Data Management and Mining in Internet Ad Systems

S. Muthukrishnan
Rutgers U., and Google Inc.
smewtoo@gmail.com

1. INTRODUCTION

New systems produce data that often present new challenges for data management and mining problems. For example, inventory and sales data led to emphasis on data mining problems such as association rule mining; analysis of Internet Packet traffic logs led to data stream management systems; and, growing markup publication systems led to challenges addressed by semi-structured data management.

Here we are inspired by systems that have emerged in the past decade that enable advertisements (ads) on the Internet. These systems handle billions of transactions every day involving millions of users, websites and advertisers, and are the basis for billions of dollars worth industry. They crucially rely on real-time collection, management and analysis of data for their effectiveness. Further, they represent unusual challenges for data analysis: nearly all parties in Internet ad systems from marketeers to publishers use active, selfish strategies that both help generate new data as well as distort data produced due to their selfish strategies. Mining such data while cognizant of the inherent game theory is a great research challenge. Finally, Internet ad systems use Information Retrieval, Auction and Game Theory, Machine Learning and Optimization Algorithms, and data analysis systems have to be compatible with these methods.

The tutorial will provide an overview of Internet ad systems and discuss in detail both data management as well as data mining tasks that arise:

1. In the first part, we will describe different Internet ad systems and discuss issues in managing the data that arises in them as well as various tools. (Section 2)
2. The second part will be on the data mining problems that arise, many unique to these systems. (Section 3)

More details below.

2. INTERNET AD SYSTEMS

Activities on the Internet can be abstracted as interactions due to three parties. There are

- the *users* who access various web pages and perform activities from searching to browsing,
- the *publishers* who control the web pages and generate the content in them, and
- are *advertisers* who wish to get the attention of the users using the publishers as the channel for placing ads on the pages.

Precisely what ads show when a user accesses a page is a detailed process. There are at least two distinct types of ads:

- *Sponsored Search Ads.* The publisher may be a search engine such as Google, Yahoo or Bing, and users visit these sites to pose queries. Then, ads — typically textual — are targeted to the queries. There is a sophisticated language for advertisers to specify queries and criteria on which they should be matched; search engines do text processing and match ads to queries. Among the matching ads, the search engines choose a subset and rank them for presentation to the user. The number of decisions that have to be made include how to rank ads, how to price the ads, and how many to show. These are decisions are handled via *auctions*, where advertisers bid and the search engines sell off slots. Economics Theory provides guidance on how to design and evaluate auctions: sponsored search has converged to what is called *Generalized Second Price* (GSP) auctions. Further, advertisements typically do not pay just to have auctions shown (pay-per-impression), but only when users click on them (pay-per-click) or when users buy (pay-per-conversion). Advertisers have sophisticated campaigns that target thousands of keywords and critera, have total budgets, and even set up multiple campaigns to experiment with the system to test marketing hypotheses and learn. The ad systems run auctions in real time for each query, billions of times a day, tracking budgets, accountable clicks, and so on; they also estimate quality of ads through empirical measurements, that is crucially used in the auctions.
- *Display Ads.* There are other publishers that get a lot of traffic, such as youtube or nytimes and others. Then, ads – typically images, video and rich media — are targeted to the demographics and other characteristics of the users. Typically publishers and advertisers negotiate contracts offline, and these are typically

Proceedings of the VLDB Endowment, Vol. 3, No. 2

charged per impression. The contracted ads are delivered and accounted by online systems. There are many challenges of forecasting traffic, determining suitable prices, allocating ads online to meet many fairness and optimization criteria, and others. Increasingly, these display ads are sold via *ad exchanges*. When a viewer accesses a webpage, the publisher contacts the exchange where in real time solicits bids from multiple ad networks and based on an auction, returns the winning ad for the viewer. Hence the ad placement becomes more automatic with ad exchanges and decided in real time. Ad exchanges are essentially like financial exchanges except these trade in ads rather than other instruments. There are many challenges in ad exchanges, including what is a suitable auction for ad exchanges, understanding the game theory of intermediaries like ad networks buying on behalf of advertisers, how publishers tradeoff exchange inventory from other alternative sales channels and many others.

There is an increasing amount of publications available about sponsored search and display ads. For example, for sponsored search, the basic auctions were analyzed in [6, 1, 3] and some of the open problems are in [4]. The course at Stanford [2] covers information retrieval and machine learning aspects of Internet ad systems. For display ads, a description of ad exchanges can be found in [5].

In this first part of the tutorial, we will describe the sponsored search and display ad systems (including ad exchanges) in detail. We will discuss the data management issues that arise in sponsored search: in search engine companies to support sponsored search and in advertisers and search engine optimizers that represent them for campaign management. We will also discuss the data management issues that arise in display ads: tools for publishers to optimize their ad inventory, advertisers to manage campaigns across publishers, as well as tools for ad exchange to manage the ad market. The crux of these data management issues involves truly massive data and the need for truly real time analysis (in less than a second).

3. DATA MINING PROBLEMS

While Internet ad systems have been explored by many research communities – Economists, Algorithmicists, Auction and Game Theorists, Machine Learners, etc — data mining community has not embraced these challenges as much, and part of the goal of this tutorial is to bring challenging and crucial data mining problems to the attention of the VLDB community. In this second part of the tutorial, we will describe a set of such problems in detail, focusing on summarizing prior work, and showing where new research is needed. Here are a few examples:

- *Estimating Landscapes.* In order for advertisers to be able to reason through their campaign goals and cost, search engines provide *landscapes*: these are functions of bid that show for each potential bid, the estimated number of impressions (or expected clicks, conversions and so on) one is likely to get if they participated in all the auctions that took place, say during the day. This of course depends on the queries, others' bids, and the auction mechanism. There are two approaches to estimating these functions: one is to use stochastic assumptions and forecast for the future, and the other is to simulate the past auctions and calculate what-if scenarios with different bids. Both these approaches have many challenges, for example with keywords that get very few impressions (the "long tail"). We will describe these challenges.

- *Attribution.* When a user clicks on an ad, how should one attribute the click and what instigated the click? A first order assumption is that the ad on which the user clicked on was the one that got users' attention. However, in many instances, it is likely that an ad that was shown to a user a while ago had residual impact, and indirectly led to the click under consideration. Ads have long term impact and properly modeling and understanding this impact is a challenge. In the tutorial, we will describe graphical models for user behavior that how data can be mined to learn these models, and the impact measures for ads one can derive and validate. There are many other attribution problems in Internet ad systems, such as, when an advertiser sees changes in performance in terms of say the number of clicks for their budget, to what factors should they attribute the change?

- *Price Guidance.* Publishers and advertisers in display ads have to price bundles of goods (eg., slots in headlines page versus sports etc). Likewise, search engines have to set "reserve prices" to increase their revenue. These are instances where data mining is needed using Economic principles to estimate the value of goods. What makes this challenging that the market will react to prices rapidly and will be gamed by participants: therefore, price estimation methods have to be able to model and react to the underlying game theory.

There are many other examples of classical data mining tasks from clustering to detecting anomalies and changes in Internet ad systems, but tasks such as the ones above are fairly unique to ad systems. Tutorial will be partly based on papers found at [8] and [7].

4. REFERENCES

[1] M. O. B. Edelman and M. Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review.*, 97:242–259, 2007.

[2] A. Broder and V. Jasifovski. Introduction to computational advertising. `http://www.stanford.edu/class/msande239/`

[3] A. G. G. Aggarwal and R. Motwani. Truthful auctions for pricing search keywords. *ACM Conference on Electronic Commerce.*, pages 1–7, 2006.

[4] S. Muthukrishnan. Internet ad auctions: Insights and directions. In *Proc. ICALP*, pages 14–23, 2008.

[5] S. Muthukrishnan. Ad exchanges: Research issues. In *Proc. WINE*, pages 1–12, 2009.

[6] H. Varian. Position auctions. *Intl J of Industrial Organization*, 2006.

[7] *Ad Exchange Research.* `http://sites.google.com/site/adexchangeresearch/`

[8] *Algorithms Research Site.* `http://sites.google.com/site/algoresearch/`