

Modeling User Click Behavior in Sponsored Search

Vibhanshu Abhishek, Peter S. Fader, Kartik Hosanagar
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
{vabhi, faderp, kartikh}@wharton.upenn.edu

Abstract

There has been significant recent interest in studying consumer behavior in sponsored search environments. Sponsored search is the fastest growing form of advertising on the Internet. A number of factors have contributed to this growth. The ads tend to be highly targeted and offer a higher return on investment for advertisers compared to other marketing methods. In addition the large audience it offers has led to a wide-spread adoption of search engine advertising.

When a user issues a query on the search engine, sponsored results are displayed alongside organic search results. The organic search results are links relevant to the query and are ranked in order of their relevance. The sponsored results are ads submitted by advertisers. The advertisers submit bids for keywords that are relevant to them, along with these ads. When a user enters a query, the search engine identifies the advertisers bidding on keywords closely related to the query and uses data on bids and ad quality/performance to rank order the ads that appear in the list of sponsored results. The most widely used pricing model is the *pay per click* model, in which the advertiser pays only when a user clicks on his ad. The advertiser's cost per click or *cpc* is determined using a generalized second price auction, i.e. whenever a user clicks on an ad in position, the advertiser pays an amount equal to the minimum bid needed to secure that position.

One of the attractive features of sponsored search is that it is a highly measurable form of advertising. Data on consumer click and purchase patterns have been used to study consumer behavior and advertiser strategies. Several researchers have built random utility models to study the effect of ad position, keyword length, presence or absence of brand name, etc. on the clickthrough rate (*ctr*) of the ad. Rutz and Bucklin [6] propose a hierarchical bayesian model to study the conversion performance of individual keywords. They show that the model adequately addresses the sparse data problem while accounting for keyword heterogeneity. Ghose and Yang [4] use hierarchical bayesian models to understand the relationship between different metrics such as *ctr*, conversion rates, bid prices and keyword ranks using the advertiser's aggregate data. They also show that the advertisers are not bidding optimally to maximize their profits. Recent work by Agarwal et. al [1] shows that although the *ctr* decreases with position, the *conversion rate* often increases and then decreases. They show that the topmost position is not necessarily the revenue maximizing position. These models have been estimated on aggregated data that catalogue advertiser's bid, average position and total impressions, clicks and cost on a daily basis for keywords in the advertisers sponsored search campaign.

The position of an ad varies across impressions during a day, but since these use aggregate data, their models assume the mean position during the day is the actual position. This aggregation of data can lead to potential biases in the estimation of parameters and ultimately affect the conclusions from these studies. The problem of aggregation bias has been addressed earlier in some detail but there is no definite answer. Kelejian [5] discusses why, under certain conditions, aggregation bias might occur and proposed a test for the existence of this bias. Allenby and Rossi [2] present an

analytical proof for the non existence of aggregation bias in nested logit models when the estimate is performed on store level scanner data.

In this paper we conclusively show that applying standard models (e.g., logistic regression) on such aggregated datasets can lead to biased estimation of the parameters of a random utility model. There is sufficient variability in the position of the ad during the day which is ignored in these models. In order to prove the existence of the bias, we first prove that the average position is less in convex order than the position for a particular impression. A simple position based model is used to show that if the data generating process was logit in nature then the logit function estimated on the aggregate dataset is always biased. Furthermore, the effect of position on *ctr* is underestimated.

Unfortunately, advertisers and researchers do not have access to the complete data-set and need to use the aggregate data as the search engines report only the aggregate level performance of keywords in the advertiser’s campaign. We propose a probabilistic model for consumer behavior that accounts for variation in position. Probabilistic models are extensively used in Marketing [3] and to our knowledge this paper is the first application of these techniques in the sponsored search literature. There can be several drivers for the variation in position. We model the variation that occurs due to the change in advertisers’ bids and the consumer queries. A continuously updating GMM estimator is used to estimate the parameters of the model. Preliminary results show that if our model is the data generating process then unbiased estimates of the parameters can be obtained from the aggregate data. Finally we evaluate the performance of our model on real world datasets and determine if our model is a better predictor of click through and conversion rates than logistic regression.

Sponsored search is increasingly becoming an important medium for reaching out to the consumers and hence the industry and researchers need to understand how to exploit this medium of advertising. In this work we show that currently used techniques might be inappropriate given the aggregate nature of the data and propose an alternative method that can be adequately used to study the drivers of sponsored search.

References

- [1] A. Agarwal, K. Hosanagar, and M. D. Smith. Location, location, location: An analysis of profitability and position in online advertising markets. *SSRN eLibrary*, 2008.
- [2] G. M. Allenby and P. E. Rossi. There is no aggregation bias: Why macro logit models work. *Journal of Business and Economic Statistics*, pages 1–14, 1991.
- [3] P. S. Fader and W. W. Moe. Dynamic conversion behavior at e-commerce sites. *Management Science*, pages 326–335, 2004.
- [4] A. Ghose and S. Yang. An empirical analysis of search engine advertising: Sponsored search and cross-selling in electronic markets. *SSRN eLibrary*, 2007.
- [5] H. H. Kelejian. Aggregated heterogeneous dependent data and the logit model: A suggested approach. *Economic Letters*, pages 243–248, 1995.
- [6] O. J. Rutz and R. E. Bucklin. A model of individual keyword performance in paid search advertising. *SSRN eLibrary*, 2007.