Decision Support Systems 56 (2013) 379–386

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Online learning for auction mechanism in bandit setting

Di He [a], Wei Chen [b], Liwei Wang [a,*], Tie-Yan Liu [b]

[a] Key Laboratory of Machine Perception, MOE, School of Electronics Engineering and Computer Science, Peking University, 100871 Beijing, PR China
[b] Microsoft Research Asia, Beijing, PR China

ARTICLE INFO

Article history:
Received 7 November 2012
Received in revised form 11 June 2013
Accepted 12 July 2013
Available online 26 July 2013

Keywords:
Armed bandit problem
Mechanism design
Online advertising

ABSTRACT

This paper is concerned with online learning of the optimal auction mechanism for sponsored search in a bandit setting. Previous works take the click-through rates of ads to be fixed and known to the search engine and use this information to design optimal auction mechanism. However, the assumption is not practical since ads can only receive clicks when they are shown to users. To tackle this problem, we propose to use online learning for auction mechanism design. To be specific, this task corresponds to a new type of bandit problem, which we call the armed bandit problem with shared information (AB-SI). In the AB-SI problem, the arm space (corresponding to the parameter space of the auction mechanism which can be discrete or continuous) is partitioned into a finite number of clusters (corresponding to the finite number of rankings of the ads), and the arms in the same cluster share the explored information (i.e., the click-through rates of the ads in the same ranked list) when any arm from the cluster is pulled. We propose two upper-confidence-bound algorithms called UCB-SI1 and UCB-SI2 to tackle this new problem in discrete-armed bandit and continuum-armed bandit setting respectively. We show that when the total number of arms is finite, the regret bound obtained by UCB-SI1 algorithm is tighter than the classical UCB1 algorithm. In the continuum-armed bandit setting, our proposed UCB-SI2 algorithm can handle a larger classes of reward function and achieve a regret bound of $O(T^{2/3}(d\ln T)^{1/3})$, where d is the pseudo dimension for the real-valued reward function class. Experimental results show that the proposed algorithms can significantly outperform several classical online learning methods on synthetic data.

1. Introduction

Nowadays, online advertising has become one of the most profitable business models for internet companies. Sponsored search, as a major type of online advertising, is a revenue powerhouse for search engines. Keyword auction is the central mechanism in sponsored search, which determines the ads to be present to the users (which we call ranking) and the per-click prices to charge the corresponding advertisers (which we call pricing).

The keyword auction mechanism works as follows. When a web user submits a query to search engine, the search engine not only delivers organic search results to him/her, but also shows real-time sponsored search results, i.e., advertisements (see Fig. 1). As a dominant industry practice, the search engine will charge an advertiser only when a user clicks on his/her ad. This is referred to as cost-per-click pricing rule (CPC). Generalized Second Price Auction (GSP) [1,2] is a widely used mechanism for CPC, in which the ads are ranked according to a function of the ad quality and its bid price, and the per-click price of a displayed ad equals the minimal bid price for the owner of the ad to maintain the current rank position. Different ways of computing quality score have been used in the literature, for example, Yahoo! once used a constant quality score in early 2000s [1], and Google uses the predicted click-through rate nowadays [2].

With different quality score functions, different ads will be shown to the users and advertisers will receive different charged prices if their ads are clicked. Thus the quality score function will highly affect the search engine's performance, e.g., revenue. In the literature, there are several works on revenue maximization for the search engine, from either machine learning or game theory perspective. In [3,4], the authors assume that the click-through rate and bidding price of each ad are known and fixed, and then propose a machine learning approach to find the revenue-optimal quality score function based on historical log data. In [5,6], the authors assume that the advertisers have full information, and the click through-rate of the ads are known to the search engines, then different quality score functions can be compared with respect to the worst-case revenue in symmetric Nash equilibrium or Bayesian Nash equilibrium (Fig. 2).

However, in most of the previous works, a key assumption is that the click-through rates of all ads are known to the search engine and never changed. This is seldom true in practice. In real applications, there are usually hundreds of advertisers bidding on one keyword, and only a small number of ads can be shown on the search result page and receive clicks. If one ad has never been shown to the users, the probability of the ad being clicked cannot be observed; Furthermore, even if one ad has been shown to users in history, the probability of it being clicked is difficult to estimate because the variance is large if the number of

* Corresponding author. Tel.: +86 10 6275 6657; fax: +86 10 6275 5569.
E-mail address: wanglw@cis.pku.edu.cn (L. Wang).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.07.004

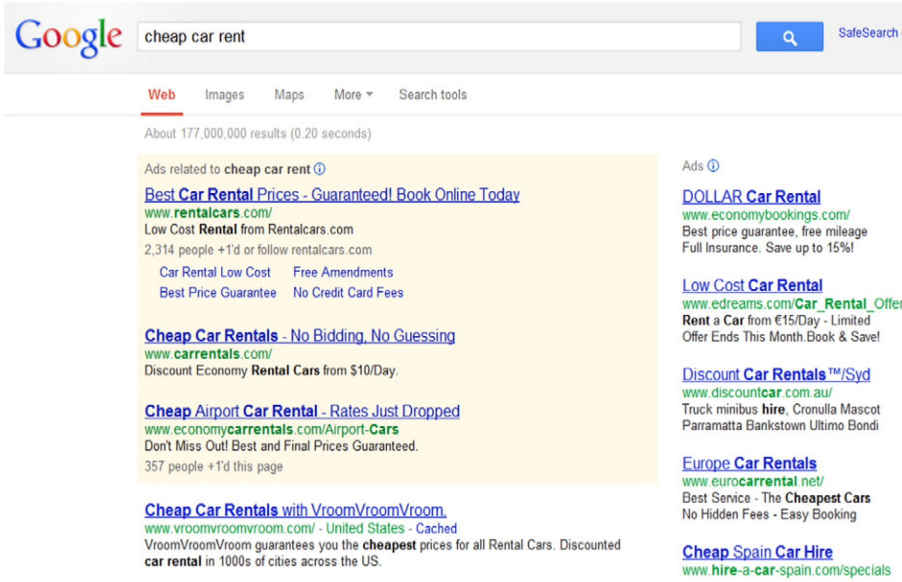

Fig. 1. The displayed ads for query “cheap car rent”.

observations is small. Such limitations make the pervious machine learning methods or game theoretic analysis for revenue maximization not practical.

To tackle this problem, we propose online exploring the users' click behaviors as well as using the explored information for auction mechanism design. In particular, we propose to user the bandit algorithms, where the quality score function in GSP mechanism corresponds to the arm, and the performance (e.g., search engine revenue) corresponds to the reward function.

Our setting has several new features. First, in our task, while the quality score function (arm) may come from a continuous function class, only the top-ranked ads will have impact on the revenue (since only these top-ranked ads will be shown to the users). This makes the reward function discontinue. Furthermore, the reward function is non-convex due to the complex per-click pricing scheme. This type of reward function has never been studied in literature of bandit problems. Second, the arms in our new bandit problem share the explored information. No matter how different the quality score functions are, as long as the ranked list of ads produced is the same, users will give the same kinds of feedback (because users can only see the ranked list of ads, but not the scores for these ads). Therefore, those arms that produce the same ranking result have high dependency on their rewards, and the explored information about user clicks can be shared among them.

Considering the aforementioned uniqueness, we propose a new type of armed bandit problem, called the armed bandit problem with shared information (AB-SI). Specifically, the AB-SI problem has a two-layer structure. In the first layer, there are $K(K < \infty)$ clusters, with arms putting into different clusters according to their dependencies. In the second layer, within a cluster, the number of arms may be finite or infinite, when exploring any arm in a given cluster, the obtained information can be shared with other arms in the cluster.

To handle the aforementioned problem, we propose a stepwise online–offline learning algorithm, which we call the UCB-SI algorithm in general (which can be regarded as a generalization of the standard UCB algorithm [7]). In the online learning phase of the algorithm, one of the clusters is selected according to the best empirical performance of the arms in the cluster as well as a confidence value, then the best

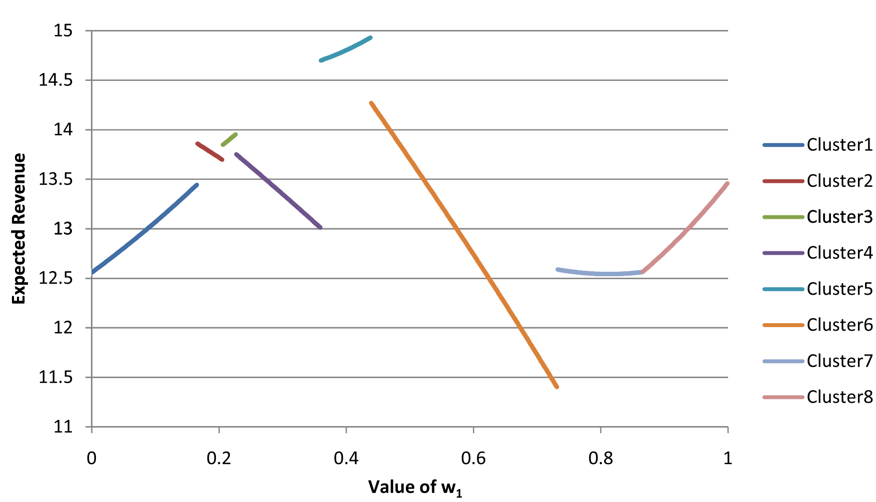

Fig. 2. The dependency between expected revenue and arms.