Visual Analysis on Online Display Advertising Data

Ling Huang*

Genome from Yahoo!

Abstract

In recent years online display advertising has grown at a rapid pace. Genome from Yahoo! is the big data buying solution for online display advertising. The goal of our platform is to identify the best opportunity to display an ad to a user who is most likely to take a desired action. Our system contains websites which are visited by several million users per day. The number of attributes related to user events is also of the order of several thousand. Visual analysis has emerged as a powerful technique to facilitate demonstrating data, filtering extreme cases and outliers, exploiting data details, and identifying data analysis tasks. With respect to large-scale online data, the paper presents some use cases on visual analysis at Genome from Yahoo!

Keywords: High-Dimensional data, illustrative visualization, statistical graphics, visualization system and toolkit design.

1 Introduction

Online display advertising has become a billion dollar business in recent years. More and more advertisers are reaching out to their customers through online channels. At Genome from Yahoo!, the display advertising server operates billions of advertising targeting events per month. More specifically, there are hundreds of millions of users coming across thousands of websites in our system who are targeted by hundreds of advertisers every day. Advertisers have been able to buy display advertising either by cost-per-impression (CPM), cost-per-click (CPC) or cost-per-action (CPA). The campaign performance directly depends on how well the click-through-rate (CTR) or conversion-rate (CVR) is respectively. For each advertiser campaign, through our systems, all internal and external information of users are collected and hundreds of machine learning models are developed to identify users with a high interest in the brand or a high likelihood of purchasing products in campaigns. As a prerequisite, comprehensively exploring and analyzing user activities become the first crucial step in our platform. However, skyrocketing data volume and complex associations of different categorical data in our platform make it challenging to obtain tangible user insights.

Sheneiderman [1] presented the information seeking mantra as, “Overview first, zoom and filter, then details on demand”. It is a useful guideline for big data analysis, which recommends exploring the overall picture and drilling down for details as needed. Following this guideline, we will present the visualization platforms at Genome from Yahoo! in the rest of the paper. First, a tree-structure system, the patented Open Segment Manager (OSM) is introduced. The OSM platform is used to collect, store and integrate all data using a massive category hierarchy. Then we address how to incorporate OSM hierarchies to visualize and investigate data with different levels of granularity and check details of different level user behaviors and user interests by three use cases.

2 Category Hierarchies of the OSM System

At Genome from Yahoo!, all data are grouped by the hierarchical tree-structure in OSM. There are two types of data in our system: internal data and external data. Internally there are three key components in the display advertising server: user, advertiser, and publisher. Conceptually, the user refers the people who see the ad. User data is identified and stored as browser cookies in the server. The advertiser is the entity charged with executing an advertising campaign. The publisher indicates the online website inventory. Meanwhile, there are external data sources from about 25 third-party data providers, which provide thousands of features of users, advertisers and publishers. Figure 1 shows a 0.1% of the OSM taxonomy tree at Genome from Yahoo!.

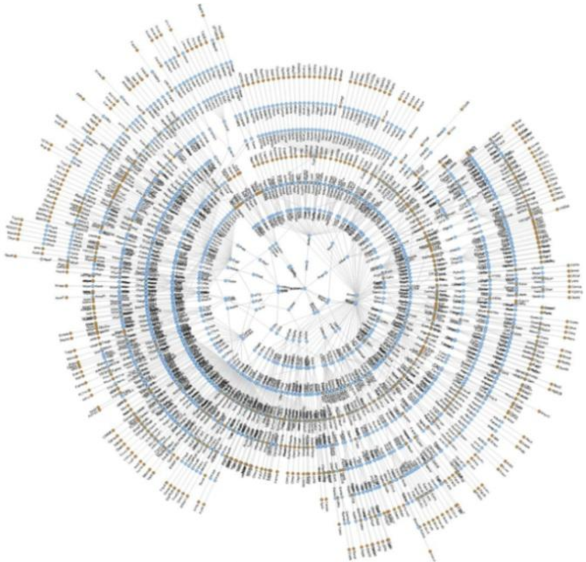

Figure 1: 0.1% of OSM Taxonomy Tree

We can explore and investigate data with different levels of granularity and check details of different level of user behaviors. For example, each advertiser typically runs a series of campaigns using different line items in which to display the ads, each with different markets, targeting methods and goals. This means that there is an advertiser path from the tree root to the tree leaf: Root → Advertiser → Campaign → Line Item → Ad. Similarly, each publisher contains a set of websites, each with different sections. A publisher path from the tree root to the tree leaf is: Root → Publisher → Website Category → Website → Section. As for third-party data, a path of a third party data provider is: Root → Third Party Data Node → Individual Third Party Data Node → Third Party Category → Third Party Attributes. Therefore, using the OSM taxonomy tree, our system is able to collect, consolidate, and store all internal and external data.

Because entities under the same branches of the OSM taxonomy tree share similar characteristics and are incorporated with OSM hierarchies, we are able to identify the appropriate clusters of entities, group the events by the clusters, aggregate the

* Ling Huang is with Yahoo!. Email: lingh@yahoo-inc.com

IEEE Symposium on Large Data Analysis and Visualization 2013
October 13 - 14, Atlanta, Georgia, USA
978-1-4799-1658-0/13/$31.00 ©2013 IEEE

occurrences of the similar events and exploit the patterns and corresponding characteristics efficiently.

3 VISUALIZING AND EXPLORING AD SYSTEM PERFORMANCE

3.1 Time Series Analysis

Our platform records and stores the vast volume of temporal user events. An event corresponds to a user action, aka impression, click, conversion. Every event in our system has a time stamp, which records when the event happened. Our platform processes about 100 billion events, which include about 6 billion impressions from 2 billion unique active users every month.

As for the overview of each campaign, the time serials of aggregated entity counts reveal the time ordering and the volumes of the events for the specific interest clusters. It reveals the sense of the data scale and complexity of the data. Meanwhile, time serial analysis can be implemented for each campaign to understand the internal structure of the data including trend, seasonality, auto-correlation, irregularity, etc. Figure 2 shows the analytic results for a sample time serials using the forecast package in R [2]. Outputs include auto-correlation function (ACF), partial auto-correlation function (PACF), Periodogram, and the best fitted model ARIMA (0, 1, 1) with the forecasted 80% and 95% prediction intervals. The in-depth time serial analysis results can be applied to business needs such as website inventory forecasts, supply and demand management, or dynamic anomaly events detections.

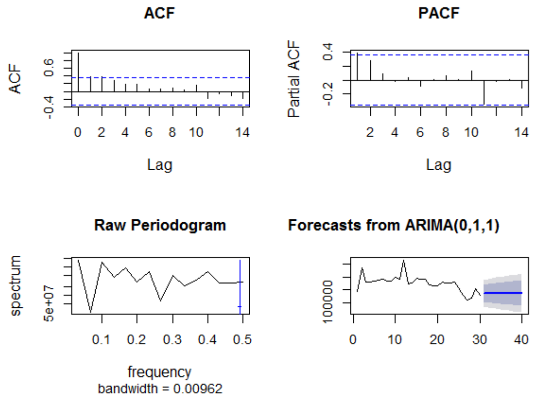

Figure 2: Time Serial Analysis for the Daily Impressions

3.2 Anomaly Detection

Anomaly detection is crucial to big data visualization and analysis. Constructing and developing ad serving systems at scale require more than just brilliant algorithmic design or advanced statistical models. It is important to monitor and evaluate the performances of each and every campaign.

The CVR of a campaign is an unbiased estimate by using the proportion of the targeted users who convert within a short period after serving an ad. In order to filter outliers and extreme cases of many campaigns, we use nonparametric loess smoothers. Figure 3 shows the scatterplot of total impressions and CVRs of 100 advertiser campaigns in one month. The blue line shows the loess smoother, and the 95% confidence bands of the loess smoother are shaded by using grey area. The pattern of the blue line indicates the larger the impressions delivered in a campaign, the higher CVRs can achieve in certain range. Furthermore, campaigns within the 95% confidence bands follow the overall performance trend, but campaigns outside the 95% confidence bands could be under-performing campaigns or over-performing campaigns, indicating possible extreme cases and outliers worthwhile of further study.

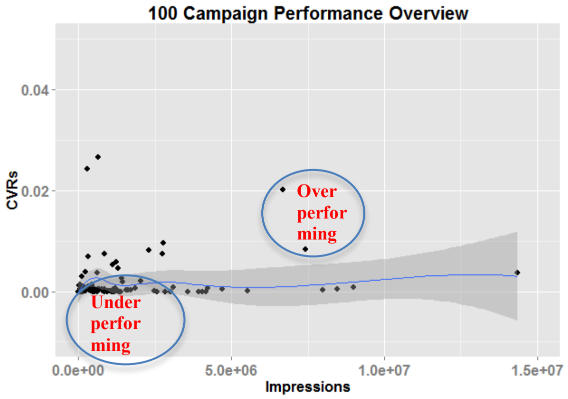

Figure 3: Scatterplot of Total Impressions and CVRs of 100 Advertiser Campaigns in a Month.

As another use case of statistics analysis, Shewart Quality Control Charts (QCC) provide a means to detect when a time varying continuous process exceeds its historic process variation. Figure 4 shows the analytic results of the QCC using the qcc package in R [3]. A process is in control if all points charted lie randomly between the lower control limit and the upper control limit. Group 4 and 6, marked in red, were outside the range and detectably different from the others, so the delivery process was not stable, and needs analysis or intervention to remedy the out-of-control process.

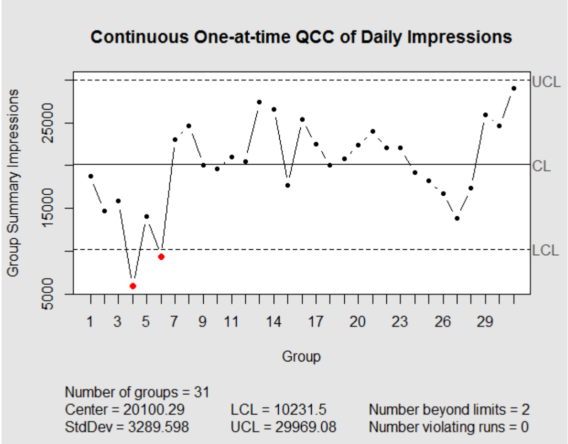

Figure 4: Quality Control Charts of the Daily Impressions

At Yahoo!, the computing environment includes about 40,000 servers which together form a suite of clusters. In future, we actively work on develop user-friendly interactive visualization tools on Yahoo! cluster and support sophisticated analysis tasks to provide more tangible business insights.

REFERENCES

[1] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages, VL '96*, pages 336-343, 1996.

[2] R. J. Hyndman and Y. Khandakar. Automatic Time Series Forecasting: The forecast Package for R. In *Journal of Statistical Software,* volume 27, issue 2, pages 1–22, July 2008.

[3] L. Scrucca. qcc: an R package for quality control charting and statistical process control. In *R News,* vol 4/1, pages 11-17, 2004.