Reinforcement Learning in Online Advertising: Challenges, Prospects, and Trust

Jingwen Cai jingwenc@cs.umu.se

Johanna Björklund johanna@cs.umu.se

Department of Computing Science, Umeå University, 90187 Umeå, Sweden

Abstract

The central decision-making processes involved in online advertising are often supported by Reinforcement Learning (RL), which serves to optimise long-term accumulative rewards through interactions with evolving environments. While RL's potential in various real-world applications has been reviewed in extant survey works, the specific ways RL algorithms address online advertising challenges remain unchartered. Therefore, this paper reviews RL applications in this practice area, identifying core challenges and key issues including trust concerns. We categorize reviewed work based on problem domains and propose potential directions for future research. Our goal is to bridge the cross-disciplinary gap in this field, offering perspectives and guidance for researchers and practitioners.

Keywords: Reinforcement Learning; Online Advertising; Systematic Survey

1. Introduction

Online advertising is dependent on the programmatic ecosystem for the targeting, personalization, and delivery of ads. At the heart of the problem are complex decision-making processes involving stochastic and often nonstationary user preferences. These characteristics align with the strengths of Reinforcement Learning (RL), the goal of which is to maximize long-term cumulative rewards through repeated interaction with, e.g., users or trade desks. As a result, RL has been applied to many online advertising problems to optimize services with respect to different needs. Although existing surveys have reviewed RL's potential across a range of real-world applications, the specific ways in which RL algorithms address challenges in online advertising remain uncharted. In literature, online advertising systems are frequently addressed within the broader framework of recommender systems (RSs); however, they present unique challenges that demand specialized analysis. Moreover, feedback signals play a crucial role in RL, but the difficulty in directly linking user purchases to displayed ads increases the complexity of the problem. To address this gap we review influential literature on RL applications in online advertising. We follow the PRISMA[1] protocol and focus on studies that explicitly formalize their problems in online advertising contexts, or that use advertising simulations and datasets in their experiments. In doing so, we identify the core issues addressed by RL and present the primary algorithmic solution approaches.

1. Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) is a set of guidelines to enhance the transparency and completeness of reporting in systematic reviews and meta-analyses.

2. Related Work

Previous surveys can be broadly categorized into two groups: those emphasizing reinforcement learning and those focusing on advertising. The former seeks to provide comprehensive reviews of RL techniques for decision-making within information-seeking systems, with recommender systems often being a central focus. For instance, Afsar et al. (2022) classify RL methods in RSs into shallow and deep RL-based approaches and propose a framework unifying these problems based on four main components: state representation, policy optimization, reward formulation and environment building. In line with the above discussion, online advertising is covered by the authors under the broader context of recommender systems. Similarly, Zhao et al. (2019) review the application of deep RL across three types of information-seeking systems. Although online advertising is treated separately, the discussion remains limited to two primary marketing strategies, namely guaranteed delivery and real-time bidding, leaving room for further exploration of the topic.

In contrast, the second group of reviews centers specifically on online advertising, covering a wider range of technical aspects within this domain. Due to this shift in focus, the approaches reviewed are not limited to reinforcement learning but extend to machine learning in general. For example, in work by Choi and Lim (2020), 23 machine learning-based strategies for target advertising are investigated and categorized into user-centric and content-centric approaches. However, given the broad and loosely defined scope of machine learning and target advertising, the limited selection of papers fails to deliver a comprehensive systematic review. In comparison, there are other reviews that narrow their scope to specific issues in online advertising. For instance, Gharibshah and Zhu (2021) provide a comprehensive literature review regarding user response prediction in online advertising and Yang and Zhai (2022) offer a classification of state-of-the-art models for addressing click-through rate (CTR) prediction problems in advertising. However, the scope of these reviews differs from that pursued here, which centers on RL applications in online advertising.

3. Method

In our search for relevant literature, we used the databases provided by six well-established publishers in computer science: Elsevier, ACM, IEEE, Springer, Tayler & Francis, INFORMS, Sage, NeurIPS and Emerald (see Table 1, Line 1–6). We then ran a complementary search using three scholarly databases (Web of Science, Google Scholar, and Semantic Scholar) to find matches by provided by smaller publishers. Our search phrase was `‘‘reinforcement learning’’ AND ‘‘online advertising’’`.

After excluding webpages, workshop introductions and summaries, books, book reviews, tutorials, technical reports, keynotes, extended abstracts, opinion papers, and duplicates, 425 research papers remained. These were then further filtered for relevance based on their titles, abstracts, and introductions, forming an initial literature pool. We subsequently examined the references cited within these papers and, when relevant, added them to the pool. This process was repeated iteratively until the pool stabilized, ultimately comprising 163 ranked articles. In reviewing these works, we focused on the problems, solutions, contextual features, reward functions, benchmarks, datasets, and metrics discussed. Finally, we synthesized our findings, and this paper presents a summary with respect to key challenges, prospects, and issues of trust.

Table 1: The publisher and scholarly databases used in the literature search. For illustration, we include the number of articles contributed by each publisher for the search phrase ‘‘reinforcement learning’’ AND ‘‘online advertising’’.

Database	Link	Search results
Elsevier	https://www.sciencedirect.com/	82
ACM Digital Library	https://dl.acm.org/	246
IEEE Xplore	https://ieeexplore.ieee.org	26
SpringerLink	https://link.springer.com/	82
Taylor & Francis Online	https://www.tandfonline.com/	13
INFORMS	https://pubsonline.informs.org/	32
Sage Journals	https://journals.sagepub.com/	6
NeurIPS	https://neurips.cc/	75
Emerald	https://www.emerald.com/insight/	5
Web of Science	https://www.webofscience.com	
Google Scholar	https://scholar.google.com/	
Semantic Scholar	https://www.semanticscholar.org/	
Total		546

4. Central Challenges in RL for Online Advertising

In this section, we categorized the reviewed studies based on the challenges addressed, together with the RL approaches used to tackle them.

4.1. Optimization of Recommendations

Consistent with recommender systems more broadly, algorithms for online advertising are designed to pursue promotional objectives by shaping users' preferences and consumption behavior through recommended items. Accordingly, a significant body of the literature is devoted to the challenges associated with the optimization of such recommendation processes. The problem is often formalized as a bandit problem, aiming to maximize reward payoffs. For instance, Li et al. (2010a) assume a linear relationship between observed environment features and the expected payoff of a recommended ad. The authors use the Upper Confidence Bound approach to select the ad that yields the highest expected reward. Building on this, Zeng et al. (2016) incorporate time-varying features for recommendation optimization. In the study by Pei et al. (2019), the authors propose a value-aware RL algorithm to directly optimize profits, where the features used for describing each item are economic metrics instead of precision metrics.

Reward functions are central to reinforcement learning. In online advertising, user responses are the most commonly employed form of reward. Accurately modeling these responses, however, is highly challenging, often resulting in less informative reward signals and suboptimal recommendation performance. To overcome this problem, Chen et al. (2021) propose a generative inverse RL approach that learns directly from observed user behaviors without defining explicit reward functions. It assumes the existence of an expert model with sufficient knowledge of the environment to consistently make optimal recommendations. The RL model's goal then shifts to approximate the expert's policy, reframing

recommendation optimization as an automatic policy learning problem. This approach improves adaptability to complex real-world problems and enhances learning efficiency with the guidance of the assumed expert. However, careful consideration is necessary for formalizing the expert model, as the performance of the expert policy directly influences the quality of the final recommendations.

4.2. Optimization of Bids

In online advertising, an impression refers to an opportunity to display an ad to a user. Real-Time Bidding (RTB) allows advertisers to compete for these impressions programmatically through automated auctions. To maximize profits in this competitive environment, advertisers must optimize their bidding strategies: deciding when to bid, how much to bid, and which ads to promote.

Traditional bidding systems are often static, but the dynamic nature of online advertising requires adaptive approaches. Consequently, several recent works attempt to derive an optimal bidding strategy through RL. For instance, Cai et al. (2017) formulate the bidding problem as a sequential Markov Decision Process and uses dynamic programming to learn optimal bidding strategies, considering remaining bidding times and budgets. However, model-based RL methods like this can be inefficient and computationally expensive due to the need to store state transitions. To address this, Wu et al. (2018) further propose a model-free RL approach to improve efficiency in learning the optimal bidding strategy. In this approach, instead of deriving specific discrete bids, a bidding scaling parameter is learned by a deep Q-network to dynamically generate continuous optimal bids.

Most bidding optimization solutions are discussed under budget constraints, and to generalize these solutions for broader real-world applications, He et al. (2021) develop a unified framework using the Deep Deterministic Policy Gradient (DDPG) method to optimize bidding strategies for various advertiser demands. In the study by Yang et al. (2025), the authors propose a Proximal Policy Optimization (PPO)-based framework for budget allocation in a multi-channel advertising environment. The framework incorporates a multi-channel cooperative distribution model that integrates user behavioral coefficients with the marginal contributions of different channels to reduce redundancy and improve allocation efficiency. Moreover, beyond bidding strategy optimizations, other works also explore RTB auction mechanisms. The predominant approach in RTB is Generalized Second-Price (GSP) which ranks candidate ads based on a fixed set of predefined metrics, including their bids, and the top-ranked ad wins the auction with a cost of the second-highest bid. Zhang et al. (2021) point out that traditional GSP mechanisms usually use static ranking functions that focus solely on a single metric. Therefore, the authors propose a new ranking method, using DDPG to enhance efficiency by accounting for multiple, potentially conflicting, metrics that influence ad actions.

4.3. Off-Policy Evaluation

Off-policy evaluation (OPE) is a common technique in RL-based decision-making. The idea is to evaluate candidate strategies using historical logged data, offering a cost-efficient alternative to online evaluation. In advertising, online A/B testing is a common method to assess new strategies, but it can be prohibitively expensive to test each new strategy this

way. In addition, new strategies are not guaranteed to be better, so testing them directly online may risk having an undesired influence on users, potentially harming advertisers' long-term profits such as brand reputation.

A challenge with off-policy evaluation is that user responses are only available for actions recorded in the logged data, and that data may be biased. To address this, Langford et al. (2008) discuss OPE in a contextual bandit setting. The authors show that policy evaluation is not tractable if it relies on current contexts and propose a value calculator for context-independent policies. However, in the real world, policies are often likely to be closely related to various contexts, so the practicality and validity of the method discussed in this paper remain unclear. Strehl et al. (2010) extend this work by introducing a lower-error-bounded policy value estimator that also incorporates context features. The estimator is reported to perform well on Yahoo! datasets, but its performance depends on many parameter choices.

User responses are also available in the bandit dataset created by Saito et al. (2021) which contains public data collected from a Japanese e-commerce platform. The authors also provide open-source Python software to implement various bandit policies, offline policy learning methods, and benchmarks for several policy value estimators. To improve offline RL policies, Li et al. (2024) propose a trajectory-wise framework that injects noise into policy parameters and uses a novel Robust Trajectory Weighting method to address insufficient exploration and inadequate exploitation in the offline training.

4.4. Other Challenges

In addition to the three main challenges discussed above, there are several other issues. For instance, Ma et al. (2018) investigate data poisoning attacks, where a potential attacker manipulates the training data to influence the performance of the resulting models, causing financial losses or social risks. Therefore, understanding these offline data poisoning attacks is beneficial for developing a robust and defendable model. The authors simulate such attacks using Upper Confidence Bound methods, demonstrating their effectiveness in the experiments with small and undetectable poisoning modifications.

On another front, Hill et al. (2017) study ad layout optimization problems, focusing on multivariate optimization of interactive layouts in large decision spaces. Rather than optimizing the individual components of an ad layout in isolation, they use a parametric Bayesian model to capture the interaction between any two layout components. The authors then formulate the optimization task as a combinatorial bandit problem, employing Thompson sampling to balance exploration and exploitation. To identify the optimal solution, they use a greedy hill-climbing method that navigates the decision space in real-time. Similarly, Wang et al. (2021) propose a three-phase framework for ad text generation. The first two phases involve pre-training a Transformer-based language model on unsupervised data and fine-tuning it with supervised data collected from expert advertisers. Then in the final phase, model-based RL methods are used to refine the model through interactions with users, so that the generated ad text becomes as natural and pleasing as possible. Beyond texts, ad image generation also falls within this broader category of ad layout optimization. Chen et al. (2025) incorporate Multimodal Large Language Models (MLLMs) for CTR-driven ad image generation. After pre-training the model, they introduce a novel

reward model that jointly considers multimodal features and user preferences, and develop a product-centric preference optimization strategy to further improve model efficiency.

5. Prospects and Trust

The study of reinforcement learning in decision-making can be traced back decades, but the emergence of real-time bidding has significantly accelerated its development. We are now seeing a greater focus on problems motivated by real-world scenarios, which are dynamic, vast, and intricate in nature. While this is scientifically interesting and encourages the development of sophisticated algorithms, it also raises concerns regarding efficiency, computational resources, and trustworthiness. Metrics such as error bounds, convergence speed, and user experiences become integral to evaluating performances (Li et al., 2010b; Pei et al., 2019; Chen et al., 2025). There is also a trend exploring more comprehensive and unified solutions that, e.g., consider weighted combinations of metrics (Lacerda, 2017; He et al., 2021). In this case, the cross-disciplinary gap between RL and advertising is shrinking, with economic considerations increasingly represented in the algorithms. Examples include multi-objective optimization, ad layout design, and multi-agent bidding strategies.

Despite these efforts, significant challenges still exist, particularly the pressing need for open-source datasets and libraries. Advertising data often involves sensitive user or brand information, leading to limited availability of open-source datasets, which are always encoded and incomplete. Consequently, many studies rely on private datasets (Zhao et al., 2018; Pei et al., 2019; Liao et al., 2022; Li et al., 2024), hindering reproducibility and comparative evaluation. For instance, research on balancing exploration and exploitation may have inconsistent performance comparisons, partly due to testing on different private datasets (Li et al., 2010a; Afsar et al., 2022).

Lastly, as the advertising industry adapts to data protection regulations, there is a growing need for transparent and trustworthy solutions. For instance, demographic bias in personalized advertising systems remains a concern, and fairer solutions such as those proposed by Timmaraju et al. (2023) are needed to foster a more equitable advertising ecosystem. Contextual advertising has emerged as an alternative targeting approach that does not rely on personal data, while ad allocation and layout design increasingly emphasize the quality of interaction. As the user experience becomes ever more critical, it is essential to consider what defines a good advertising algorithm from an interdisciplinary perspective.

6. Conclusion

In this paper, we review the applications of reinforcement learning in online advertising, identifying key problems and future research directions in the field. Core challenges such as RTB optimization demonstrate RL's ability to address dynamic advertising problems. As the interest in creating scalable, dynamic, and trustworthy online advertising systems increases, there is a pressing need for more public datasets and the development of interdisciplinary, user-centric, and transparent RL approaches. These advancements are essential for fostering more effective and ethical advertising strategies in the future.

References

M. Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Comput. Surv.*, 55(7), December 2022. ISSN 0360–0300. doi: 10.1145/3543846. URL https://doi.org/10.1145/3543846.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 661–670, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346757. doi: 10.1145/3018661.3018702. URL https://doi.org/10.1145/3018661.3018702.

Xiaocong Chen, Lina Yao, Aixin Sun, Xianzhi Wang, Xiwei Xu, and Liming Zhu. Generative inverse deep reinforcement learning for online recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 201–210, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482347. URL https://doi.org/10.1145/3459637.3482347.

Xingye Chen, Wei Feng, Zhenbang Du, Weizhen Wang, Yanyin Chen, Haohan Wang, Linkai Liu, Yaoyu Li, Jinyuan Zhao, Yu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, Jingping Shao, Yuanjie Shao, Xinge You, Changxin Gao, and Nong Sang. Ctr-driven advertising image generation with multimodal large language models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2262–2275, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714836. URL https://doi.org/10.1145/3696410.3714836.

Jin-A Choi and Kiho Lim. Identifying machine learning techniques for classification of target advertising. *ICT Express*, 6(3):175–180, 2020. ISSN 2405-9595. doi: https://doi.org/10.1016/j.icte.2020.04.012. URL https://www.sciencedirect.com/science/article/pii/S2405959520301090.

Zhabiz Gharibshah and Xingquan Zhu. User response prediction in online advertising. *ACM Comput. Surv.*, 54(3), May 2021. ISSN 0360–0300. doi: 10.1145/3446662. URL https://doi.org/10.1145/3446662.

Yue He, Xiujun Chen, Di Wu, Junwei Pan, Qing Tan, Chuan Yu, Jian Xu, and Xiaoqiang Zhu. A unified solution to constrained bidding in online display advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 2993–3001, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467199. URL https://doi.org/10.1145/3447548.3467199.

Daniel N. Hill, Houssam Nassif, Yi Liu, Anand Iyer, and S.V.N. Vishwanathan. An efficient bandit algorithm for realtime multivariate optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1813–1821, New York, NY, USA, 2017. Association for Computing Machinery.

ISBN 9781450348874. doi: 10.1145/3097983.3098184. URL https://doi.org/10.1145/3097983.3098184.

Anisio Lacerda. Multi-objective ranked bandits for recommender systems. *Neurocomput.*, 246(C):12–24, jul 2017. ISSN 0925–2312. doi: 10.1016/j.neucom.2016.12.076. URL https://doi.org/10.1016/j.neucom.2016.12.076.

John Langford, Alexander Strehl, and Jennifer Wortman. Exploration scavenging. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 528–535, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390223. URL https://doi.org/10.1145/1390156.1390223.

Haoming Li, Yusen Huo, Shuai Dou, Zhenzhe Zheng, Zhilin Zhang, Chuan Yu, Jian Xu, and Fan Wu. Trajectory-wise iterative reinforcement learning framework for auto-bidding. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4193–4203, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645534. URL https://doi.org/10.1145/3589334.3645534.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670, New York, NY, USA, 2010a. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, and Rong Jin. Exploitation and exploration in a performance based contextual advertising system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 27–36, New York, NY, USA, 2010b. Association for Computing Machinery. ISBN 9781450300551. doi: 10.1145/1835804.1835811. URL https://doi.org/10.1145/1835804.1835811.

Guogang Liao, Ze Wang, Xiaoxu Wu, Xiaowen Shi, Chuheng Zhang, Yongkang Wang, Xingxing Wang, and Dong Wang. Cross dqn: Cross deep q network for ads allocation in feed. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 401–409, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512109. URL https://doi.org/10.1145/3485447.3512109.

Yuzhe Ma, Kwang-Sung Jun, Lihong Li, and Xiaojin Zhu. Data poisoning attacks in contextual bandits. In *Decision and Game Theory for Security: 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29–31, 2018, Proceedings*, pages 186–204, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01553-4. doi: 10.1007/978-3-030-01554-1_11. URL https://doi.org/10.1007/978-3-030-01554-1_11.

Changhua Pei, Xinru Yang, Qing Cui, Xiao Lin, Fei Sun, Peng Jiang, Wenwu Ou, and Yongfeng Zhang. Value-aware recommendation based on reinforcement profit maximization. In *The World Wide Web Conference*, WWW '19, pages 3123–3129, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313404. URL https://doi.org/10.1145/3308558.3313404.

Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/33e75ff09dd601bbe69f351039152189-Paper-round2.pdf.

Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pages 2217–2225, Red Hook, NY, USA, 2010. Curran Associates Inc.

Aditya Srinivas Timmaraju, Mehdi Mashayekhi, Mingliang Chen, Qi Zeng, Quintin Fettes, Wesley Cheung, Yihan Xiao, Manojkumar Rangasamy Kannadasan, Pushkar Tripathi, Sean Gahagan, Miranda Bogen, and Rob Roudani. Towards fairness in personalized ads using impression variance aware reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4937–4947, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599916. URL https://doi.org/10.1145/3580305.3599916.

Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. Reinforcing pretrained models for generating attractive text advertisements. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 3697–3707, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467105. URL https://doi.org/10.1145/3447548.3467105.

Di Wu, Xiujun Chen, Xun Yang, Hao Wang, Qing Tan, Xiaoxun Zhang, Jian Xu, and Kun Gai. Budget constrained bidding by model-free reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 1443–1451, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271748. URL https://doi.org/10.1145/3269206.3271748.

Mengfei Yang, Qiong Cao, Lingyun Tong, and Jiawen Shi. Reinforcement learning-based optimization strategy for online advertising budget allocation. In *2025 4th International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID)*, pages 115–118, 2025. doi: 10.1109/ICAID65275.2025.11034517. URL https://doi.org/10.1109/ICAID65275.2025.11034517.

Yanwu Yang and Panyu Zhai. Click-through rate prediction in online advertising: A literature review. *Inf. Process. Manage.*, 59(2), March 2022. ISSN 0306–4573. doi: 10.1016/j.ipm.2021.102853. URL https://doi.org/10.1016/j.ipm.2021.102853.

Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16,

pages 2025–2034, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939878. URL https://doi.org/10.1145/2939672.2939878.

Zhilin Zhang, Xiangyu Liu, Zhenzhe Zheng, Chenrui Zhang, Miao Xu, Junwei Pan, Chuan Yu, Fan Wu, Jian Xu, and Kun Gai. Optimizing multiple performance metrics with deep gsp auctions for e-commerce advertising. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 993–1001, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441771. URL https://doi.org/10.1145/3437963.3441771.

Jun Zhao, Guang Qiu, Ziyu Guan, Wei Zhao, and Xiaofei He. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1021–1030, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219918. URL https://doi.org/10.1145/3219819.3219918.

Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei 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. *SIGWEB Newsl.*, 2019 (Spring), July 2019. ISSN 1931–1745. doi: 10.1145/3320496.3320500. URL https://doi.org/10.1145/3320496.3320500.