\mathcal{A}_t^{\text{pos}} = \emptyset$  or  $t > K$ . In this way, we collect a trajectory  $\tau_b = \{(s_t, a_t = (a_t^{\text{exp}}, a_t^{\text{pos}}), r_t, s_{t+1})\}_{t=0}^{|\tau_b|-1}$ .

PPO is employed for conducting the learning process that amounts to optimizing the following loss function:

$$\mathcal{L} = \mathcal{L}^{\text{Actor}} + \lambda_1 \mathcal{L}^{\text{Critic}} + \lambda_2 \mathcal{L}^{\text{Entropy}} \quad (4)$$

where  $\lambda_1 > 0$  and  $\lambda_2 > 0$  are trade-off coefficients,  $\mathcal{L}^{\text{Actor}}$  is the loss from the actor:

$$\begin{aligned} \mathcal{L}^{\text{Actor}} &= -\mathbb{E}_{\mathcal{D}} \left[ \min \left( w^{\text{exp}} A(s, a^{\text{exp}}), \text{clip} \left( w^{\text{exp}}, 1 - \epsilon, 1 + \epsilon \right) A(s, a^{\text{exp}}) \right) \right. \\ &\quad \left. + \mathbb{I} [a^{\text{exp}} = 1] \cdot \min \left( w^{\text{pos}} A(s, a^{\text{pos}}), \text{clip} \left( w^{\text{pos}}, 1 - \epsilon, 1 + \epsilon \right) A(s, a^{\text{pos}}) \right) \right], \end{aligned}$$

where for each  $\tau \in \mathcal{D}$  and for each tuple  $(s, (a^{\text{exp}}, a^{\text{pos}}), r, s') \in \tau$ ,  $w^{\text{exp}} = \frac{\pi(a^{\text{exp}} | s)}{\pi_{\text{old}}(a^{\text{exp}} | s)}$  and  $w^{\text{pos}} = \frac{\beta(a^{\text{pos}} | s)}{\beta_{\text{old}}(a^{\text{pos}} | s)}$  denotes the probability ratios between new policies and old policies on the sub-action  $a^{\text{exp}}$  and  $a^{\text{pos}}$ , respectively. The ‘clip’ operator with parameter  $\epsilon$  decides how far the new policy can deviate from the old policy.  $\mathbb{I}[\cdot]$  is an indicator function which splits  $\mathcal{L}^{\text{Actor}}$  into two parts: the first part corresponds to sub-action  $a^{\text{exp}}$  and the second part corresponds to sub-action  $a^{\text{pos}}$ . The second part is dependent on the value of  $a^{\text{exp}}$ : when  $a^{\text{exp}} = 1$  (i.e., the item will be exposed at  $a^{\text{pos}}$ -th slot), the

**Algorithm 1** Training Panel-MDP with PPO

---

**Input:** Initialized Actor  $\pi, \beta$ , Critic  $V$

- 1: **repeat**
- 2:    $\mathcal{D} \leftarrow \emptyset$
- 3:   **for**  $b = 0, \dots, B$  **do**
- 4:     Initialize:  $s \leftarrow [L, 0, \emptyset]$ ;  $\mathcal{A}^{\text{exp}} \leftarrow \{0, 1\}$ ;  $\mathcal{A}^{\text{pos}} \leftarrow \{(m, n) | m = 1, \dots, M; n = 1, \dots, N\}$ ;  $\tau_b \leftarrow \emptyset$
- 5:     **for**  $t = 0, 1, \dots$  **do**
- 6:       Sample sub-actions  $a^{\text{exp}} \sim \pi, a^{\text{pos}} \sim \beta$  {Eq. (2)& (3)}
- 7:       Execute  $a = (a^{\text{exp}}, a^{\text{pos}})$ , observe reward  $r$ , and new state  $s'$  {Sec. 3.1}
- 8:       Update  $\mathcal{A}^{\text{pos}}$  {Eq. (1)}
- 9:       Append  $(s, (a^{\text{exp}}, a^{\text{pos}}), r, s')$  to  $\tau_b$
- 10:     **end for**
- 11:      $\mathcal{D} \leftarrow \mathcal{D} \cup \{\tau_b\}$
- 12:   **end for**
- 13:   **for all**  $\tau \in \mathcal{D}$  and  $0 \leq t < |\tau|$  **do**
- 14:     Compute rewards-to-go  $R_t$ 's
- 15:     Compute advantages  $A(s, a^{\text{exp}})$  and  $A(s, a^{\text{pos}})$  using GAE method [12], based on current value function  $V$
- 16:   **end for**
- 17:   Update Actor-Critic by applying SGD w/ Adam to Eq. (4)
- 18: **until** convergence
- 19: **return**  $\pi, \beta, V$

---

part will be counted in the total loss. Otherwise, this part will be ignored because the sub-action is actually not issued.

As for  $\mathcal{L}^{\text{Critic}}$ , it is defined as the mean-squared loss:

$$\mathcal{L}^{\text{Critic}} = \mathbb{E}_{\tau \sim \mathcal{D}} \mathbb{E}_t [(V(s_t) - R_t)^2],$$

where  $R_t = \sum_{k=t}^{|\tau|-1} \gamma^{k-t} r_k$  is the discounted sum of future reward.

Besides, we also apply an entropy-based regularizer on  $\pi(a^{\text{pos}}|s)$  to encourage the agent exploration on the different positions:

$$\mathcal{L}^{\text{Entropy}} = -\mathbb{E}_{\mathcal{D}} [\beta(a^{\text{pos}}|s) \log \beta(a^{\text{pos}}|s)].$$

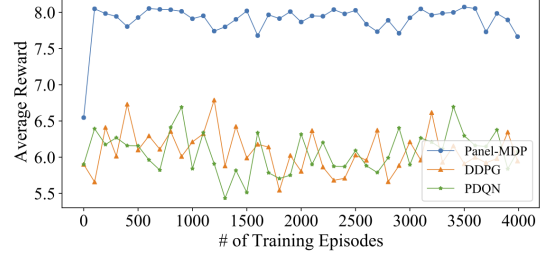
Standard PPO-clip with gradient descent [13] is applied to optimize the parameters. In the online re-ranking phase, after receiving the initial ranking list  $L$  and state initialization, Panel-MDP constructs a grid-based result panel by repeatedly executing actions through maximizing the learned policies  $\pi$  and  $\beta$ , updating the state and actions space, and moving to the next iteration. This procedure ends until the result panel is full or all items are processed.

## 4 EXPERIMENTS

We conducted experiments to test our Panel-MDP.

To the best of our knowledge, there is no publicly available model specially designed for recommending grid-based results in the literature. Therefore, we adapted existing re-ranking models to generate grids, by simply placing the outputted item list in a predefined order(left-to-right and top-to-down). The chosen re-ranking models include the supervised model **miDNN** [22], and RL-based re-ranking model of **DDPG** [9] and **PDQN** [24].

**Experiments on simulated environment:** We constructed a simulation environment to test whether Panel-MDP and other RL-based algorithms can capture the user activity patterns. In the



**Figure 1: Performance curves of Panel-MDP and the RL-based baselines on simulated environment.**

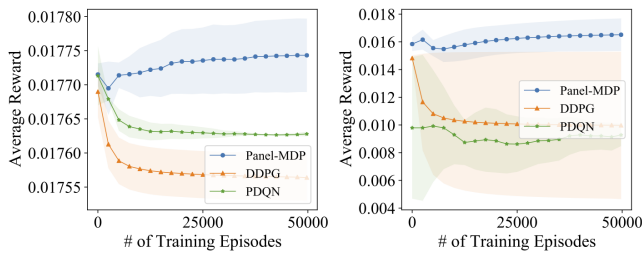
simulated environment and at every session, the agent will receive a user and a ranking of 20 candidate items (represented as features). The goal is to re-rank the items in a  $4 \times 3$  grid. The user will provide feedback to each item based on the given grid in terms of his preference both on items and positions. Figure 1 demonstrates the performance curves of Panel-MDP and the baselines of DDPG and PDQN, in terms of cumulative average reward w.r.t. number of training episodes. We can observe that Panel-MDP achieves highest cumulative average reward at very beginning of the training episodes, verified the effectiveness of Panel-MDP for capturing the special user activity patterns on grid-based result panel.

**Experiments on real dataset:** We also conducted experiments on a dataset collected from a popular e-commerce app in Oct. 2021, where the recommendation results are displayed as  $2 \times 3$  grid-based panels. For each user request, the early stages provided 16 ranked candidate items. The users' feedback (i.e., purchase) was collected as the labels. Users' profile (e.g., hashed user id, non-sensitive user feature etc.) and the items' profile (e.g., category, price, etc.) were stored for calculating the user and item embeddings. After pre-processing, the dataset contains 20,000 user requests for training and 1,000 for testing. We tested Panel-MDP on two tasks. The first

**Table 1: Performance comparison between Panel-MDP and the baselines in terms of averaged reward and AUC. ‘\*’ means the improvements over the best baseline are statistical significant (t-tests and  $p$ -value  $< 0.05$ ).**

Model	Task 1: Re-org	Task 2: Select&Re-org	
	Average Reward	Average Reward	AUC
miDNN	0.0177	0.0178	0.7524
DDPG	0.0175	0.0145	0.6613
PDQN	0.0177	0.0163	0.7524
Panel-MDP (Ours)	<b>0.0179*</b>	<b>0.0179*</b>	<b>0.7980*</b>

task (called **Re-Org**) assumes that the early stages only provide 6 ranked items (the top 6) and the model re-organizes them to a  $2 \times 3$  grid, by setting sub-action  $a_t^{\text{exp}} = 1$  for all  $t$ . The second task (called **Select&Re-Org**) assumes that the early stages provide 16 ranked items and the model needs to select 6 items and re-organize them to a  $2 \times 3$  grid. As for evaluation, we use a user simulator trained on large-scale user-system interactions. Given a user, the context, and



**Figure 2: Performance curves of Panel-MDP and the RL-based baselines on real dataset. Left: Re-org; Right: Select&Re-org. The shading areas represent the standard deviations under running episodes.**

the result panel, the simulator outputs a *score* denoting the purchase probability that is treated as the *reward* from environments, and the average reward is reported. For Select&Re-ORG, we also used AUC for evaluation.<sup>1</sup>

Table 1 reports the best performance in terms of the average reward on the test set. From the results, we can see that Panel-MDP outperformed all the baselines on both of the tasks, indicating the effectiveness of Panel-MDP in re-organizing the items to 2D grid-based result panels. In the task of Select&Re-org, the action  $a^{\text{EXP}}$  is the key for exploration on selection. Compared to other RL baselines that have shown a decline in performance on this task, exploration allows Panel-MDP to achieve a better average reward and AUC.

Figure 2 illustrates the performance curves of Panel-MDP and the RL-based baselines in terms of cumulative average reward w.r.t. number of training episodes on the test set. Panel-MDP consistently outperformed the baselines of DDPG and PDQN, especially in the task of Select&Re-org. This is because both DDPG and PDQN lack the ability to capture the correspondence between positions. Their performances decline fast, indicating that the configuration of Panel-MDP is suitable for generating 2D result panels.

## 5 RELATED WORK

Re-ranking considers the potential impact of the interrelationship of the candidate items on user preference [7, 15, 16]. Bello et al. [1] considered the influence on the previous recommended items into final-stage re-ranking models; Zhuang et al. [22] proposed to fully capture the mutual information of the items in candidate sets. Pei et al. [11] proposed to further extract information in the recommended list using transformer structure.

Recently, reinforcement learning (RL) methods have been widely employed in recommendation to enhance long-term engagement or boost various objectives [3, 4, 10, 20]. Shani et al. [14] first applied MDP to recommendation. Cai et al. [2] defined a general framework for RL recommender systems. Huang et al. [6], Zou et al. [23] both utilized RL framework to model the change of users' long-term engagement and preference. Zhao et al. [21] proposed a page-wise RL-based recommendation model.

Compared to the re-ranking, much less research efforts have been

<sup>1</sup>In the Re-ORG task, AUC cannot be calculated because all the candidate items were treated positive(i.e., there is no score in RL recommender systems).

spent on presenting results other than lists, e.g., grid-based result panels. Xie et al. [17] studied the characteristics of user attentions on a 2D image search result page, and found different behavioral patterns like “middle bias”, “row skipping” and “slow decay”. Guo et al. [5] propose a de-biasing method on 2D result pages.

## 6 CONCLUSION AND FUTURE WORK

This paper proposes a novel MDP-based model that formalizes the generation of a grid-based result panel from a list of ranked items as a discrete-time sequential decision making, where the time steps and actions are respectively defined as the positions in the input list and slots in the result panel. The model, called Panel-MDP, employs PPO to implement the Actor-Critic and optimize the model based on users' feedback. Experimental results on a real-world dataset collected from a popular e-commerce app showed its effectiveness.

Panel-MDP is a preliminary research on re-ranking within grid-based panels. However, the experiments were carried out in a restricted setting with small grids. To further substantiate its efficacy, additional baseline re-ranking methods could be incorporated. Future research will delve deeper into understanding user preferences in grid-based panels and will involve comparative experiments with a broader range of baseline methods on more intricate grids.

## ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (No. 62376275), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (23XNKJ13). This work was supported by Alibaba Group through Alibaba Innovative Research Program. This work was funded by Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class” Initiative, Renmin University of China. The work was partially done at Beijing Key Laboratory of Big Data Management and Analysis Methods.

## REFERENCES

- [1] Irwan Bello, Sayali Kulkarni, Sagar Jain, Craig Boutilier, Ed Chi, Elad Eban, Xiyang Luo, Alan Mackey, and Ofer Meshi. 2018. Seq2slate: Re-ranking and slate optimization with RNNs. *arXiv preprint arXiv:1810.02019* (2018).
- [2] Qingpeng Cai, Aris Filos-Ratsikas, Pingzhong Tang, and Yiwei Zhang. 2018. Reinforcement mechanism design for fraudulent behaviour in e-commerce. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 957–964.
- [3] Sirui Chen, Yuan Wang, Zijing Wen, Zhiyu Li, Changshuo Zhang, Xiao Zhang, Quan Lin, Cheng Zhu, and Jun Xu. 2023. Controllable Multi-Objective Re-ranking with Policy Hypernetworks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3855–3864.
- [4] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. 2021. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*. 445–453.
- [5] Ruocheng Guo, Xiaoting Zhao, Adam Henderson, Liangjie Hong, and Huan Liu. 2020. Debiasing grid-based product search in e-commerce. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2852–2860.
- [6] Liwei Huang, Mingsheng Fu, Fan Li, Hong Qu, Yangjun Liu, and Wenyu Chen. 2021. A deep reinforcement learning based long-term recommender system. *Knowledge-Based Systems* 213 (2021), 106706.
- [7] Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, and Ming Yue. 2017. Learning to diversify search results via subtopic attention. In *Proceedings of the 40th international ACM SIGIR Conference on Research and Development in Information Retrieval*. 545–554.
- [8] Yvonne Kammerer and Peter Gerjets. 2010. How the interface design influences

- users' spontaneous trustworthiness evaluations of web search results: Comparing a list and a grid interface. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. 299–306.
- [9] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
- [10] Shuchang Liu, Qingpeng Cai, Zhankui He, Bowen Sun, Julian McAuley, Dong Zheng, Peng Jiang, and Kun Gai. 2023. Generative Flow Network for Listwise Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1524–1534.
- [11] Changhua Pei, Yi Zhang, Yongfeng Zhang, Fei Sun, Xiao Lin, Hanxiao Sun, Jian Wu, Peng Jiang, Junfeng Ge, Wenwu Ou, et al. 2019. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM conference on recommender systems*. 3–11.
- [12] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438* (2015).
- [13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [14] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. 2005. An MDP-based recommender system. *Journal of Machine Learning Research* 6, 9 (2005), 1265–1295.
- [15] Koen Verstrepen and Bart Goethals. 2015. Top-n recommendation for shared accounts. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 59–66.
- [16] Fan Wang, Xiaomin Fang, Lihang Liu, Yaxue Chen, Jiucheng Tao, Zhiming Peng, Cihang Jin, and Hao Tian. 2019. Sequential evaluation and generation framework for combinatorial recommender system. *arXiv preprint arXiv:1902.00245* (2019).
- [17] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating examination behavior of image search users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 275–284.
- [18] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Yunqiu Shao, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Grid-based evaluation metrics for web image search. In *Proceedings of the 2019 World Wide Web Conference*. 2103–2114.
- [19] Jun Xu, Zeng Wei, Long Xia, Yanyan Lan, Dawei Yin, Xueqi Cheng, and Ji-Rong Wen. 2020. Reinforcement learning to rank with pairwise policy gradient. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 509–518.
- [20] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. 2023. PrefRec: Recommender Systems with Human Preferences for Reinforcing Long-term User Engagement. (2023).
- [21] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 95–103.
- [22] Tao Zhuang, Wenwu Ou, and Zhirong Wang. 2018. Globally optimized mutual influence aware ranking in e-commerce search. *arXiv preprint arXiv:1805.08524* (2018).
- [23] Lixin Zou, Long Xia, Zhuoye Ding, Jiaying Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2810–2818.
- [24] Lixin Zou, Long Xia, Pan Du, Zhuo Zhang, Ting Bai, Weidong Liu, Jian-Yun Nie, and Dawei Yin. 2020. Pseudo Dyna-Q: A reinforcement learning framework for interactive recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 816–824.