Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230304

A study of real-time advertising bidding based on reinforcement learning

Jiayuan Zhang

Suzhou City University, Suzhou, 215104, China

2630282232@qq.com

Abstract. The purpose of real-time bidding (RTB), which has grown to be a significant component of Internet advertising, is to properly forecast the likelihood of a user clicking on an ad and choose the best bid for that impression. Reinforcement learning (RL) has emerged as a promising approach for RTB bidding strategies due to its ability to learn from feedback and optimize performance over time. This paper surveys RL-based real-time advertising bidding, introduces several main reinforcement learning-based real-time advertising bidding strategies, and explains the advantages of each of these strategies. Furthermore, the author analyzes the current trends and which strategies are combined and why. Overall, this survey sheds light on the potential of RL-based bidding to enhance the effectiveness and efficiency of RTB advertising while also offering information about the current state of the field and future research directions.

Keywords: real-time bidding, reinforcement learning, bidding strategy, bid optimization.

1. Introduction

RTB is the current mainstream Internet advertising model. RTB is different from the traditional advertising model of selling advertising space on a daily or monthly basis. Advertisers obtain the opportunity to display advertisements in the real-time bidding system through public bidding. RTB improves the accuracy of advertising by restricting various conditions. Therefore, the bidding model of RTB not only satisfies the interests of advertisers to maximize the precise investment, but also increases the value of advertising space through bidding, increasing the income of network owners.

In the past, many existing research problems regarded the bidding strategy as a static optimization strategy [1], but the complex situation in reality is that there are tens of millions of competitors participating in the bidding for the same advertising space, and due to the existence of temporary, the possibility of changing the delivery plan leads to a highly dynamic bidding environment, so the bidding strategy cannot meet the needs of advertisers. Real-time advertising bidding algorithms based on RL have become an active research topic due to their high efficiency and flexibility in dealing with complex data characteristics and environmental changes. Therefore, this paper aims to provide researchers with an overview of RL-based real-time advertising bidding algorithms by conducting a comprehensive survey and analysis of existing RL-based real-time advertising bidding algorithms. This paper will cover the business process of RTB advertising, reinforcement learning basics, applications of reinforcement learning in real-time ad bidding, and current challenges and future directions.

DOI: 10.54254/2755-2721/8/20230304

2. Background information

2.1. Real-time advertising bidding

Real-time advertising bidding means that through the real-time bidding mechanism, advertisers can conduct real-time bidding for their advertisements, and use the algorithm system to select the advertisements that best meet the advertising positions and target audiences on the advertising exchange platform. RTB uses a second-price auction as its format [2]. Bidders can enter their own bids in real time during the real-time bidding process, and the eventual winning bidder only has to pay the second-highest bid.

The standard RTB ad distribution process is shown in Fig.1 [3]. When a user accesses a website, the ad information on the page is sent to the Ad Exchange (AdX) via the Supply-Side Platform (SSP). Simultaneously, the cookie label acquired by the web browser employed by the user is fed into a data management platform (DMP) for the purpose of analysis, and the resulting user attribute label is subsequently transmitted to AdX. AdX sends the user information to all advertisers connected to the Demand-Side Platform (DSP) of the advertising agency, initiating real-time bidding. In this process, the DSP with the highest bid and a matching user attribute label is selected to display the advertisement. The entire process, including ad request, bidding, and ad display, typically takes place within 10-100 milliseconds.

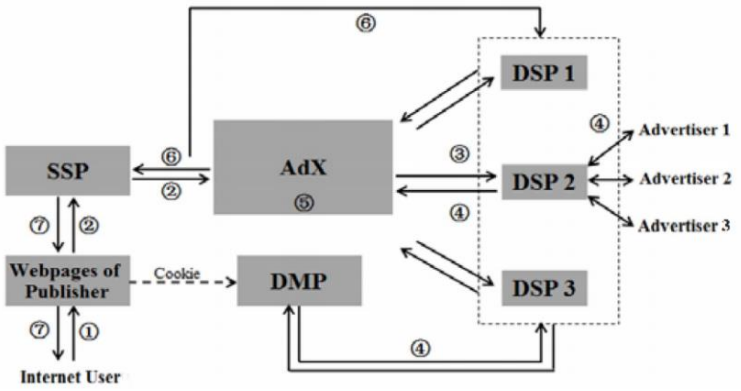

Figure 1. The business process of RTB ad delivery.

From the above example, it can be concluded that when two different people log in to the same page, they will see completely different advertisement content. And because these advertisements are delivered based on the analysis of different user behaviors, they will be more in line with the user's recent needs and interest attributes, and are basically useful information for users. Therefore, users are more willing to click on the advertisement, thereby increasing the ROI of the advertisement, which is beneficial to the advertiser. For the media, what they sell is no longer a fixed advertising space, but the users behind the advertising space, and each user is valuable. Media are no longer constrained by limited resources and thus can increase their earnings.

2.2. Reinforcement learning

Reinforcement learning is a subfield of machine learning whose main goal is to design autonomous decision-making agents so that they can learn strategies to take actions in different states to maximize cumulative rewards through interaction with the environment [4]. In reinforcement learning, the agent continuously monitors the state of the environment, interacts with it to receive rewards, and then modifies its strategy in light of the rewards and the most recent state to continuously improve its action choices. In reinforcement learning, the actions taken by the agent are not only to solve the current problem, but also need to consider the long-term reward.

In real-time advertising bidding, RL can continuously interact with the environment through environment modeling, defining reward functions, strategy learning and real-time bidding, and

DOI: 10.54254/2755-2721/8/20230304

feedback based on actual revenue, gradually optimizing the bidding strategy, so as to achieve better advertising delivery effect and higher returns.

3. The application of reinforcement learning algorithms in real-time advertising bidding

3.1. Value-based algorithms

The algorithms use a reinforcement learning algorithm to estimate the expected value of each ad slot and select the ad with the highest expected value to serve. Commonly used value function algorithms include Q-learning and SARSA.

Q-learning is an incremental update-based dynamic programming algorithm that uses a state-action value function (Q-function) to represent the expected reward of taking each action in each state [5]. During the learning process, the agent updates the Q-function by interacting with the environment and chooses the next action according to the updated Q-function. In Q-learning, the update of the Q function is based on the Bellman equation. That means, when the agent moves from the current state to the next state, it uses the Q-function to choose the best next action [6]. Then, the Q-function updates the current state's Q-value with the next state's reward and the expected maximum Q-value. By continuously iteratively updating the Q-function, the agent can eventually learn to choose the best action in the Markov decision process. For environments where the model is unknown or complex, Q-learning, as a model-free reinforcement learning algorithm, can only interact with the environment to obtain rewards and the next state. However, because the Q-learning algorithm assumes that there is only one optimal action, regardless of the impact of multiple actions, the Q-learning algorithm may overestimate the Q values of all optimal actions, resulting in the algorithm failing to converge. In addition, the Q-learning algorithm may lead to high computational complexity for problems with large state and action spaces, which is difficult to implement.

To address these issues with the Q-learning algorithm, the SARSA algorithm introduces a more conservative learning strategy. Different from Q-learning, the SARSA algorithm focuses on the Q-value of the next state and the next action after taking the current action in the current state and following a certain strategy, and then uses these values to update the Q value of the current action in the current state [7]. This strategy makes the SARSA algorithm more conservative because it avoids overestimating the Q value of the optimal action in the current state. At the same time, the SARSA algorithm is more suitable for use in the exploration process, because it can consider the impact of random actions, rather than just focusing on the current optimal action.

3.2. Policy gradient algorithm

Unlike value-based algorithms that learn the optimal value function, this algorithm directly optimizes the policy function to obtain a better policy and maximize rewards. Policy Gradient is the most basic algorithm in policy-based algorithms.

In the real-time advertising bidding problem, each ad request is designed for multiple ad slots and bidders, so ad delivery can be regarded as a continuous control problem. Specifically, the strategy function takes the current ad request, bidding information, and historical data as input, and outputs an ad delivery probability distribution. This delivery probability distribution can be regarded as a strategy for advertising delivery. Through the policy gradient algorithm, this policy function can be optimized so that the selected advertisement can maximize the overall revenue.

However, the policy gradient algorithm can deal with highly continuous, non-linear policy spaces by directly modeling and optimizing policies. But since the policy gradient algorithm directly learns the policy, it usually requires more training data to achieve good performance. Second, the policy gradient algorithm is prone to get stuck in a locally optimal solution because it only optimizes the policy function without considering the value function.

DOI: 10.54254/2755-2721/8/20230304

3.3. Actor-critic

The Actor-Critic (AC) framework is a framework that combines the advantages of both value-based and policy-based algorithms. Within the AC framework, the Actor component is tasked with acquiring and refining strategies, while the Critic component focuses on acquiring and refining value functions. These two components work in tandem to improve the overall quality of decision-making. Through continuous iterative learning, the AC framework can learn the best advertising selection strategy to maximize advertising revenue and return on investment. Unlike the Policy Gradient algorithm, which can only be updated for each round, the AC framework can be updated according to time steps, so the AC framework can use reward signals in a timely manner to adjust policies and value functions.

Representative algorithms in AC include Deep Deterministic Policy Gradient (DDPG) and Trust Region Policy Optimization (TRPO). DDPG is an algorithm that combines deep neural network and deterministic policy gradient, which can deal with the problem of continuous action space and high-dimensional state space, and also solves the problem of difficult convergence of traditional Q-learning algorithm in high-dimensional state space. DDPG uses the experience playback mechanism and the idea of the target network to improve the stability and convergence speed of the algorithm through offline training and update of the target network [8]. TRPO is a policy optimization algorithm based on the trust region, which proposes a monotonous policy improvement method [9], that is, every time the policy is updated, the reward function is guaranteed to be monotonous, so that the algorithm is easier to converge. During the optimization process, TRPO constrains the update of the strategy to ensure that each update is carried out within a certain trust region, thereby avoiding the problem of policy degradation caused by an excessively large update step.

4. RL-based real-time advertising bidding strategy

4.1. Model-free reinforcement learning strategy

Markov decision making (MDP) provides a widely used framework for modeling agent-environment interactions. In the context of MDP, it is possible to define either a state-value function or a state-action-value function in order to assess the effectiveness of a given strategy and make decisions accordingly by maximizing the value function. This method of value iteration [10] can obtain the optimal strategy in a relatively short time, and the convergence of the algorithm is widely guaranteed theoretically.

The strategy adheres to the concept of DRLB [11], and uses a model-independent RL framework to learn the optimal bidding factor generation strategy. This strategy first models the auction process as an MDP. The environment for the RTB system comprises both the RTB market and Internet users. At the start of each epoch, the bidding agent perceives the environment and determines its current state. Based on this, the agent takes an action, which corresponds to the bidding factor for that particular period. Subsequently, the agent computes the bid price for each received bid request utilizing a specific formula throughout this time period. At the end of the epoch, the agent uses actual costs and user feedback, including clicks, to generate instantaneous rewards. The optimal bidding factor generation strategy is then learned through the TD3 algorithm to maximize cumulative reward across the entire advertising period. Specifically, the advertising period is divided into time slots, and the optimal bidding factor for each slot is computed using the Double Delay Deep Deterministic Policy Gradient (TD3) algorithm [12]. The proposed strategy is aimed at adapting dynamically to the RTB environment by leveraging the estimated value of each impression and its associated bid factors for the corresponding period.

4.2. Model-based reinforcement learning strategy

Model-based reinforcement learning models [13] apply the MDP framework and aim to address the huge state-space transition dynamics that model-free methods may suffer from when learning bidding policies directly from experience. Therefore, the model strategy captures state transition and reward functions by modeling sales competition and user clicks respectively, and then uses dynamic

DOI: 10.54254/2755-2721/8/20230304

programming to derive an optimal bidding strategy. In addition, deep neural networks are used to process high-dimensional inputs and outputs to optimize placement decisions. Specifically, state value approximation using neural networks can effectively handle scalability issues caused by large auction volumes and campaign budgets [13]. This approach is designed to fit the differential between two continuous values.

4.3. Stochastic optimal control model strategy

Scholars proposed a new stochastic optimal control model [14], which can be used to solve the optimal bidding strategy problem in various real-world scenarios. This model models the auction sequence as a Poisson process [15], where the asking price at each auction is represented by a random variable that can follow almost any probability distribution. This model can be used to solve some practical problems, such as how to maximize the purchase of inventory with a given amount of cash under the framework of audience strategy. For computational considerations, this model tends to transform the problem into a minimization problem and introduces a value function to solve the problem. In this model, jumps in the Poisson process represent occurrences of auction requests, along with markers for some other variables, such as the best bids of other bidders and the occurrence of conversions. Through this stochastic optimal control model, bidding strategy problems can be better understood and solved in reality.

5. Conclusion

This paper introduces the application of reinforcement learning in real-time advertising bidding, which can improve advertising performance and reduce costs by learning optimal strategies. Specifically, through the learned optimal strategy, user needs can be better matched, and the click-through rate and conversion rate of advertisements can be improved. Even though real-time advertising bidding is facing many challenges at this stage, such as the challenge of data sparsity, future research can start with data preprocessing, data sampling and weighting, multi-modal data fusion, multi-task learning, and the fusion of other algorithms to improve Accuracy and advertising efficiency. In addition, this paper summarizes the existing research progress and main applications of reinforcement learning-based advertising real-time bidding strategies, and analyzes the advantages and limitations of different strategies to help readers better understand the practical applications and possible challenges of these methods. Finally, this article helps researchers and practitioners better understand the application of reinforcement learning in real-time advertising bidding and promote their innovative thinking.

References

[1] Zhang, W., Yuan, S. and Wang, J. (2014). Optimal real-time bidding for display advertising. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1077-1086.

[2] Edelman, B., Ostrovsky, M. and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. American economic review 97(1), 242-259.

[3] Yuan, Y., Wang, F. and Li, J., et al. (2014). A survey on real time bidding advertising. Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics. IEEE, 418-423.

[4] Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996). Reinforcement learning: A survey. Journal of artificial intelligence research 4, 237-285.

[5] Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. Machine learning 8, 279-292.

[6] Ohnishi, S., Uchibe, E. and Yamaguchi, Y., et al. (2019). Constrained deep q-learning gradually approaching ordinary q-learning. Frontiers in neurorobotics 13, 103.

[7] Wang, Y., Ye, Z. and Wan, P., et al. (2019). A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks. Artificial intelligence review

DOI: 10.54254/2755-2721/8/20230304

51, 493-506.

[8] Tian, S., Li, Y. and Zhang, X., et al. (2023). Fast UAV path planning in urban environments based on three-step experience buffer sampling DDPG. Digital Communications and Networks.

[9] Meng, W., Zheng, Q. and Shi, Y., et al. (2021). An off-policy trust region policy optimization method with monotonic improvement guarantee for deep reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems 33(5), 2223-2235.

[10] Chatterjee, K. and Henzinger, T. A. (2008). Value iteration. 25 Years of Model Checking: History, Achievements, Perspectives, 107-138.

[11] Wu, D., Chen, X. and Yang, X., et al. (2018). Budget constrained bidding by model-free reinforcement learning in display advertising. Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 1443-1451.

[12] Liu, M. J., Li, J. X. and Hu, Z. N., et al. (2020). A dynamic bidding strategy based on model-free reinforcement learning in display advertising. IEEE Access 8, 213587-213601.

[13] Cai, H., Ren, K. and Zhang, W., et al. (2017). Real-time bidding by reinforcement learning in display advertising. Proceedings of the tenth ACM international conference on web search and data mining, 661-670.

[14] Fernandez-Tapia, J., Guéant, O. and Lasry, J. M. (2017) Optimal real-time bidding strategies. Applied mathematics research express 2017(1), 142-183.

[15] Garman, M. B. (1976). Market microstructure. Journal of financial Economics 3(3), 257-275.