Real-time User Modeling and Prediction: Examples from YouTube

Dr. Ramesh R. Sarukkai
YouTube/Google Inc
1600 Amphitheatre Parkway
Mountain View
CA-94043, USA
sarukkai@google.com

ABSTRACT

Real-time analysis and modeling of users for improving engagement, and interaction is a burgeoning area of interest with applications to web sites, social networks and mobile applications. Apart from scalability issues, this domain poses a number of modeling and algorithmic challenges. In this talk, as an illustrative example, we present DAL, a system that leverages real-time user activity/signals for dynamic ad loads, and designed to improve the overall user experience on YouTube. This system uses machine learning to optimize for user activity during a visit and helps decide on real-time advertising policies dynamically for the user. We conclude the talk with challenges and opportunities in this important area of real-time user analysis and social modeling.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information filtering; I.2.6 [Artificial Intelligence]: Learning— Parameter learning.

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Real-time analysis, user modeling, advertising, optimization.

1. INTRODUCTION

The internet has seen tremendous opportunity and growth over the past two decades – transforming from an information search system to a vital cloud infrastructure that powers a variety of different information needs, social activities and interactive services. Use cases today are extensive ranging from communication, storage, information access, media/video consumption, and commercial services. This ocean of information – whether it be websites, videos, images, or just sheer volume of data that has been made accessible by the internet has but compounded with the emergence and growth of cloud services, cheaper storage, faster CPUs, proliferation of mobile/tablet devices and social sharing. With such continued usage of web powered services and activities, optimizing for user time is vital – real-time data and modeling plays an integral role towards accomplishing that goal. Specifically, we are moving from an era of comprehensiveness to an era of precision for our users – find the right information, content or action for our users, at the right time, in the right context. Advertising systems are also an important part of this equation. In this talk, we will expand on these concepts and go into one application area – Dynamic Ads for YouTube.

2. DYNAMIC AD LOADS

One of the benefits of advertising on YouTube is that we provide users free access to interesting video content, while creating a viable ecosystem whereby content creators benefit from the advertising revenue. In this regard, it is important to ensure that the overall advertising experience is optimal from the end user experience, not just revenue – there are many dimensions to this problem of picking and showing the right ad such as targeting, relevance optimization or just the diversity of creatives participating in the ad auction. We expand this optimization criteria to take into account real-time user engagement signals. How do we tailor such real-time metrics into advertising and combine these optimizations?

We focused on one specific problem: deciding when to show an in-stream advertisement? Traditional media handles this by applying pre-defined breaks evenly distributed over the course of a show. This approach is not effective on the web, especially for short-form content, both due to the dynamic nature of video consumption as well as the diversity of content. An alternate approach to this problem is to take into account users active state and real-time interactions with the site, holistically rather than treat each video view and advertising event in isolation. This dynamic advertising load system (DAL) uses per visit user data and state to train models that predict user response to ads. At runtime, these models are evaluated to predict user response to ads, and determine whether to show ads as the user navigates from one video to the next in his/her session. The tradeoff here involves balancing user happiness (as measured by predicted time spent watching a video) versus the revenue opportunity for our content creators, in a unified manner modeled with a joint cost function. We trained such models using large amounts of data, and evaluated this model using real traffic. The net outcome was a win-win situation, whereby we are able to optimize both for the needs of the user experience (as measured by time spent watching videos) and the overall advertising revenue.

3. CONCLUSION

With the deluge of information, media, and services fueled by the cloud, coupled with a multitude of distractions for the user (devices, apps, sites, social networks), we are in a world where the deciding commodity is user time and attention. Using real-time data and modeling to tailor web applications and ads to balance such user satisfaction metrics with product/revenue goals allows us to improve the overall experience for users. As an illustrative example, we presented the Dynamic Ad Loads system, an algorithmic approach that uses real-time user data to optimize for user watch time with advertising opportunity. We believe this area is fertile with latent ideas and potential applications to many other domains in the web.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.