Chapter 1
Click-Through Rate Estimation for Rare Events in Online Advertising

Xuerui Wang
Yahoo! Labs, USA

Wei Li
Yahoo! Labs, USA

Ying Cui
Yahoo! Labs, USA

Ruofei Zhang
Yahoo! Labs, USA

Jianchang Mao
Yahoo! Labs, USA

ABSTRACT

In online advertising campaigns, to measure purchase propensity, click-through rate (CTR), defined as a ratio of number of clicks to number of impressions, is one of the most informative metrics used in business activities such as performance evaluation and budget planning. No matter what channel an ad goes through (display ads, sponsored search or contextual advertising), CTR estimation for rare events is essential but challenging, often incurring with huge variance, due to the sparsity in data. In this chapter, to alleviate this sparsity, we develop models and methods to smoothen CTR estimation by taking advantage of the natural data hierarchy or by clustering and data continuity in time to leverage information from data close to the events of interest. In a contextual advertising system running at Yahoo!, we demonstrate that our methods lead to significantly more accurate estimation of CTRs.

DOI: 10.4018/978-1-60960-189-8.ch001
INTRODUCTION

The Internet revolution has driven a transformation of how people experience information, media and advertising. Web advertising did not even exist twenty years ago, but nowadays it has become a vital component of the modern Internet, where advertisements are delivered from advertisers to users through different channels such as display ads, sponsored search and contextual advertising. Based on a recent study by Interactive Advertising Bureau and PricewaterhouseCoopers International, online advertising revenue in the first half of 2009 reached over 10.9 billion US dollars (IAB & PwC, 2009). Recent trends have also witnessed that larger and larger share of advertisers’ budgets are devoted to the online world, and online advertising spending growth has greatly outpaced some of the traditional advertising media, such as radio and magazine.

There are many pricing models for online advertising. For example, in a pay-per-impression campaign, the advertisers pay for the number of ad exposures. This is a very popular model in display ads. Another example is pay-per-action, where the advertisers pay only if the ad leads to a sale or similar transaction. Both models have their own limitations. The pay-per-impression model does not consider ad performance while the pay-per-action model often has difficulty in tracking user transactions. Therefore, the success and effectiveness of online advertising campaigns is prevalently measured by whether users click on ads. This pricing model is called pay-per-click, where the advertisers pay a certain amount for every click. As a result, the click-through rate (CTR), defined as a ratio of number of clicks to number of impressions of an ad for a query or on a page, is one of the most useful and informative metrics to measure user response in online advertising. The use of CTR in online advertising is everywhere. For many advertising systems, CTR is one important factor to determine what advertisements to be displayed and in what order.

In addition, advertisers often plan their budget based on historical CTRs or predicted CTRs.

In online advertising, a serving event refers to the showing of an ad in response to a user query, which can be a search query in sponsored search or a Web page in display and contextual advertising. The serving frequency varies for different queries and ads. Some popular ones account for a large fraction of Internet traffic while the overwhelming counterparts are extremely rare events, whose behaviors follow a typical power-law pattern. Even being infrequent, the revenue from the rare events generates a several-billion-dollar business, which makes estimating CTR for such rare events an extremely important and practical problem. Unfortunately, due to the sparseness, the CTR estimation for rare events is notoriously challenging and extremely unreliable, often with huge variance. There are two scenarios that are especially difficult to deal with. The first one is when we observe a click out of only few serving events. A direct estimation of the CTR is often much higher than the true CTR. Another case is when we observe no clicks at all out of thousands of serving events. This is very common in contextual advertising because many page/ad CTRs are on the magnitude of 0.1% or even lower. Again, a direct estimation is not desirable here. In order to address these issues, we need to take into account that rare events are not strictly independent to each other. Instead, the existing correlations or associations among events can be leveraged into the CTR estimation for rare events.

First, in many scenarios, attributes in event data are naturally organized in hierarchies, and/or can be clustered via data-driven methods. An illustrative example of a page to publisher hierarchy is shown in Figure 1. As we can see, the publisher “yahoo_food_specialday” (special day site for Yahoo! Food) contains many pages that are similar to each other, such as foods for birthday and anniversary. These pages differ in content but share some common themes in the big picture, and we would imagine CTRs on these