Sequential Advertising Agent with Interpretable User Hidden Intents

Extended Abstract

Zhaoqing Peng
Alibaba Group

Junqi Jin
Alibaba Group

Lan Luo
University of Southern California

Yaodong Yang, Rui Luo
University College London

Jun Wang
University College London

Weinan Zhang
Shanghai Jiao Tong University

Miao Xu, Chuan Yu
Alibaba Group

Tiejian Luo
Univ. of Chinese Academy of Sciences

Han Li, Jian Xu, Kun Gai
Alibaba Group

ABSTRACT

Online advertising campaigns are typically launched for a customer across multiple touch points (scenarios) before the conversion of his final purchase. To maximize the advertisers' revenue, it requires the platform to develop its advertising strategy based on the consumers' behavioral trajectories in the previous scenarios. Meanwhile, it is also critical to maintain the interpretability of the models on the conversion rate; however, modern reinforcement learning based solutions fail to do so due to their black-box modeling on the consumer intents. In this paper, we model consumer's purchase intention as a latent variable, and formulate the advertising problem as a partially observed Markov Decision Process (POMDP). We incorporate the expectation maximization (EM) algorithms for solving the optimal POMDP. Our extensive experiments based on large-scale real-world data demonstrate that our method provides superior performance over several baselines. Apart from the improved advertising performance, our method is able to offer interpretation on the attribution of the conversion.

ACM Reference Format:
Zhaoqing Peng, Junqi Jin, Lan Luo, Yaodong Yang, Rui Luo, Jun Wang, Weinan Zhang, Miao Xu, Chuan Yu, Tiejian Luo, and Han Li, Jian Xu, Kun Gai. 2020. Sequential Advertising Agent with Interpretable User Hidden Intents. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

Online advertising is a marketing paradigm that leverages the Internet, either through PC or mobile phones, to target the audience, with the goal to motivate their conversion of purchases [13]. Typically, advertising campaigns interact with the consumers multiple times until the conversion of purchase across different scenarios. To maximize the total advertising income, it is of merchant's great interest to jointly optimize the sequential advertising strategies across different scenarios for each individual consumer.

As Fig.1 depicted, we show an example of the sequential advertising strategy on an advertisement (hereafter **ad**), launched by an advertiser X. In the beginning, the consumer saw X's display ad at time t_1 in scenario No.1, and then he clicked on the same ad for details at time t_2. When the consumer switched to the scenario No.2 and saw X's ad again, he finally made a purchase at time t_3 to t_4. How do we attribute the final conversion to the displayed ads? Does the advertising effects in scenario No.2 contribute to that in scenario No.1? In this paper, we try to figure out these questions so that the future advertising strategies with similar user trajectories could be improved.

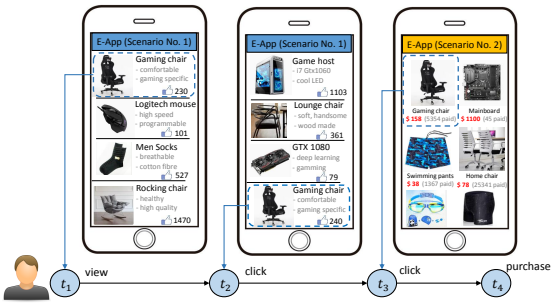

Figure 1: Consumer trajectories on certain advertising campaign across different scenarios.

So far, the solutions to the multi-scenario advertising have been focused on the attribution-based methods and the optimization-based methods [5, 13]. Attribution-based methods pay attention to the analysis of how to assign the credits to the previous ad displays before the conversion, while without optimizing their advertising strategies. Optimization methods usually formulate the problem with reinforcement learning (RL) [1–4, 6]. But they do not explicitly model the interacting environment, i.e., users' purchase intentions; this may incur difficulties in interpreting the attribution of the conversions.

In this paper, we consider not only the optimization of the advertising strategy but also the attribution through modeling the consumer conversion intentions. Obviously, this intention cannot be directly observed, but we consider it as a latent variable that can be inferred from large historical observation data. As such, we formulate the problem as a Partially Observable Markov Decision Process (POMDP). The significant advantage over the previous methods [2–4, 6] is that our POMDP-based approach offers more

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

interpretability by inferring how probable a user will be in each hidden state and which state the user will transit to; this helps us analyze the attribution of certain ads exposure, and improves the future advertising policies.

Optimizing advertising strategy by POMDP is, however, difficult since the POMDP model in e-commerce platform is not available in advance. To tackle this issue, we derive an EM-based method to estimate the parameters of POMDPs by learning from large-scale real online data. Given the learned parameters, we now can infer the probability distribution of the latent states (defined as beliefs), and then optimize the advertising policy based on those beliefs. However, exact methods for the policy learning are intractable [10], so we adopt a variant of Smooth Partially Observable Value Approximation (SPOVA) [11] to approximate the belief value function. In this manner, we can implement SPOVA with neural networks to improve the generalization and the learning efficiency of the advertising decisions. Based on these, we propose a POMDP-based approach, namely Deep Intents Sequential Advertising (DISA), which extends and adapts POMDP to a real-world advertising problem. While the majority of the previous studies on POMDPs are mostly theoretical in nature, [7, 8, 12], our study is developed in the context of a realistic industrial setting. The results of simulations and online experiments demonstrate our method's superiority over several baselines.

2 PROBLEM DEFINITION

Formally, given a sequence of requests from a consumer, our problem is defined as a sequential decision process to determine the appropriate ad items to maximize the advertisers' revenue with fewer budgets. At each time-step t, the agent infers a probability distribution b_t (belief) over all user hidden states and decides on the optimal action a_t based on b_t. The current belief b_t is produced with the previous b_{t-1} and a_{t-1} using the Bayes rule:

$$b_t(s') = \rho O(s', a_{t-1}, o_t) \sum_{s \in \mathcal{S}} T(s, s', a_{t-1}) b_{t-1}(s) \tag{1}$$

where $T(s, s', a)$ is the transition function, $O(s', a, o)$ is the observation function, ρ is the normalized factor, and $b(s)$ represents the probability that a consumer hidden state is in state s. After estimating b_t, the agent has to learn the mappings from beliefs to actions, denoted by a policy $a_t = \pi(b_t)$, and the Bellman equation for POMDP [10] is defined as:

$$V^*(b_t) = \max_{a_t} [r_t + \gamma \sum_{o \in O} T(o_t | a_t, b_t) V^*(b_{t+1})] \tag{2}$$

where $V^*(b_t)$ is the belief value function with an optimal policy π^*. The reward is defined as the advertiser's revenue subtracting the budget cost, given by $r_{t,i} = \lambda price_i y_t - bid_i x_t$ where λ is a positive hyper-parameter to adjust the weight between earned revenue and cost. The objective of learning is to find optimal policies to maximize the expected return of each ad item.

3 EXPERIMENT

Our experiments are conducted over the dataset collected from Taobao display ad system. The following algorithms are compared with our method (DISA) with the same settings: 1) Manual bid [6]: it is the real bid strategy conducted according to humans' experience,

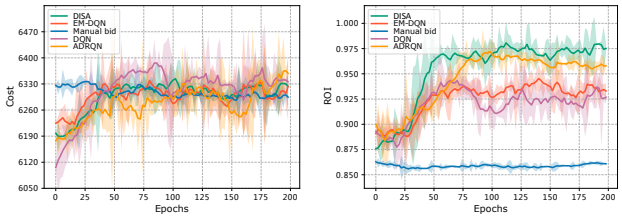

Figure 2: The learning curves of cost and ROI.

2) DQN [9]: it is a model-free RL algorithm, 3) ADRQN [14]: it is a model-free POMDP where the latent state is implicitly captured and modeled by the LSTM, 4) EM-DQN: it is a variant of DQN where its input is the beliefs of DISA rather than observations.

Each method is evaluated by the ROI indicator (revenue/cost) and the average rewards (the advertisers' profits). Fig.2 shows that DISA outperforms all the others in ROI while achieving almost the same cost as other baselines. These results indicate the superiority of DISA as it not only helps advertisers earn more income per budget cost but also improve profits. Compared with DQN, a higher ROI of EM-DQN shows the benefits of inferring beliefs over the behavior-action mappings (black-box) in model-free fashion. Furthermore, DISA also demonstrate its advantage of the belief value approximation in SPOVA over the general neural network (pure belief-action mappings) in EM-DQN.

Interpretability. Essentially, the EM in our DISA learns a mapping from high-dimensional historical observations to a compressed belief state, and this mapping is reflected in the learned parameters. By analyzing the $T(s'|s_i, a)$, $O(o|s_i, a)$ and $b_0(s_i)$ parameters, we can know how each state connects with different observations, so as to further interpret the property of each state s_i. According to our experiments, we explain the characteristics of each state: 1) state s_3 is an awareness state since the users under s_3 are observed to have little advertising exposure and clicks, 2) state s_2 is an interest state because we observe a large number of user browsing and click behaviors in this state, 3) compared with state s_2, users in s_1 start to actively search for their interested items across different scenarios, thus we label s_1 as an active state.

Based on the interpretable state, we cast light on the learned strategies by a few case studies, through which we find that our DISA successfully learns to select the ad item with a potentially higher reward to win the bidding. In addition, we also get important insights from DISA that: 1) users under an active state are more likely to convert than an interest state as well as an awareness state, and 2) for a category of items, an optimal advertising strategy is to guide a user to the active state while he/she switches to a certain scenario.

4 CONCLUSIONS

In this paper, we proposed a POMDP-based solution to the problem of multi-scenario sequential advertising. Our method incorporates the consumers' latent intentions in optimizing the advertising strategy. We evaluated our models in the context of realistic production setting; the results have validated the superiority of the proposed algorithm on performance against several RL baselines. We found that the model can correctly infer users' latent intentions.

REFERENCES

[1] Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. 2017. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 661–670.

[2] Shi-Yong Chen, Yang Yu, Qing Da, Jun Tan, Hai-Kuan Huang, and Hai-Hong Tang. 2018. Stabilizing reinforcement learning in dynamic environment with application to online recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1187–1196.

[3] Jun Feng, Heng Li, Minlie Huang, Shichen Liu, Wenwu Ou, Zhirong Wang, and Xiaoyan Zhu. 2018. Learning to Collaborate: Multi-Scenario Ranking via Multi-Agent Reinforcement Learning. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1939–1948.

[4] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement Learning to Rank in E-Commerce Search Engine: Formalization, Analysis, and Application. *arXiv preprint arXiv:1803.00710* (2018).

[5] Wendi Ji and Xiaoling Wang. 2017. Additional Multi-Touch Attribution for Online Advertising.. In *AAAI*. 1360–1366.

[6] Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. 2018. Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising. *arXiv preprint arXiv:1802.09756* (2018).

[7] M Mahmud. 2010. Constructing states for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 727–734.

[8] Andrew Kachites McCallum and Dana Ballard. 1996. *Reinforcement learning with selective perception and hidden state*. Ph.D. Dissertation. University of Rochester. Dept. of Computer Science.

[9] et al. Mnih, Volodymyr. 2015. Human-level control through deep reinforcement learning. *Nature 518, no. 7540 (2015): 529* (2015).

[10] Kevin P Murphy. 2000. A survey of POMDP solution techniques. *environment* 2 (2000), X3.

[11] Ronald Parr and Stuart Russell. 1995. Approximating optimal policies for partially observable stochastic domains. In *IJCAI*, Vol. 95. 1088–1094.

[12] Andres C Rodriguez, Ronald Parr, and Daphne Koller. 2000. Reinforcement learning using approximate belief states. In *Advances in Neural Information Processing Systems*. 1036–1042.

[13] Xuhui Shao and Lexin Li. 2011. Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 258–264.

[14] Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. 2018. On improving deep reinforcement learning for pomdps. *arXiv preprint arXiv:1804.06309*.