Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.33189

Development of Intelligent Advertising Decision Based on Reinforcement Learning

Zichun Hu

Department of Queen Mary School Hainan, Beijing University of Posts and Telecommunications, Lingshui, China
jp2024213928@qmul.ac.uk

Abstract. With the rapid development of the digital advertising industry, the advertising system has become a link between the market and consumers, and also a win-win situation. The traditional advertising recommendation method can only complete the matching of advertising in a short time, which cannot meet the current dynamic matching needs, and the matching efficiency is low. Based on the application of reinforcement learning, a machine learning method in advertising recommendation, this paper analyzes the modeling methods in this field through Multi-armed Bandit, DRL deep reinforcement learning (Deep Q-Network (DNQ), Advantage Actor Critic (A2C), Proximal Policy Optimization (PPO)), policy gradients and Actor Critic algorithms, and summarizes the common data sets in this field. Finally, this article will further analyze the current problems in this field and propose corresponding solutions. The analysis results show that at present, the reinforcement learning of advertising decision-making has achieved some results, but there are still many problems, such as multi-objective imbalance, exploration and utilization difficult to balance, and privacy and fairness. The use of quantitative multi-objective learning framework and user-oriented exploration strategy, the application of Federated learning and differential privacy will be the feasible methods to solve the problem.

Keywords: strengthen learning, intelligent advertising decision, multi-armed bandit, DRL deep reinforcement learning, policy gradient and actor critical algorithm.

1. Introduction

With the rapid development of the digital economy, online advertising has become an important part for enterprises to achieve precision marketing and business growth, for consumers to obtain information, and for the circular development of the social economy. According to relevant data, the scale of the global advertising market has exceeded US $1.19 trillion in 2025, with a year-on-year increase of 8.9% [1]. Among them, personalized advertising revenue customized by combining users' historical search records, context information and shopping preferences accounts for more than 60% of the total revenue of digital advertising [2]. The core of the advertising system is to accurately deliver the content that users need according to their preferences, so as to maximize the brand revenue.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.33189

The traditional advertising recommendation system can generally be divided into two categories. One category is content-based advertising recommendation and collaborative filtering (mainly used before 2010), which recommend advertising content for users according to their superficial preferences. There is usually a cold start problem (for new users, there is a lack of search data, and advertising content cannot be accurately delivered); The other is the recommendation method based on supervised learning. Its main feature is that it takes user click through rate and satisfaction as the main evaluation criteria. However, it is often difficult to adapt to the rapidly changing user scenarios, resulting in irrelevant recommended content and lack of information [3].

Reinforcement learning (RL) maximizes the cumulative rewards by studying the interaction between decision-makers and the environment, which is suitable for the flexibility, dynamic and long-term characteristics of advertising recommendation. This paper will analyze and summarize the current mainstream modeling methods and commonly used data sets, analyze the existing problems, and provide feasible solutions to these difficulties. These discussions and analyses are of great significance to promote the application of reinforcement learning in personalized advertising recommendation and improve advertising revenue and user experience.

2. Advertising recommendation method based on reinforcement learning

2.1. Advertising recommendation based on multi-armed bandit

Multi-armed Bandit is the mainstream method for intelligent advertising recommendation in the cold start phase. For example, the DeepLinUCB multi-armed bandit machine algorithm uses the representation ability of deep neural network to transform the original context information in the return function, predict the reward expectation of advertising decision through the context characteristics, and then realize the advertising recommendation in the cold start stage [4]. Its work flow chart is shown in Figure 1. This method has been adopted by the Taobao cold start decision module. In addition, the EvoClusterBandit framework realizes active environment partition through incremental clustering, completes region specific optimization by combining with the upper confidence limit algorithm, and avoids strategy fragmentation through cross regional knowledge migration, and then combines unsupervised clustering with context slot machines. Its workflow is shown in Figure 2. EvoClusterBandit framework provides an effective solution for dynamic decision-making under an unbalanced environment in the cold start phase, and better captures the potential changes of user preferences in the cold start phase [5].

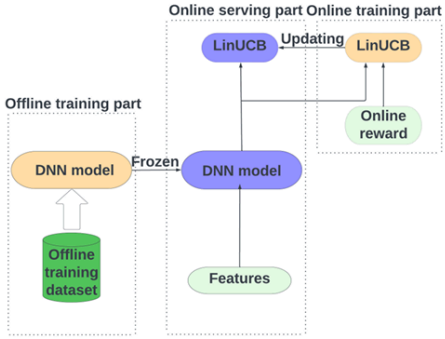

Figure 1. DeepLinUCB multi-armed bandit machine algorithm work flow chart [4]

DOI: 10.54254/2753-8818/2026.33189

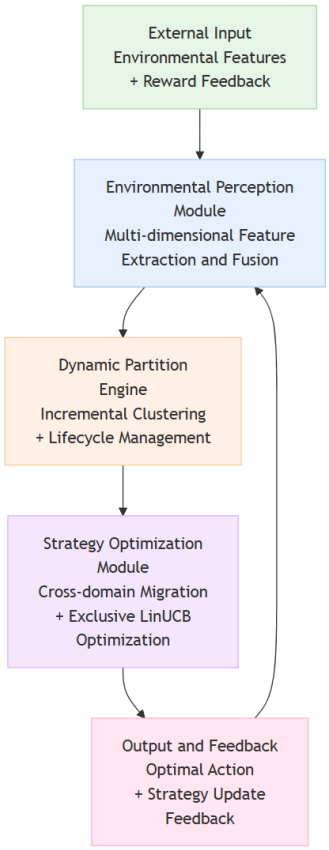

Figure 2. EvoClusterBandit framework work flow chart [5]

2.2. Advertising recommendation based on DRL deep reinforcement learning

With the deepening of reinforcement learning research, DRL deep reinforcement learning is becoming an important method to deal with personalized advertising decision-making. Its core is to regard recommendation as a sequential decision-making problem, which can optimize long-term benefits and apply to dynamic scenarios with user preferences changing. For example, the hybrid model of Collaborative Filtering (CF) and Reinforcement Learning (RL) is reduced to processing by using matrix analysis through integrating data preprocessing, algorithm interaction, evaluation feedback and other links, realizing the combination of static preference mining and dynamic real-time optimization [6]. The execution flow chart is shown in Figure 3. In addition, the multi task off-line reinforcement learning model (MTORL) designs a causal state catcher, introduces causal attention mechanism, and uses the multi task learning method to decode behaviors and benefits. It can simultaneously solve the two problems of channel recommendation and budget allocation while capturing the changes of users' interests and preferences over time [7]. Its implementation flow chart is shown in Figure 4.

DOI: 10.54254/2753-8818/2026.33189

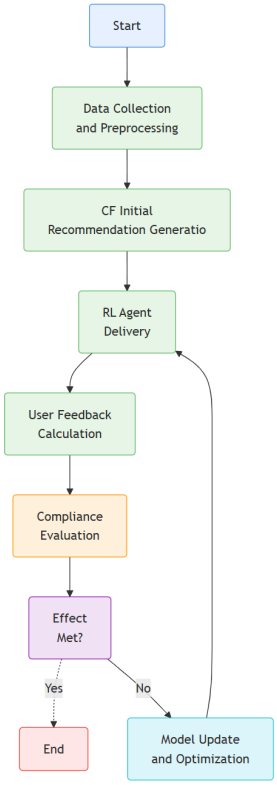

Figure 3. Collaborative filtering and reinforcement learning hybrid model flow chart [6]

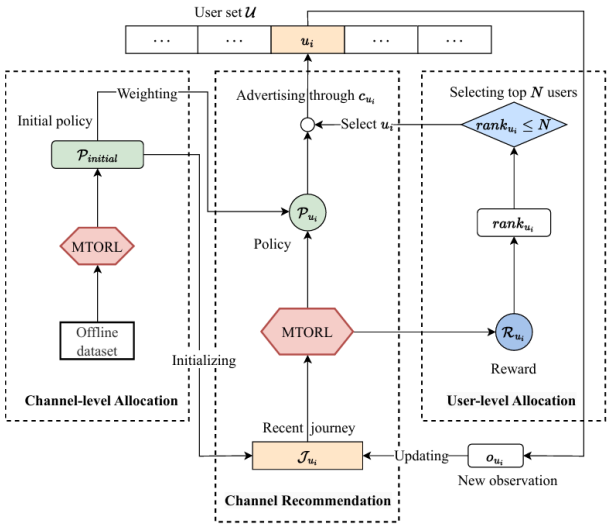

Figure 4. Multi task off-line reinforcement learning model flow chart [7]

2.3. Advertising recommendation based on strategy gradient and actor critical algorithm

The application of real-time bidding scenarios in advertising recommendation is very extensive. In short, it means that at the moment when an advertising space is visited by a user, multiple merchants bid simultaneously to place it, and the one with the highest price wins. For example, HMMCB, as a hierarchical multi-agent learning framework, can dynamically allocate budgets among channels by

DOI: 10.54254/2753-8818/2026.33189

building a diffusion model constrained by the cost of a single click. Its workflow architecture is shown in Figure 5. At the same time, it solves the problem of offline learning extrapolation error caused by distributed external work, which is suitable for real-time bidding scenarios [8]. In addition, the actor critic reinforcement learning (ACRL) model learns the Gaussian distribution by introducing additional context auxiliary information of media and users to simulate the audience's decision-making behavior in a more realistic bidding scenario. It realizes the quantitative evaluation of advertising exposure opportunities. Its workflow architecture is shown in Figure 6. This model has been deployed on the Tencent online display advertising platform and has produced practical application value [9].

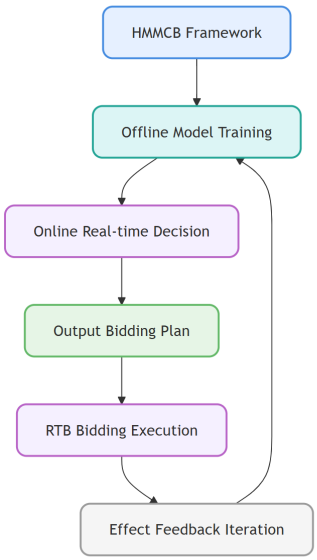

Figure 5. HMMCB framework work flow chart [8]

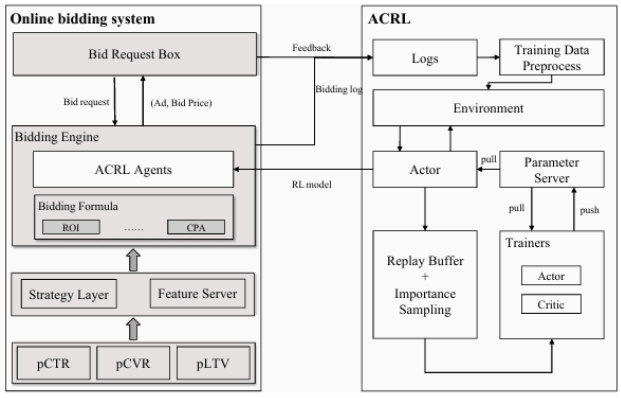

Figure 6. ACRL model work flow chart [9]

3. Advertising recommendation dataset

In reinforcement learning, a data set is a tuple of recorded data including status, actions, rewards, next status, and whether to end when the decision-maker interacts with the environment. The data

DOI: 10.54254/2753-8818/2026.33189

sets commonly used in this field mainly include Taobao-MM data set, Tenrec data set and CriteoPrivateAds data set.

The Taobao-MM data set is a large-scale data set with long behavior sequence and multimodality in the field of advertising recommendation. It is built based on Taobao advertising business log, covering 99 million samples, 8.79 million users, 35.4 million unique products and 128-dimensional SCL embedded vector. Its data is composed of user characteristics (user ID, age, gender, city of residence), product characteristics (product ID, category, city of origin) and click labels (1 represents that the user clicked on the advertisement, 0 represents that the user did not click on the advertisement).Taobao-MM as a privacy desensitization public data set, provides important data support for multimodal recommendation and user interest preference modeling [10].

The Tenrec data set is a large-scale, multi-purpose recommendation system benchmark data set developed by Tencent and other institutions. It solves the problems of small scale and single scene of the traditional data set, adapts to the research of multiple types of recommendation tasks, and has been open-source and strictly privacy desensitized. Its core overall scale covers about 5.02 million users and a total of 1.4 billion and interactive behaviors, including user ID, item ID, clicks, exposure and other information. This data set is suitable for the research of advertising recommendation Click-Through Rate (CTR) prediction, cold start and other fields [11].

CriteoPrivateAds data set is the world's richest real-world anonymous bidding data set released by the French Criteo company. Based on the anonymous version of Criteo's production environment log, it contains 100 million advertising display data, covers the production log for 30 consecutive days, and has more than 100 related features, including line ID, advertising campaign ID, and publisher ID. It supports multi-model training and configuration testing, making the performance of the model more suitable for the real production environment [12].

4. Challenges and solutions

According to the existing mainstream methods and data sets, it can be applied to modules such as accelerated cold start, real-time dynamic bidding, and long-term revenue optimization, but there are still problems to be solved in other aspects.

4.1. Multi-objective disequilibrium problem

In the intelligent decision-making of advertising, the system needs to take into account multiple competitive optimization goals such as click-through rate and conversion rate, which have huge differences in feedback frequency. This will lead to the model automatically being biased to the goals that are easier to optimize and more rewarding in training, while ignoring other key goals. Finally, the goal is unbalanced. For example, many advertisements only care about the indicator of click-through rate, but often ignore issues such as fairness and diversity. This will result in the proliferation of headline parties and low-quality advertisements, leading to the decline of user experience.

Using a quantitative multi-objective reinforcement learning framework and optimizing the design of the reward function will be a feasible solution to the problem of multi-objective imbalance. The next item recommendation task is modeled as a multi-objective Markov decision process, and multiple objectives are optimized at the same time. Furthermore, differentiable rewards are designed for each goal, so that the rewards of each goal are linked with each other, and the weight can be adjusted flexibly. Different goals can be adjusted by adjusting the weight vector at different demand stages, so as to achieve flexible control of multi-objective balance [13].

DOI: 10.54254/2753-8818/2026.33189

4.2. Exploring and utilizing is hard to balance

Exploration means launching new advertisements and new people to collect feedback and discover potential high-value opportunities; Utilization refers to the continuous delivery of the currently known maximum click through rate, conversion rate and maximum revenue advertising portfolio based on historical data. Due to the continuous changes in user interest and market environment, as well as the delay in the real rewards of advertising transformation, the model is vulnerable to external interference and noise. The model cannot judge whether a certain exploration is really valuable, which leads to the fact that exploration and utilization cannot be combined, resulting in sharp fluctuations in short-term earnings and serious waste of budget.

Using user-oriented exploration strategy will be the core solution to solve the difficulty of balancing exploration and utilization. To put it simply, users are divided into three categories according to their activity, high active users, medium active users and low active users. Different exploration intensities are matched for different groups, such as high exploration for low active users and high utilization for high active users. At the same time, the dual core component architecture is used to accurately quantify the uncertainty of different user groups while focusing on serving user groups with specific activity, taking into account the diversity of exploration and the stability of training [14].

4.3. Privacy and fairness issues

Privacy problem means that the system over collects and stores user behavior data for training and optimization strategies; Fairness refers to the systematic and reproducible discrimination of advertising decision-making systems against groups of different genders, regions and ages. This will cause the identification of users' identities and the accurate construction of portraits, and damage the fairness of users' choices and market competition.

Federated learning and differential privacy will be important solutions to privacy and fairness. The data is available and invisible through federated learning. All parties do not share the original user advertising behavior data, but only send the parameters updated by the model to the central server. The server generates a global recommendation model after weighted average of all parameters, and then sends it to all parties for iterative optimization. Differential privacy limits the fluctuation range of the model gradient, and then adds controllable Gaussian noise to the gradient, so that the uploaded parameters cannot be restored to the user's specific behavior information [15]. While improving the advertising effect, it also adheres to the bottom line of privacy and the principle of fairness.

5. Conclusion

The summary analysis shows that reinforcement learning is a long-term revenue optimized machine learning method, which is closely related to the long-term and dynamic demand of intelligent decision-making in advertising recommendation. Its mainstream methods, Multi-armed Bandit machines, DRL deep reinforcement learning, strategy gradient and actor critic algorithm, have been applied in important fields, but there are still many difficulties in the design of reward function, the balance between exploration and application, and the protection of user privacy. In order to solve the above problems, this paper reviews the multi-objective learning framework, user-oriented exploration strategy, differential privacy and federated learning, so as to achieve the coordination and unity of user demand, business value and social value in the era of digital economy.

DOI: 10.54254/2753-8818/2026.33189

With the continuous development of reinforcement learning and large model technology, future research will focus on the semantic understanding ability of using large language model to strengthen the accuracy of the advertising intelligent decision context system, and accurately capture the deep meaning of the text and the user's intention. In the future, with the continuous improvement of AI ethics and privacy protection, fairness constraints should also run through the whole process of model creation, making the intelligent decision-making of advertising recommendation develop continuously and stably.

References

[1] WARC. (2025). Global ad trends: Media's new normal. https: //www.warc.com
[2] Dentsu. (2026). Global ad spend forecast 2026. https: //www.dentsu.com
[3] Zhu, L., Sun, M., & Yu, L. (2023). Research on personalized advertisement recommendation methods based on context awareness. Journal of Advanced Computing Systems, 3(10), 39–53.
[4] Shi, Q., Xiao, F., Pickard, D., et al. (2023). Deep neural network with LinUCB: A contextual bandit approach for personalized recommendation. In Companion proceedings of the ACM Web Conference 2023 (pp. 778–782).
[5] Ren, H. (2025). EvoClusterBandit: Adaptive partitioned bandit algorithm for dynamic environments with latent variable modeling. ITM Web of Conferences, 80, Article 01011.
[6] Iyer, D., Sharma, D., Singh, N., et al. (2022). Enhancing digital advertising through AI-powered personalization: Leveraging reinforcement learning and collaborative filtering algorithms. International Journal of AI ML Innovations, 11(8).
[7] Liu, L., Wang, W., Zhang, C., et al. (2025). Multi-task offline reinforcement learning for online advertising in recommender systems. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Vol. 2, pp. 4635–4646).
[8] He, S., Yu, C., Lin, Q., et al. (2024). Hierarchical multi-agent meta-reinforcement learning for cross-channel bidding. IEEE Transactions on Knowledge and Data Engineering, 37(3), 1241–1254.
[9] Yuan, C., Guo, M., Xiang, C., et al. (2022). An actor-critic reinforcement learning model for optimal bidding in online display advertising. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 3604–3613).
[10] Wu, B., Yang, F., Chan, Z., Gu, Y. R., Feng, J., Yi, C., ... & Zheng, B. (2025). MUSE: A Simple Yet Effective Multimodal Search-Based Framework for Lifelong User Interest Modeling. arXiv preprint arXiv: 2512.07216.
[11] Yuan, G., Yuan, F., Li, Y., et al. (2022). Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. Advances in Neural Information Processing Systems, 35, 11480–11493.
[12] Sebbar, M., Odic, C., Léchine, M., Bissuel, A., Chrysanthos, N., d'Amato, A., ... & Vono, M. (2025). CriteoPrivateAds: A Real-World Bidding Dataset to Design Private Advertising Systems. arXiv preprint arXiv: 2502.12103.
[13] Stamenkovic, D., Karatzoglou, A., Arapakis, I., et al. (2022). Choosing the best of both worlds: Diverse and novel recommendations through multi-objective reinforcement learning. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (pp. 957–965).
[14] Zhang, C., Chen, S., Zhang, X., et al. (2024). Reinforcing long-term performance in recommender systems with user-oriented exploration policy. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1850–1860).
[15] Adnan, M., Kalra, S., Cresswell, J. C., et al. (2022). Federated learning and differential privacy for medical image analysis. Scientific Reports, 12(1), Article 1953.