Journal of Systems Engineering and Information Technology (JOSEIT)
Volume 4, No. 1, April (2025), Pages 1-6
e-ISSN: 2829-310X | 10.29207/joseit.

Enhancing News Recommendations with Deep Reinforcement Learning and Dynamic Action Masking

Dong Sang-hong[1], Ahn Jun-soo[2]
[12]School of Software Engineering, Chengdu University of Information Engineering, Chengdu, China

[*]*dongsang211@gmail.com*

Abstract. A news recommender system is crucial for the transmission of news in new media. A deep reinforcement learning-based recommender system is suggested to integrate the characterization capabilities of neural networks with the strategic selection capabilities of reinforcement learning to enhance news recommendation efficacy. Dynamic action masks enhance the capacity to assess short-term interests of users. An optimized caching mechanism improves the efficiency of the experience cache, and a reward design characterized by region masking accelerates model training, thereby enhancing the performance of the recommender system for news recommendations. Experimental results indicate that the recommendation accuracy of the proposed model on the news dataset is on par with that of prevalent neural network recommendation techniques and surpasses existing state-of-the-art algorithms in ranking performance

Keywords: *news recommendations; enhanced learning; dynamic masking; advantageous caching; intrinsic rewards*

How to cites:
Dong Sang-hong, & Ahn Jun-soo. (2025). Enhancing News Recommendations with Deep Reinforcement Learning and Dynamic Action Masking. Journal of Systems Engineering and Information Technology (JOSEIT), 4(1), 1-6. https://doi.org/10.29207/joseit.v4i1.6536

Received by the Editor: 2025-04-13
Final Revision: 2025-04-24
Published: 2025-04-28

SDGs contributed:

© Dong et al (2025) | This is an open-access article under the CC BY 4.0 License
Publisher: Ikatan Ahli Informatika Indonesia

1. Introduction

News recommendations are typically modeled sequentially [1], [2]. However, in real-world recommendation scenarios, the user selection of news topics exhibits irregular diversity [3], and users generally do not click on similar news items consecutively. This behavior differs from that observed in sequential recommendation scenarios such as business class recommendations [4], [5]. Consequently, conventional news topic recommendation models, which are based on historical data and utilize supervised learning techniques, cannot adapt to the variability in user preferences over time [6], [7]. This limitation arises from the fact that these models are not designed to recognize short-term changes in news propensity on a session-by-session basis. By contrast, recommendation models built on supervised learning techniques are more adapted to static, highly associative recommendation scenarios, such as those found in business class recommendations [8]. However, these models are not adapted to dynamic and rapidly changing recommendation scenarios, which are characteristic of news recommendations.

Despite significant advances in recommendation systems, a critical research gap persists in the news-recommendation domain. Traditional supervised learning approaches fail when confronted with distinct temporal characteristics of news consumption patterns. Unlike product recommendations, in which user preferences tend to remain relatively stable, news preferences exhibit pronounced variability and contextual dependency. This variability manifests as rapid shifts in interest triggered by evolving world events, temporal factors, and the inherently ephemeral nature of news content [9]. Conventional models anchored to static representations of user preferences derived from historical interactions struggle to capture these dynamic preference transitions. The disconnect between the fluid nature of news consumption and the rigid framework of traditional recommendation models creates a substantial efficacy gap, particularly in scenarios in which timeliness and contextual relevance are paramount. This study aims to bridge this gap by developing models capable of adapting to the continuous evolution of user preferences in news consumption contexts.

News recommendation models require sequential decision-making capabilities to enhance their applicability to news recommendation [10], [11]. To address this challenge [12] employs deep reinforcement learning was employed to address the to address this challenge [12] employs deep reinforcement learning was employed to

address the news-recommendation task and enhance its operational efficiency. However, the convergence speed of the model remains a concern, and the utilization of greedy algorithms results in an elevated repetition rate. This paper proposes a dynamic action-masking deep reinforcement learning (DAMDRL) approach for modeling a news recommendation system. This approach is based on the deep Q network (DQN) method, which has been shown to enhance the efficiency and accuracy of deep reinforcement learning algorithms.

The mechanisms proposed in this study are strategically designed to address the unique challenges that define the news recommendation landscape. Unlike product recommendations, in which user preferences remain relatively stable and items maintain relevance for extended periods, news consumption exhibits distinct temporal dynamics and diversity requirements. The ephemeral nature of news content, where articles rapidly lose relevance within days or hours, demands recommendation systems capable of adapting to this accelerated lifecycle. Our dynamic action masking approach continuously refines the available action space, ensuring that timely content receives appropriate consideration, while obsolete articles are systematically filtered out of recommendation candidates.

This study proposes the use of dynamic masking to represent the action space in the deep reinforcement learning recommendation method. In addition, this study also proposes the implementation of a priority caching mechanism to reinforce certain experiences stored in the cache, thus accelerating model convergence. The main contribution of this paper is the proposal to use dynamic masking to represent the action space in the deep reinforcement learning recommendation method and the use of a priority caching mechanism to strengthen certain cached experiences, thus accelerating model convergence. Implementation of an exploration method to avoid loops to increase the efficiency of the recommendation system in exploring new items, achieving an optimal balance between the utilization rate and exploration nature.

2. Methods

This study delineates the recommendation process of a recommender system as an act of suggesting an item to a user with the intention of receiving a reward for that action within an environment shaped by the user's historical data, while also striving to modify the user's historical data to facilitate subsequent actions [13], [14]. The entire procedure generates the sequence (s_t, a_t, r_t, s_{t+1}), with the meanings of the elements in the series delineated as follows. The state space S comprises the users' positive feedback, negative feedback, implicit feedback, and items for recommendation, defining the state $s_t \in S$, where s_t represents the user's history data at time t [19]. Action space A comprises the small-scale to be discharged, derived after recalling the sequence of items delineating the action $a_t \in A$, where at represents the behavior of recommending an item to the user in the state s_t. The base reward $_{rt} \in \{0,1\}$ is assigned a value of 1 if the user accepts the recommended item and 0 if the user rejects it. The aggregate reward is represented by $R = r + r^{in}$. Alongside the base reward, the intrinsic reward r^{in} is employed to reduce the exploration of previously established regions by the intelligences [20]

2.1 Improvement of DQN Algorithm and Two-Layer Deep Q-Network Algorithm

The DQN algorithm evaluates the reward of a recommender system for suggesting item a to a user in state s by learning the action-value function Q (s, a). Temporal difference (TD) is a technique employed to approximate the value function of a policy [15], [16]. Conventional value-based reinforcement learning techniques (e.g., Q-learning) employ a tabular method to estimate the TD error; however, this approach is impractical for recommender systems with extensive action spaces. To resolve this issue, the DQN technique employs a neural network for function approximation to represent the action-value function. DQN network. Two issues exist: firstly, correlations among reinforcement learning interaction experiences violate the assumptions of data independence and homogeneous distribution necessary for constructing neural networks; secondly, DQN algorithms typically overestimate exploration values, rendering some of them invalid. To address these two issues, a two-layer deep Q-network (double DQN, DDQN) employs an experience replay mechanism and a target network in both approaches.

Empirical playback: To integrate Q-learning with deep neural networks, the DQN algorithm employs an empirical playback method that involves maintaining a playback buffer that stores quaternion data (s_t, a_t, r_t, s_{t+1}) sampled from the environment. During the training of the Q-network, a random selection of data was drawn from this playback buffer. The target network experiences instability in training because the T D error target is derived from the neural network's output, which fluctuates with each update of the network parameters. To address this issue, DQN establishes a target network aligned with the Q network, stabilizes the TD target within the target network, and refreshes the target network after a specified iteration interval.

2.2 DAMDRL Overall Architecture

The architecture of the DAMDRL model primarily consists of an embedding word-embedding module, a user historical behavior sequence storage module, an experience acquisition module, and a double-layer Q network output module, as shown in Figure 1. Initially, all candidate recommendation items were initialized and defined

by the embedding word-embedding module, and the environment was established. Subsequently, the experience acquisition module engages with the environment to gather necessary experience data. Finally, the DDQN was employed for exploratory learning, and behavioral sequence data from each exploration were recorded.

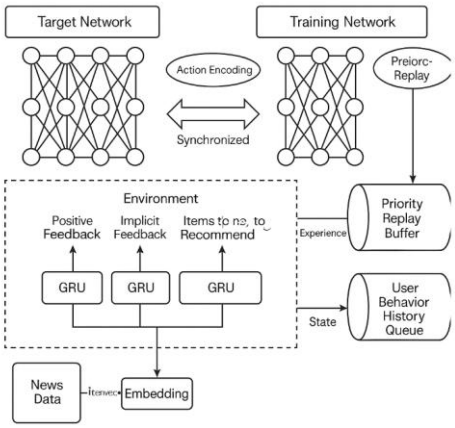

Figure 1. Architecture of the DAMDRL model[17]

Prior recommendation systems predominantly emphasize positive user feedback, neglecting negative and implicit behavioral feedback. In the context of news recommendations, positive feedback alone does not ascertain user interest in the news, and such feedback typically encounters issues of data sparsity, resulting in limited incentives for reinforcement learning. This study aims to model user behavior using positive, negative, and implicit feedback. Implicit feedback encompasses the temporal aspects of news and user actions such as deep reading, retweeting, and commenting, and eventually integrates numerous feedback sources with the items to be recommended to establish the recommendation environment.

In this study, we employed a GRU network to extract three components of features, specifically the sequence of user activities. The feature extraction of the three vectors is limited to a length of k. Feature extraction initially employs item2vec to generate the news vector representation, and subsequently encodes the user's feedback sequences with three distinct GRU encoders. Ultimately, the recommended items are encoded alongside the behavioral sequences of the user to create the environmental feedback necessary for reinforcement learning, as illustrated in Fig. 2. The encoding of a user's behavioral sequence is used to generate the environmental feedback necessary for reinforcement learning [18].

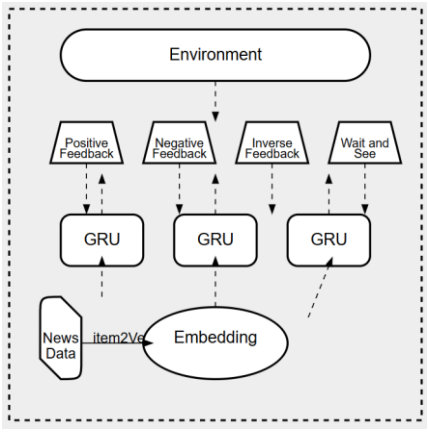

Figure 2. Sequential encoding of user behavior

In deep reinforcement learning, to acquire an independent and identically distributed dataset, user behavior that is continually correlated is segmented using the approach of "initially gathering an experience pool and subsequently conducting computations." This study employed an autonomous storage pool to preserve the distribution of user behavior data. Deep reinforcement learning employs an experience-pooling approach to store empirical data as quaternions (s_t, a_t, r_t, s_{t+1}). Subsequently, samples are taken based on a specific strategy, and the neural network parameter θ is adjusted to approximate the Q-function. $Q(s, a;\theta)$ was approximately equal to $Q^*(s, a;\theta)$. This study employed a method of constructing an empirical pool with preference caching to attain non-random data sampling [24]. Preferential caching assesses the significance of each empirical data source and computes the sampling probability based on the evaluated importance metric when the data enter the queue, determined as follows:

$$P(i) = \frac{p_i}{\sum_k p_k} \quad (1)$$

The value-based reinforcement learning system calculates the TD error at each step [19], [20], which is particularly appropriate for continuous recommendation prediction contexts. During empirical updating, data with the highest error are favored for replay because of their higher probability. δ_i represents the TD error of the i_{th} experience, indicating that the modification of each sample in the experience pool is governed by the regularized weight of the TD error. This updating strategy renders the empirical data interdependent, resulting in the repeated utilization of high-weight data after several rounds, whereas low-weight data progressively diminishes. This research addresses the cache update imbalance issue in experience playback by introducing the coefficients γ and bias β, which guarantee that all experience data are updated during Q-network training and that low-priority data are retained. The probability of a dominant cache is expressed as follows:

The dual-layer Q-network module utilizes enhanced experience for training, calculating the current user's numerous feedback sequences, and the prospective value functions between items $Q^*(s, a, \theta)$. This module involves reinforcement learning algorithms that utilize dominant playback experiences. The user experience data are extracted from the cache pool and the user feedback history cache, and thereafter predicted, and the parameter θ is changed based on the prediction outcomes. This study employed a masking mechanism to ensure the non-repetitiveness of the recommended items in the recommender system. Specifically, it initializes a mask vector $M = \{m_1, m_2, m_3, ..., m_k \mid m_i \in \{0,1\}\}$, where mask length k corresponds to the size of the action space. Following the computation of the action Q-network, the action space is concatenated with M, allowing the determination of the final available actions, which are then selected based on their Q values.

3. Results and Discussion

Three assessment metrics were selected to assess the quality of the recommendation lists: hit ratio (HR), mean reciprocal rank (MRR), and normalized discounted cumulative gain (NDCG). The HR indicates the inclusion of the item clicked by the user within the recommendation sequence, as defined in Equation (3).

$$\mathrm{HR}@k = \frac{1}{k}\sum_{i=1}^{k}\mathrm{hit}(i) \tag{3}$$

Here, k represents the length of the recommendation list, and the hit(i) function denotes the presence of the selected item in the recommendation sequence, yielding a value of 1 if it is present and 0 otherwise. The MRR indicator signifies the placement of the recommended project within the user recommendation list, highlighting the positional relationship. The definition is presented in equation (4), where k represents the length of the recommendation list and signifies the position of the user's actual accessed item inside the recommendation list; if the item is absent from the recommendation sequence, then p is infinite.

$$\mathrm{MRR}@k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{p_i} \tag{4}$$

Discounted cumulative gain (DCG) indicates the presence of the user's preferred goods in the highest ranks of the recommendation list. DCG requires normalization for direct comparison. Organize all recommended products in a specified order, select the top K items, and compute their DCG. Subsequently, the DCG is divided by the DCG corresponding to the desired state to obtain NDCG, a value ranging from 0 to 1, defined as

$$\mathrm{NDCG}@k = Z_k\sum_{i=1}^{k}\frac{2^{r_i}-1}{\log_2(i+1)} \tag{5}$$

r_i represents the "rank relevance" of the i_{th} position, where r_i equals 1 if the item at that place is included in the test set, and 0 if it is not; z_k is a normalization coefficient representing the inverse of the cumulative summation calculation as expressed in equation (5) under optimal conditions (when $r_i = 1$).

Table 1 presents the experimental outcomes of the DAMDRL model alongside all baseline approaches on the news dataset evaluated using three metrics: HR@10, MRR@10, and NDCG@10. Table 1 indicates that when HR@10 was used as the assessment metric, the two supervised learning models utilizing neural networks GRU4Rec and SLi-Rec exhibited superior performance, whereas the unaltered DQN model demonstrated a diminished hit rate. The DAMDRL model enhanced the DQN model. The sequence-oriented measures MRR@10 and NDCG@10 indicate that DAMDRL significantly outperformed the three neural network models, NCF, GRU4Rec, and SLi-Rec, together with the unenhanced DQN reinforcement learning model. Experiments demonstrate that the DAMDRL model introduced in this study exhibits strong performance in terms of recommendation accuracy and ranking.

Table 1. Comparative experimental results

Methodologies	HR@10	MRR@10	NDCG@10
SVD++	0.38105	0.19401	0.11597
NCF	0.59171	0.38729	0.39812
GRU4Rec	0.73882	0.48101	0.41101
SLi-Rec	0.68491	0.42635	0.47539
DQN	0.31023	0.43916	0.44379
DAMDRL	0.60177	0.49348	0.50181

This study introduces versions of the DAMDRL model, namely DAMDRL - 1 and DAMDRL - 2, to assess the influence of each module on the recommendation effect. The DAMDRL-1 model employed an average sampling method for experience acquisition to assess the influence of dominant sampling on the model. The DAMDRL-2 model employs only a greedy algorithm, omitting the loop-avoidance exploration strategy, to assess the effect of the inference exploration approach on the model. The ablation experiments utilized an identical dataset and evaluation measures as in the comparative experiments. The experimental results are presented for three models: DAMDRL, DAMDRL-1, and DAMDRL-2.

Table 2. Results of ablation experiments

Methodologies	HR@10	MRR@10	NDCG@10
DAMDRL-1	0.59021	0.47834	0.41032
DAMDRL-2	0.58911	0.48921	0.37079
DAMDRL	0.60177	0.49348	0.50181

In metrics HR@10 and MRR@10 (Table 2), the application of the underlying sampling method and exploration strategy did not significantly alter the recommendation accuracy; however, in the ablation experiment utilizing NDCG@10 as the metric, a substantial variation in accuracy was observed. This experiment demonstrates that the DAMDRL framework introduced in this study enhances the recommendation hit rate through a reinforcement learning approach, with key improvements in the caching and exploration strategies primarily targeting the sorting performance post-recommendation.

4. Conclusions

This paper proposes dynamic action-masking deep reinforcement learning to enhance recommender systems' capacity to address fluctuating user expectations in news recommendation scenarios, utilizing the deep reinforcement learning DQN method to better assess users' short-term interests. Reward design, characterized by dynamic masking and regional masking, accelerates model training and enhances the performance of the recommender system in the news recommendation sector. The experimental findings indicate that the recommendation accuracy of DAMDRL on the news dataset is on par with contemporary neural network recommendation methods and surpasses existing state-of-the-art recommendation algorithms in terms of ranking performance. Future studies should explore the feasibility of implementing reinforcement learning in the domain of conversational news recommendations.

References

[1] Y. Ding, B. Wang, X. Cui, and M. Xu, "Popularity prediction with semantic retrieval for news recommendation," *Expert Syst Appl*, vol. 247, 2024, doi: 10.1016/j.eswa.2024.123308.

[2] D. R. Liu, Y. Huang, J. J. Jhao, and S. J. Lee, "News recommendations based on collaborative topic modeling and collaborative filtering with generative adversarial networks," *Data Technologies and Applications*, vol. 58, no. 1, 2024, doi: 10.1108/DTA-08-2022-0315.

[3] Z. Y. Poo, C. Y. Ting, Y. P. Loh, and K. I. Ghauth, "Multi-Label Classification with Deep Learning for Retail Recommendation," *Journal of Informatics and Web Engineering*, vol. 2, no. 2, 2023, doi: 10.33093/jiwe.2023.2.2.16.

[4] C. Wu, F. Wu, Y. Huang, and X. Xie, "Personalized News Recommendation: Methods and Challenges," *ACM Trans Inf Syst*, vol. 41, no. 1, 2023, doi: 10.1145/3530257.

[5] M. Li and L. Wang, "A Survey on Personalized News Recommendation Technology," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2944927.

[6] Y. Pan, "Design and research of news recommendation system based on perceptron model in big data era," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, 2024, doi: 10.2478/amns.2023.1.00365.

[7] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, 2022, doi: 10.47065/bits.v4i1.1670.

[8] M. Zhang, G. Wang, L. Ren, J. Li, K. Deng, and B. Zhang, "METoNR: A meta explanation triplet oriented news recommendation model," *Knowl Based Syst*, vol. 238, 2022, doi: 10.1016/j.knosys.2021.107922.

[9] C. Song, K. Shu, and B. Wu, "Temporally evolving graph neural network for fake news detection," *Inf Process Manag*, vol. 58, no. 6, p. 102712, Nov. 2021, doi: 10.1016/j.ipm.2021.102712.

[10] T. Qi, F. Wu, C. Wu, Y. Huang, and X. Xie, "Privacy-preserving news recommendation model learning," in *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, 2020. doi: 10.18653/v1/2020.findings-emnlp.128.

[11] W. Zhang, "Design of news recommendation model based on sub-attention news encoder," *PeerJ Comput Sci*, vol. 9, 2023, doi: 10.7717/PEERJ-CS.1246.

[12] G. Zheng *et al.*, "DRN: A deep reinforcement learning framework for news recommendation," in *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, 2018. doi: 10.1145/3178876.3185994.

[13] D. Andra and A. B. Baizal, "E-commerce Recommender System Using PCA and K-Means Clustering," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, pp. 57–63, Feb. 2022, doi: 10.29207/resti.v6i1.3782.

[14] N. Yanes, A. M. Mostafa, M. Ezz, and S. N. Almuayqil, "A machine learning-based recommender system for improving students learning experiences," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3036336.

[15] L. Luo, N. Zhao, Y. Zhu, and Y. Sun, "A* guiding DQN algorithm for automated guided vehicle pathfinding problem of robotic mobile fulfillment systems," *Comput Ind Eng*, vol. 178, 2023, doi: 10.1016/j.cie.2023.109112.

[16] Y. Zhang, C. Li, G. Zhang, R. Zhou, and Z. Liang, "Research on the Local Path Planning for Mobile Robots Based on PRO-Dueling Deep Q-Network (DQN) Algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140842.

[17] X. Lu *et al.*, "Deep Augmented Metric Learning Network for Prostate Cancer Classification in Ultrasound Images," *IEEE J Biomed Health Inform*, vol. 29, no. 3, pp. 1849–1860, Mar. 2025, doi: 10.1109/JBHI.2024.3396424.

[18] W. Shu, K. Cai, and N. N. Xiong, "A Short-Term Traffic Flow Prediction Model Based on an Improved Gate Recurrent Unit Neural Network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16654–16665, Sep. 2022, doi: 10.1109/TITS.2021.3094659.

[19] G. Pang *et al.*, "Efficient Deep Reinforcement Learning-Enabled Recommendation," *IEEE Trans Netw Sci Eng*, vol. 10, no. 2, 2023, doi: 10.1109/TNSE.2022.3224028.

[20] A. Brini, G. Tedeschi, and D. Tantari, "Reinforcement learning policy recommendation for interbank network stability," *Journal of Financial Stability*, vol. 67, 2023, doi: 10.1016/j.jfs.2023.101139.