The Importance and the Complexity of Online Advertising

Timothy C. Owen and William Kahn
January, 2012
Travelers Insurance, Hartford, CT

The Situation:

A major source of funding for the internet is advertising. In the United States, in 2010, $26 billion was spent on online advertising. This represents about 15% of all the money spent in advertising including TV, newspapers, magazines, radio, and billboards. In some countries, such as the UK, more advertising money is spent on the internet than on TV. The price charged for most of this internet advertising is directly tied to performance. Many major companies, including Google and Yahoo, are primarily funded through advertising.

To calibrate one's sense of scale, $26B is larger than the sum of NASA's $19 billion and the National Science Foundation's $7B budgets. Online advertising is big business. But it is just now starting to receive serious analytic attention. The simultaneous optimization of total return for marketers, vendors, and users is still far from being figured out. And the data size, routinely 10TB per year even for modest sized marketers, remains a technical challenge.

We'll describe here one particular puzzle: the attribution problem. An internet user is exposed to a sequence of advertising. At some point the user may take an action—perhaps clicking on a banner ad, searching for the product, visiting the merchant's website, or making a purchase. For simplicity, let us for right now simply consider the action of interest to be visiting the merchant's web site.

The question, which is of simultaneous interest to the merchant, the vendor, and the user is , "What characteristics of the sequence of banner ads was influential in guiding the user to the website?" To a surprising degree there is alignment of interest between all three parties. The merchant wants to spend as little money as possible. The vendor that displays the merchant's ad on their website wants to give users the most pleasant experience possible. The user wants to be exposed to as little marketing "noise," and as much marketing "signal," as possible.

The Business Question:

What sequence of banner ads works best?

The Data:

The basic version of the data a merchant collects is a record for each user (ignoring for now the difference between users and computers/browsers). Every banner ad a merchant serves to a user is tracked. (For discussion right now, though not in industrial practice, we will ignore complexities such as computers that do not allow cookies or computers that delete cookies quickly). More advanced data structures, available to vendors, may include a richer set of banner exposures, such as ads for competitive or complementary products—another complexity we will ignore for now.

Typically, the characteristics of the ads and the characteristics of the users are parameterized. Parameterization is the process of defining how differences will be measured, and is fundamental to how a problem will be understood.

For the ads, features such as the total number of pixels, aspect ratio, colors, and content constitute 10 to 20 controllable features. The decision of how to parameterize an abstract idea like the ad content requires significant understanding of the product value and the user's interests. For example, an advertisement may be a "value play," e.g. advertising low price, or a "feature play," advertising performance. Many other ways of categorizing content are created by advanced merchants.

There is much insight needed to decide how best to parameterize the web site in which a banner ad is shown. Free or paid? News or social? Political? Fashion? What are the socio-demographics (age, sex, geography,…) of the website's visitors?

Features of the website's visitors considered include browser type, operating system, and geographic region (based on IP address).

Finally, there are various date and time parameters to consider. Do users at 3am local time behave the same as users at 3pm local time?

Altogether there is typically about 1KB of relevant data for every banner ad served. A moderate sized internet marketer might place a billion banner ads a month. Larger marketers are routinely 10 times larger.

So, merchants find themselves routinely managing a terabyte of data a month. For perspective, this is about three-times the data volume coming down from the Hubble space telescope. We need to analyze the data to figure out which sequence of which ads works best for different users, on different types of websites, at different times.

The Models:

The general way to parameterize this problem has not yet been figured out. The analytic face of this $26B industry is still sufficiently immature that pretty much each of the thousands of analytic teams working on the problem is inventing their own way to simplify it.

For example, after an event occurs, such as a site visit, one could attribute the value of that event only to the last banner ad displayed. In such a memory-less model, the second-to-last ad, by not being part of the model, is being assigned zero influence. A bit of a silly model, but it may be a good place to begin the thinking.

It is not enough to look only at the banner ads served before the events. One also has to look at ads served before non-events. To build a model that predicts the value of an ad in generating an event requires looking at all the ads that don't generate events as well. We may decide to build a regression model to predict, for every user, an event in the next 10 days, using the approximately 100 variables that describe the last ad served. Maybe not the most sophisticated of models, but at least we are on our way to bringing analytic rigor to this complex problem. A well-constructed model would include interactions so that we would have guidance on how to serve ads differently depending on combinations of ad, user, and location features.

One of the most critical issues we have to manage in this process is correlated predictors. While increasingly merchants and vendors are collaborating to put formally designed experiments into market, such is still the exception, not the rule. Because of this lack of quality experimentation, much of the data available is highly collinear. For example, certain classes of web sites may only be served "value ads" and others only served "performance ads." In such a case, the confounding of the data simply will not enable a decomposition of where to do what.

Of course, one actually wants to use not just the approximately 100 variables of the last ad displayed, but also the 100 variables of the second-to-last ad, and the one before that, and before that. Then one needs to parameterize the history, and sequence, of ads. For example, one could build a variable for the number of ads shown in the last 30 days. Or an indicator for if this user was ever shown a "value ad." Building the deep understanding of how to structure this analysis is the key work now being developed in analytic shops across the internet world.

The Future:

Ultimately, with an understanding of how a sequence of ads generates an event, and an understanding of the value of an event, merchants can make sensible economic bids for ad placement. In an efficient market, where the data flows are mature and the analyses sensible, internet users will be individually served the ads they actually value.

But the industry is far from this end point right now. Much more sophisticated operational capability, ever more insightful on-going experimental design, and many more generations of statistical models will be required to get internet advertising optimized. Learning how to fund the internet through a win-win-win analytically intensive process will be an active, evolving field of research for decades.

Timothy C. Owen (tcowen@travelers.com) and William Kahn (wkahn@travelers.com) work in Strategic Initiatives of Travelers Insurance where, among other responsibilities, they help manage the analytics of internet marketing.