Revenue-Optimized Webpage Recommendation

Chong Wang*, Achir Kalra†§, Cristian Borcea† and Yi Chen‡
*Dept. of Information Systems †Dept. of Computer Science ‡School of Management
New Jersey Institute of Technology, Newark, NJ 07102, USA
Email: {cw87, borcea, yi.chen}@njit.edu
§Forbes Media, 499 Washington Blvd, Jersey City, NJ 07310
Email: akalra@forbes.com

***Abstract*—As a massive industry, display advertising delivers advertisers' marketing messages to attract customers through banners shown on webpages. For publishers, i.e. websites, display advertising is the most critical revenue source. Most existing webpage recommender systems suggest webpages based on user interests only. However, the articles of interest to specific users may not be profitable to publishers. Conversely, only recommending the most profitable articles may lose publishers' user base. To address this issue, we will conduct a series of investigations and design Revenue-Optimized Recommendation, aims to recommend users webpages that optimize interestingness and ad revenue.**

Keywords—Computational Advertising, Recommender System.

I. Introduction

Online display advertising has become a billion dollar industry. In the display advertising ecosystem, advertisers, e.g. Audi, pays a publishers, e.g. Forbes, for space on webpages to display a banner during page views in order to expose their products to interested audience. A page view happens each time a webpage is requested by a user and displayed in a browser. One display of an ad is called an ad impression.

There is some existing research on optimizing advertisers' benefits in display advertising, e.g. [3] and [4]. On the other side of the table, display advertising is the most critical revenue stream of online publishers. A large proportion of profit is put into providing users better information resources and services. However, there is little research about boosting display ad revenue of publishers in real-time ad bidding process. Publishers are still seeking a way to collect more revenue in order to survive from intense business competitions.

The goal of this research is to develop a profitable strategy of ad impression yield for publishers. There are two key factors to increase ad revenue. First, publishers should extend the user base by meeting users information needs. A recommender system which suggests webpages of interest can attract more users. Having more visits, publishers yield more ad impressions. Ad revenue thus tends to ramp up. Second, publishers also enhance the revenue got from individual page views. In our context, each page view's revenue is defined as the average profit the publisher gets from all impressions on the page.

Our preliminary study based on a large publisher's one-month proprietary user log reveals that 1) Webpages with different topics tend to make different ad revenue. 2) Different users have different most profitable topics, i.e. although on average the business channel makes the most revenue, it is probably not held for individual users. Hence, to maximize ad revenue, publishers may choose to only recommend a given user articles with the most profitable topic. However, in this way, publishers will lose user base and thus hurt long-term revenue. On the other hand, if suggesting only the most interesting articles, publishers ad revenue is not optimized because they may not be the most profitable ones. Therefore, it is important to design a strategy that can recommend webpages of interest to users and bring more revenue to publishers.

Designing such recommendation is nontrivial. First, the revenue of each impression is dynamically decided in real-time by advertisers. Publishers have to passively accept any wining bids from advertisers. Second, although there are existing work on advertisers' bidding, it is almost infeasible to simulate the bidding strategies on publishers' side. Publishers do not have user private data and ad campaign targeting information that are two important factors used by advertisers to decide bid prices. Also, advertisers use diverse disclosed bidding algorithms. Third, for interestingness prediction, lacking explicit interestingness rating from users is an issue we will face. Although dwell time, i.e. page stickiness, was proposed to measure webpage relevance, it introduces much noise because users may step away from computers with pages open. This makes dwell time become an unreliable indicator. Finally, this work is a very practical activity. The outcome has to stand a test by actual users in real business environment.

We design revenue-optimized webpage recommendation that can recommend users articles of interest and meanwhile optimize publishers ad revenue. In particular, the proposed method shows a user a list of recommended articles next to the webpage article which he/she is reading. The generation of the list considers both the users interests and profitability of each web article. User interests are interred based on maximum scroll depths. The reasons are 1) Users tend to scroll deeper on interesting pages; 2) It has less noise than dwell time. 3) Ads on the pages with high max scroll depth have high chance to be shown on users' screen, which increases the effectiveness of advertisers' brand promotion. Thus, a model is trained to predict the max scroll depth of any page and user pair. In addition, using knowledge available on the publishers' side, a revenue prediction model is built to predict page view ad revenue. Finally, the ranking of recommended pages is determined by balancing both interestingness and revenue.

II. Dissertation Outline

The project is divided into several phases: First, we have proposed a probabilistic latent class model to predict how far down a user will read on a webpage, i.e. max scroll depth. This can reflect the interestingness of a page to a user. Second, we will predict ad revenue of individual page views. Page view revenue is defined as the average revenue that the publisher receives from each ads on the page. Third, a ranking score is computed for each page by averaging its interestingness and potential revenue. The final ranking list of webpages is delivered to the user. Finally, the proposed method will be evaluated in the real business test bed. We cooperate with Forbes, a well-known online publisher. Forbes will provide required datasets and evaluation environment.

III. Progress to Date

We have completed the work [1] which predicts the viewability of any given page depth, i.e the probability that any page depth of a page view will be shown on the screen. The algorithm proposed in this paper can be used to predict how far down a user will read on a webpage, i.e. the interestingness of a webpage. We define the problem by several important concepts: 1) The *scroll depth* is the percentage of a webpage content vertically scrolled by a user. It is recorded by the last row of pixels on screens. 2) The *maximum scroll depth* of a page view is how far down the page the user has scrolled in that page view. The maximum scroll depth that a user u scrolls on a page a is denoted as x_{ua}. We argue that max scroll depth can highly reflect the interest of a user in a webpage article.

Problem Definition. *Given a page view, i.e., a user u and a webpage a, our goal is to predict the probability that the max scroll depth is x_{ua}.*

The dataset we use is collected over months on Forbes' website by tracking real users' browsing behaviors. It records the information of individual page views, including user ID, user IP, URL, max scroll depth, and time. Intuitively, the characteristics of individual users and webpages can be utilized to improve the performance of max scroll depth prediction. However, the significant features, e.g. topics and user interests, are hard to be captured due to lack of data and the ambiguity of user-webpage interaction. In addition, data sparsity is another challenge. While a large publisher may have several thousands of pages, a user only visits a few.

To overcome above issues, we propose a probabilistic latent class model (PLC) that discovers latent classes of users and pages. The intuition is that different classes of webpages and users tend to generate different levels of max scroll depths. The PLC can detect classes of users and pages that share similar patterns of maximum scroll depth. The class membership of each user and page are learned from the dataset. It outputs the probability $P(x_{ua}|u,a)$, where x_{ua} is the max scroll depth that a user u reaches on a page a. The PLC works as follow:

$$P(x_{ua}|u,a) = \sum_{i=1}^{N_s}\sum_{j=1}^{N_p} P(s_i|u)P(p_j|a)P(x_{ua}|s_i,p_j) \quad (1)$$

where N_s is the number of latent user classes, and N_p is the number of latent page classes. $P(s_i|u)$ is the probability that u belongs to s_i, while $P(p_j|a)$ is the probability that a belongs to p_j. $P(x_{ua}|s_i,p_j)$ is the probability that the max scroll depth is x_{ua}. It can be approximated by the probability density function of the normal distribution (Formula 2).

$$P(x_{ua}|s_i,p_j) = \frac{1}{\sigma_{s_ip_j}\cdot\sqrt{2\pi}} * \exp\left(-\frac{(x_{ua}-\mu_{ua})^2}{2\sigma^2_{s_ip_j}}\right) \quad (2)$$

The mean of the normal distribution, μ_{ua}, is modeled by a regression model whose features are extracted from the history of u and a as well as the context of the page view, i.e., $\mu_{ua} = \sum_m^M w_{spm} f_m^{ua}$. f_m^{ua} is the mth feature and w_{spm} is the weight of the mth feature. Each pair of latent user class s_i and latent webpage class p_j has a set of $w_{s_ip_j*}$ and $\sigma_{s_ip_j}$. The M features extracted based on empirical measurement include the mean max scroll depths made by u and on a, the most recent three max scroll depths made by u and on a, interaction of the mean max scroll depth of u and that of a, user geo, and devices. Let $\mathbf{W}$ be the collection of the weight vectors w_{sp*}. $\boldsymbol{\sigma}$ is the collection of the standard deviations σ_{sp*}. The Expectation Maximization (EM) algorithm is adopted to determine the parameters ($P(s|u)$, $P(p|a)$, $\mathbf{W}$, $\boldsymbol{\sigma}$), which can maximize the corresponding likelihood function. To predict the max scroll depth of a user and page pair, we just need to find the $x_u a$ that can maximize the probability $P(x_{ua}|s_i,p_j)$. This $x_u a$ is the predicted max scroll depth of the user and page pair. The max scroll depth will be used to represent the interestingness of a page to a specific user. We have compared our model with a number of comparison systems. The results demonstrate that our method outperforms the others.

IV. Future Work

The next step is to predict ad revenue of individual page views. The main challenge is that publishers have much less information about users and targeting requirements of ad campaigns. However, advertisers who determine bid prices in the real-time possess much data including user visiting history and target audience. Also, different advertisers use diverse bidding strategies. Thus, predicting page view revenue on the publishers' side is non-trivial. [2] is the only existing work predicts ad revenue on the publishers' side. The authors use a regression model with explicit features. But still not all significant features are accessible to all publishers. Moreover, since advertisers adopt diverse real-time bidding algorithms, features are weighted variously by different advertisers. Having the interestingness and revenue of a page view, we are going to develop a webpage ranking function that can balance both factors. At last, we will conduct evaluation in the real business test bed in terms of interestingness rating and revenue lift.

Acknowledgment

This work is partially supported by National Science Foundation (NSF) CAREER Award IIS-0845647, NSF Grant No. CNS 1409523, Google Cloud Service Award and the Leir Charitable Foundations. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not reflect the views of the funding agencies.

References

[1] Wang, C., Kalra, A, Borcea, C., Chen, Y. (2015, Oct). Viewability Prediction for Online Display Ads. CIKM 2015. Melbourne, Australia.

[2] Cheng, H., Manavoglu, E., Cui, Y., Zhang, R., & Mao, J. Dynamic ad layout revenue optimization for display advertising. ADKDD 2012.

[3] Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. How much can behavioral targeting help online advertising?. WWW 2009.

[4] Weinan Zhang, Shuai Yuan, & Jun Wang. Optimal real-time bidding for display advertising. KDD 2014.