International Journal For Technological Research In Engineering
Volume 9, Issue 12, August-2022 *ISSN (Online): 2347 - 4718*

WEB AD-OPTIMIZATION TECHNIQUES AND CHALLENGES – A REVIEW

Vivek Singh[1], Charanjeet Singh[2]
Student[1], Assistant Professor[2]
Dept. of Electronics Engineering
DCRUST ,Murthal, Sonipat, India

Abstract— Online advertisements are one of the most common online platform methods and websites are the most popular place for showing advertisements online. Information searching techniques, satisfying the information need of users through personal user features i.e., services and information in timely manner as well as locally. It is very crucial in lessening the issue of information overload. With the latest good advancement in RL or "Reinforcement Learning", which is growing enthusiasm for development of RL related search methods of information? This reinforcement learning based approaches have two important advantages, first, it is possible to continually update search related data according to visitors' instantaneous responses, and second, it can increase the long duration accumulated reward expected from the past information based on user activities like click-through rate, satisfaction of the visitors, involvement of the user for long term. In this review we present an analysis of Reinforcement Learning techniques in web advertisements, optimisations.

Keywords: RL (Reinforcement Learning), UCB (Upper Confidence Bound), MDP (Markov Decision Process), DQN (Deep Q-Network), EXP3 (Exponential-weight algorithm), web-ads optimizations

1. INTRODUCTION

Massive amounts of data have been created by the World Wide Web's rapid expansion. The issue of information overload has gotten worse as a consequence. Therefore, a way to identify items that meet the needs of users' information in a timely manner and space have become very important, which has inspired three representatives' ways to search for in-depth information - search, recommendations, and online advertising. The search method releases query-related items, produces a list of values that match specific user requirement, as well as an online advertising the modification method is the same as the search and the recommendation expects the items to be presented with ads.

Advertising raises a product's profile among the general public. Using the internet, online advertising is a method of marketing and advertising that reaches people or prospective consumers. Using their mobile devices, consumers may access a broad variety of tailored marketing initiatives. On the user's device, advertisements are shown in certain areas. There are two options available to a user: either dismiss the advertising or click on it. The act of clicking on an advertising arouses curiosity. However, not all advertisements are targeted at the same demographic. Traditional advertising existed prior to the introduction of web advertising. Traditional advertising relies on the placement of advertisements in certain sections of newspapers and periodicals. Interested parties may buy this area, or it can be auctioned off, with the highest bidder taking it. Traditional advertising can only be measured by a rise in income, while with internet advertising, the first level of efficacy may be assessed by a user's mouse click.

Advertisers, publishers, and advertising agencies make up the three main players in the internet advertising system. Advertisers are companies, brands, and organizations looking to start a marketing campaign to promote their goods and/or services to the public. Publishers are the individuals or companies whose websites serve as the platform for third-party adverts. Advertisers and publishers work together via an advertising agency, which serves as an intermediary. Advertisers may use these services to create powerful marketing campaigns and track the effectiveness of their advertising. As an advertising agency, we assist publishers with the selection of which kind of advertisements they should show, where they should be placed, and at what frequency. Because of their massive user bases, Face book and Google are both publishers and advertising agencies, since these businesses have the resources and computing ability to transform user data into profit and launch targeted ad campaigns from their massive user databases.

The first step in figuring out a user's preferences is to see what he or she clicks on. Using this information, you may create personalized advertisements for your customers. This might lead to a poor return on investment since traditional advertising does not provide a means for understanding customer preferences. The success of an ad may be monitored on a weekly, monthly, or quarterly basis through online advertising. Use this information to change your target audience, budget. There is a great deal of versatility in this method. Online advertising, in addition to its flexibility, it delivers a big audience. In the 21st century, the internet has made it possible for everyone to communicate with each other on the same platform.

Both online and offline advertisements seek to determine the best audience (users) and ad matching possible given the unique contextual elements of each platform. The ultimate goal of every advertising system is to achieve this. Simply expressed, this is the comparable of creating an algorithm that accurately predicts whether a certain advertisement will be viewed positively or negatively by its intended audience based on user behaviour. According to a research, accurate forecasts of user reaction metrics may be advantageous for both

publishers and advertisers.

As a fundamental part of an online ad management system, ad placement is all about timing and making the correct selection at the appropriate moment for each ad. Classifying the target population, selecting the most effective advertising, tracking their performance, knowing when and how to distribute ads to users, and optimizing the costs borne by advertisers are all part of a successful ad campaign. Because of the ever-increasing amount, velocity, and diversity of data, computers are becoming more suited to do these jobs than people. Machine Learning is ideally suited for the goals listed above since it allows an online advertising system to learn and improve over time.

1.1 REINFORCEMENT LEARNING

Learning to associate events with actions is what reinforcement learning all about [1]. The two key components of RL are learning the mapping (policy learning) and formulating the scenarios (using mathematical models). Markov decision processes (MDP) and multi armed bandits (MAB) are the two primary contexts for issue formulation. MAB is just a straightforward model for the exploration/exploitation trade-off [2].

The Partially Visible Markov decision process (PVMDP) extends the MDP to the situation in which the system state is not always observable [3] [4] .

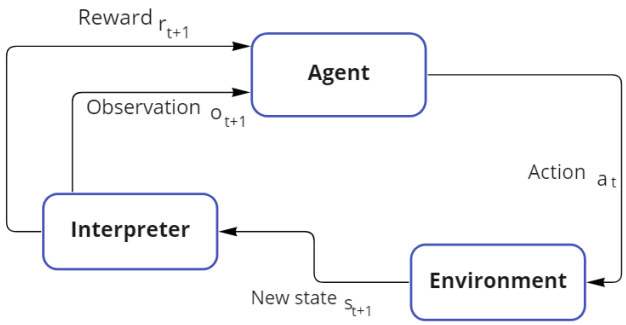

Fig 1. RL cycle.

The agent examines the condition of the environment and then tries to respond to something in the best way feasible. A state transition brought on by the action is translated into the possible reward, sent to agent along with fresh analysis of the surroundings [5].

1.2 MDP or "Markov Decision Process"

This is a classic example of linear decision making, which really is a stylized representation of RL problems [6].

The primary source of income for search results, which are frequently commercialized, is marketed research. In sponsored type of search, in addition to the natural search results, consumers also see paid adverts (ads)[7]. First and foremost, these advertisements provide customers with benefits such as product information, a sizable discount, etc. Second, given that the target demographic is reached by the marketers' marketing activities, these advertisements unquestionably add value to them[7]. Third, the search engine benefits from these advertisements as well since it will make money if consumers click on the adverts[7]. Simply Put, first of all, the advertisement is rated according to the rank score, which is calculated by multiplying the bid and ad quality score.

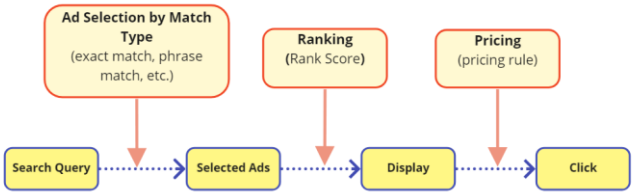

Fig 2.

It is preferable to choose ads after carefully weighing all of their potential effects, such as whether branded selection method itself can increase user clicks, marketer public assistance, and web search revenue [7].

Online advertising campaigns attracted the attention of many advertisers who are willing to advertise online business. One of the biggest problems advertisers faces, especially those who do not have more in online advertising, organizes their campaigns more effectively. To prepare the right campaign is necessary to choose the right target, so you are guaranteed a high level of acceptance users in ads. It is also required that the number of visits that meet the requirements to set up be high enough to cover advertiser campaigns.

2. EXPERIMENTAL EVIDENCES

Different categories exist for RL algorithms. A semi hierarchy of all the most popular contemporary template matching techniques is shown in Fig. 3 [5]. The algorithms are usually split into either model based and model free types before being further subdivided based on how the learning happens or what the algorithms are really learning internally [5].

Based on the sorts of experiences they may learn from; RL algorithms can be split into two groups [5]. These two machine learning algorithms are "on policy" and "off-policy" [5]. In case of the "on policy" algorithms make the assumption that the policy they are presently employing will also be utilized in the future when doing actions [5]. In order to engage in or associating with the many activities which the present sort of plan is recommending, they assess how fantastic a move or a condition is based mostly on reward obtained and the predicted gain in the next visible circumstance [5].

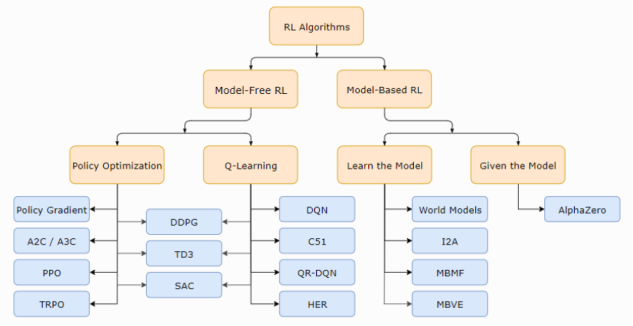

Fig 3.

Modern reinforcement learning algorithms are organised into a non-exhaustive taxonomy using OpenAI "an AI research and deployment company"[5].

This theory limits on-policy methods to using only learnings only with current policy [5] as projections are significantly dependent upon that guideline which is used to guide the actions in the future. Off-policy algorithms don't rely on this. Depending on lessons learned from past policies, non-policy approaches can help a policy. These strategies frequently draw lessons from data gathered in relation to both the present policy and a previous iteration of the current policy. In other words, as one learns, encounters are stored in memory, and as one trains, both new and stored encounters are used [5].

The Table 1 lists a few well-known or widely used model-free reinforcement learning algorithms along with some of their characteristics [5]. The types of state and action spaces that RL algorithms work well in make up their third distinguishing feature.

RL Algorithm	Policy	Action Space	State Space	Operator
Q-learning	Off-policy	Discrete	Discrete	Q-value
SARSA	On-policy	Discrete	Discrete	Q-value
SARSA - Lambda	On-policy	Discrete	Discrete	Q-value
DQN	Off-policy	Discrete	Continuous	Q-value
DDPG	Off-policy	Continuous	Continuous	Q-value
TRPO	On-policy	Continuous	Continuous	Advantage
NAF	Off-policy	Continuous	Continuous	Advantage
A3C	On-policy	Continuous	Continuous	Advantage
PPO	On-policy	Continuous	Continuous	Advantage
SAC	Off-policy	Continuous	Continuous	Advantage
TD3	Off-policy	Continuous	Continuous	Q-value

2.1 Algorithms based on Policies

Policy-based or policy optimization algorithms, in contrary to value-based algorithms, skip the intermediary step of forecasting the Q values for every single of the potential action and instead try to anticipate the best action to perform. In order to do this, value-based learning is replaced by the process of learning an assumable value for the optimum function(s) for the policies rather than the function for the state and action value. A policy with changeable parameters is often adopted, and its parameters are subsequently updated either using based or free of the gradient optimization techniques to increase the expected return. The second approach has had some success, but the former approach has demonstrated its worth and is the approach of preference for the majority of RL techniques. This is, mostly caused by the greater sampling efficiency of gradient-based approaches. The capacity to operate in large dimensional and continuous action spaces distinguishes policy-based algorithms in applications from value-based ones. Algorithms that are based on the values become infeasible when the quantity of possible actions is limitless and soon get complicated as the number of accessible actions increases. Continuous action spaces were never an issue for policy-based algorithms since, at the lowest level, they aim to create a single optimum action rather than a collection of Q-values for each potential action. The superior convergence features of policy-based actions versus value-based algorithms are another benefit. Value-based algorithms can experience significant performance oscillations while interacting, since even randomly very few or tiny changes in the predicted values may result in a significant shift in the preferred response at each stage. The majority of approaches for policy optimization, however, are gradient-based and, as a result, automatically provide updates that are smoother. The capacity to learn probabilistic policies, or policies that generate probability distribution across all potential responses provided a state as opposed to a deterministic mapping of states to actions, is a third benefit of policy-based algorithms. A stochastic policy's actual performed action is taken from a selection of the distribution it provides.

This encourages exploration naturally, therefore it is not necessary to add an additional algorithm, like the epsilon greedy approach, on atop of such algorithm to encourage exploration. Additionally, the issue of so-called perceptual aliasing is eliminated using a stochastic approach.

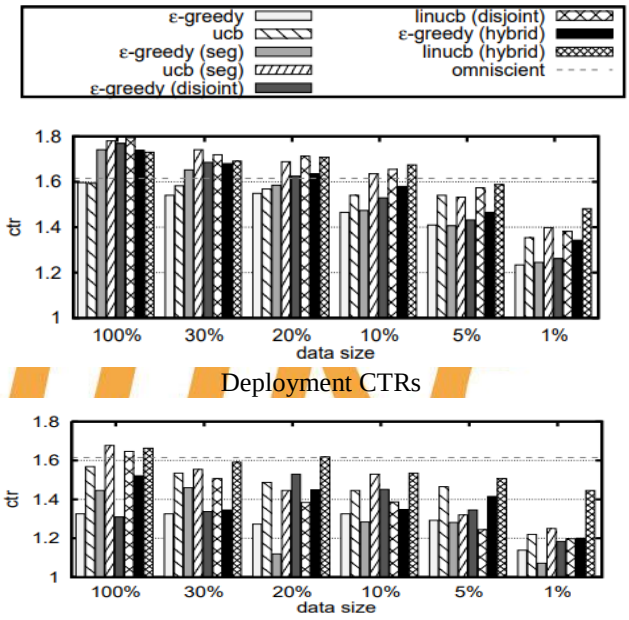

Fig 4. Assessment of CTRs with the data of variable sizes [8]

When two or more states have distinct optimum actions, yet from the viewpoint of an agent, the states appear to be identical due to incompletely visible environments, this problem is known as perceptual aliasing.

Due to the massive association between a state or even just the detection of one and a certain action, predictable methods are unable to decide the best course of action for every cognitively aliased state. Relatively, one may create a randomized rule projection that emphasizes all of the activities that are ideal in the various cognitively resampled states using policy based methodologies. Getting caught in local optima is a drawback with gradient based strategy refinement that applies to all gradient ascent and descending algorithms. This is especially true when contrasted to value-based learning. Furthermore, “policy based” background approaches frequently converge more slowly than valuation methods.

2.2 RECOMMENDING SYSTEMS USING RL

Recommendation systems' primary objective is to determine via customers' replies what they like, then to provide

predictions that correspond to their choices. This section examines how RL is changed to accommodate a number of significant functionalities in the recommendations.

2.2.1 Exploration and Exploitation Dilemma

The exploitation-exploration conundrum is a problem with traditional recommender systems. Exploration involves recommending items at random in order to gather more user feedback, whereas exploitation involves recommending items that are expected to best fit users' interests. The spatial bandit models an agent that attempts to strike a balance between the competing tasks of exploitation and research in maximizing the long term reward over a set period.

In a bandit situation, the conventional tactics to strike a balance between exploitation and exploration are "ε-greedy" [9], EXP3 [10], and UCB1 [10]. In order to boost the total amount of customer hits, a retraining method based LinUCB is presented to pick content systematically for individual users depending upon that background knowledge of the individuals and pieces. There in current feeds situation, the discovery challenge of tailored recommender system is described as a contextually bandit challenge [8].

2.2.2 Chronological Dynamics

In fixed situations where values are seen as permanent, a number of recommendation system, including clustering, information, and accruing, have been widely investigated. Due to the varying tastes of users, this assumption is frequently false, and as a result, the distribution of incentives frequently changes over time. We often give a variable reward function in the bandit situation to assess the dynamic character of the surroundings. The job of reward translation For example, inside a particle study model on the a nonlinear dynamic slide to highlight the reimbursement of a prize map task in the multiple armed bandit challenge [11], nonlinearity is understood as a gathering of probability models of particles, and indeed the feature set is defined by the deep shaped particles which have been flexibly selected. In order to track changes in the surroundings that affect earning confidence and to update the appropriate weapon selection strategy, the contextual criminal algorithm was developed [12].

2.2.3 Maintaining User Engagement

The study of a user's desired reaction to the suggestions made by the classification techniques is known as user growth [16]. Both immediate and long-term responses can be used to gauge user engagement [17]. [18] frames the problem of long-term user participation as a difficulty of successive decision-making. The agent must determine the risk of missing the customer to every repeat relying on the customer's variable reaction to prior suggestions. On the basis of the customer's transient behavior to earlier suggestions, the agent must determine the chance of losing a customer in each iteration. In realistic recommendation sessions, users will sequentially encounter a number of circumstances, such as the entrance sites and the commodity detailed page, but every case has a different suggestion strategy. [11] proposes a system based on multi-agent reinforcement learning that may simultaneously optimize a number of recommendation algorithms and capture the linear connection between distinct scenarios. To be more precise, a “model based” RL Technique is presented just to decrease amount of training data needed and carry out more precise strategy updates.

2.3 REINFORCEMENT LEARNING FOR ONLINE ADVERTISING

2.3.1 Guaranteed Delivery

Advertisements with a common topic or idea are bundled together into campaigns in guaranteed delivery, and the predetermined number of clicks or impressions is charged per campaign [12]. The most widely used Assured Deliveries methods are created using disconnected improvement approaches and then upgraded for use in an online world. Determining the ideal distribution of impressions, nevertheless, may be challenging, especially in applications where the surroundings is ambiguous. In [18], it is advised to use a multi-agent reinforcement learning technique to create cooperative rules that would enable the publisher to achieve its goal to the fullest in an unstable environment. The issue of allocating impressions was envisioned as an auction in which each contract might make a hypothetical offer for each impression. They were able to acquire the best impression allocation technique and solve the optimal contract bidding functions using this method.

2.3.2 Real Time Bids

Using RTB, a marketer may quickly make an offer to every unique view. Ad selection challenges are commonly modelled using multi-armed bandit problems with samples from each arm, quick feedback, and stationary rewards [20]; [18]; [21]. The payment functions of an MAB may vary, but it is expected that any modifications will occur gradually. On the other hand, as part of an advertising campaign, display ads are regularly introduced while others are removed. The algorithm's goal in this case is to tug the bandit's arms while staying within a certain cost in order to optimize the long term returns. Pulling the bandit's arms will result in unpredictable costs and benefits. This structure can more closely resemble Internet advertising than older works where tugging an arm was either free or had a set cost. The choice of the amount to bid is viewed in the MAB setting as a stable optimizer with two alternatives: weighing the value of each impression separately or setting a quote with each category of ad flow. Nonetheless, it would be continuous bidding for the duration of the budget for a particular advertising. As a result, studies have also been conducted on the MDP arrangement [22]. Raising the bandit's arms in just this situation will yield a range of benefits. Algorithms are utilized to predict the measured value in order to address the scalability challenge of a big bid frequency or a constrained marketing cost. Reward Net is designed to produce incentives in order to avoid reward design traps. The issue is characterized as a control challenge. A clustering approach and multi-agent bidding model that considers the bids of the other advertisers in the system are developed to handle the problem of a large number of advertisers [23].

3. CONCLUSION AND DIRECTIONS FOR THE FUTURE

In this article, we discuss knowledge acquisition from the perspective of RL. The introduction to RL based knowledge

collection querying strategies is what we commence with. Next, we examine the popular and recent algorithms for search, recommendations, and advertising, sample information seeking methods. We next go over some intriguing reinforcement learning research topics that may open up new region for information seeking research. The majority of current research, according to [13], train policies for only one situation while ignoring the preferences and actions of users in other circumstances. A poor strategy will arise from this, necessitating the use of collaborative RL frameworks that concurrently take into account of search, recommendation, and advertising scenarios. Secondly, for the recommendation systems, algorithms based upper confidence bound were more reliable. Developing a new algorithm requires significant technical work, which adds to the cost of testing it. If the algorithm is not yet developed, it may also have detrimental effects on user experience. Hence, it is essential to pre train and test new techniques and use an emulator of both the online space or an offsite evaluate the training based on previous logs preceding releasing them online. It is a growing need for further independent access RL environment for information acquisition in order to aid the RL and understanding groups and improve uniformity among offline and internet results.

REFERENCES

[1] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction Second edition, in progress."

[2] P. Varaiya and J. C. Walrand, "Multi-Armed Bandit Problems and Resource Sharing Systems," 1983.

[3] R. D. Smallwood and E. J. Sondik, "OPTIMAL CONTROL OF PARTIALLY OBSERVABLE MARKOV PROCESSES OVER A FINITE HORIZON.," Operations Research, vol. 21, no. 5, pp. 1071–1088, 1973, doi: 10.1287/opre.21.5.1071.

[4] E. J. Sondik, "OPTIMAL CONTROL OF PARTIALLY OBSERVABLE MARKOV PROCESSES OVER THE INFINITE HORIZON: DISCOUNTED COSTS.," Operations Research, vol. 26, no. 2, pp. 282–304, 1978, doi: 10.1287/opre.26.2.282.

[5] A. Mäkelä, "Deep reinforcement learning as a tool for search engine campaign budget optimization A dive into deep reinforcement learning and its application to optimizing budget allocation between search engine advertising campaigns Atte Mäkelä."

[6] Richard Bellman, dynamic programming. 2013.

[7] Q. Cui, F. S. Bai, B. Gao, and T. Y. Liu, "Global Optimization for Advertisement Selection in Sponsored Search," Journal of Computer Science and Technology, vol. 30, no. 2, pp. 295–310, Mar. 2015, doi: 10.1007/s11390-015-1523-4.

[8] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A Contextual-Bandit Approach to Personalized News Article Recommendation," Feb. 2010, doi: 10.1145/1772690.1772758.

[9] C. Watkins and R. Holloway, "Learning From Delayed Rewards A reversible MCMC model of sexual evolution: practical genetic algorithms with closed-form stationary distributions View project." [Online]. Available: https://www.researchgate.net/publication/33784417

[10] P. Auer, O. Cesa-bianchi, Y. Freund, R. E. Schapire, and S. J. Comput, "THE NONSTOCHASTIC MULTIARMED BANDIT PROBLEM *." [Online]. Available: http://www.siam.org/journals/sicomp/32-1/39837.html

[11] C. Zeng, Q. Wang, S. Mokhtari, and T. Li, "Online context-aware recommendation with time varying multi-armed bandit," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, vol. 13-17-August-2016, pp. 2025–2034. doi: 10.1145/2939672.2939878.

[12] Q. Wu, N. Iyer, and H. Wang, "Learning contextual bandits in a non-stationary environment," in 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, Jun. 2018, pp. 495–504. doi: 10.1145/3209978.3210051.

[13] J. Feng et al., "Learning to collaborate: Multi-scenario ranking via multi-agent reinforcement learning," in The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018, Apr. 2018, pp. 1939–1948. doi: 10.1145/3178876.3186165.

[14] X. Zhao, L. Xia, L. Zhang, Z. Ding, D. Yin, and J. Tang, "Deep reinforcement learning for page-wise recommendations," in RecSys 2018 - 12th ACM Conference on Recommender Systems, Sep. 2018, pp. 95–103. doi: 10.1145/3240323.3240374.

[15] X. Zhao, L. Xia, J. Tang, and D. Yin, "'Deep reinforcement learning for search, recommendation, and online advertising: a survey' by Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin with Martin Vesely as coordinator," ACM SIGWEB Newsletter, no. Spring, pp. 1–15, Jul. 2019, doi: 10.1145/3320496.3320500.

[16] M. Lalmas, H. O'Brien, and E. Yom-Tov, Measuring user engagement.

[17] S. Schopfer and T. Keller, "Long Term Recommender Benchmarking for Mobile Shopping List Applications using Markov Chains."

[18] D. Wu et al., "Budget constrained bidding by model-free reinforcement learning in display advertising," in International Conference on Information and Knowledge Management, Proceedings, Oct. 2018, pp. 1443–1452. doi: 10.1145/3269206.3271748.

[19] A. Hojjat, J. Turner, S. Cetintas, and J. Yang, "A unified framework for the scheduling of guaranteed targeted display advertising under reach and frequency requirements," Operations Research, vol. 65, no. 2, pp. 289–313, Mar. 2017, doi: 10.1287/opre.2016.1567.

[20] M. Gasparini, A. Nuara, F. Trovò, N. Gatti, and M. Restelli, "Targeting Optimization for Internet Advertising by Learning from Logged Bandit Feedback."

[21] 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, Apr.

2018, pp. 167–176. doi: 10.1145/3178876.3185994.
[22] H. Cai et al., "Real-time bidding by reinforcement learning in display advertising," in WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining, Feb. 2017, pp. 661–670. doi: 10.1145/3018661.3018702.
[23] J. Jin, C. Song, H. Li, K. Gai, J. Wang, and W. Zhang, "Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising," Feb. 2018, doi: 10.1145/3269206.3272021.