Post-Click Conversion Modeling and Analysis for Non-Guaranteed Delivery Display Advertising

Rómer Rosales
Yahoo! Labs
4401 Great America Parkway
Santa Clara, CA
romerr@yahoo-inc.com

Haibin Cheng
Yahoo! Labs
4401 Great America Parkway
Santa Clara, CA
hcheng@yahoo-inc.com

Eren Manavoglu
Yahoo! Labs
4401 Great America Parkway
Santa Clara, CA
erenm@yahoo-inc.com

ABSTRACT

In on-line search and display advertising, the click-trough rate (CTR) has been traditionally a key measure of ad/ campaign effectiveness. More recently, the market has gained interest in more direct measures of profitability, one early alternative is the conversion rate (CVR). CVRs measure the proportion of certain users who take a predefined, desirable action, such as a purchase, registration, download, etc.; as compared to simply page browsing. We provide a detailed analysis of conversion rates in the context of non-guaranteed delivery targeted advertising. In particular we focus on the post-click conversion (PCC) problem or the analysis of conversions after a user click on a referring ad. The key elements we study are the probability of a conversion given a click in a user/page context, P(conversion | click, context). We provide various fundamental properties of this process based on contextual information, formalize the problem of predicting PCC, and propose an approach for measuring attribute relevance when the underlying attribute distribution is non-stationary. We provide experimental analyses based on logged events from a large-scale advertising platform.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Commercial Services; H.4.m [**Information Systems**]: Miscellaneous; I.5.2 [**Design Methodology**]: Classifier Design and Evaluation

General Terms

Algorithms, Design, Experimentation

Keywords

Display Advertising, Non-Guaranteed Delivery, Conversion Rate, Conversion Modeling, Post-click Conversion

1. INTRODUCTION

Display advertising is a fast growing on-line advertising medium where advertisers pay publishers to place graphical ads on their web pages. There are two ad purchasing and delivery mechanisms that are commonly used today. Guaranteed Delivery (GD) is the traditional mechanism where an advertiser buys a predetermined number of impressions (5 million) at a fixed price (10 cents per impression) for a fixed time period (August 2011) from the publisher. In this model, the publisher is obligated to deliver the agreed upon number of impressions that satisfy the advertiser's targeting requirements (*e.g.,* females, between the age of 25-30, in California). The second mechanism, Non-Guaranteed Delivery (NGD), is a spot market where the advertisers can buy ad impressions one at a time. Under this mechanism every time a user loads a web page with an ad slot, an auction is run among the ads that match the targeting specifications of that particular opportunity and an ad is chosen for display.

Guaranteed delivery is preferred by risk-averse publishers and advertisers as it provides one with a steady and predictable source of income, and the other with an exact exposure rate. Spot markets, on the other hand, allow the advertisers to change their bids based on highly granular targeting profiles, and let them adapt to varying trends in traffic patterns quickly.

Spot markets also offer advertisers a wider range of payment models. Like in guaranteed display, advertisers can choose to pay per impression (CPM). This model works well for the advertisers who are trying to build brand awareness where getting the message out is the goal. However, for many advertisers, an ad impression that does not ultimately lead to a visit to their website or to a product purchase is not worth paying for. To address this concern many non-guaranteed display exchanges provide performance-dependent payment methods such as pay per click (CPC) and pay per conversion/action (CPA) [21]. In the pay per click model, an advertiser will not be charged unless the user clicks on their ad. Pay per conversion/action model reduces the risk for the advertisers even further by allowing them to pay only when the user takes an action that is of interest to them. Advertisers have complete control over the definition of these conversion actions. Example actions include but are not limited to: subscribing to an email list, adding an item to the shopping cart, or making a purchase.

In a marketplace where advertisers with different payment types will be competing for the same ad slot, the auction mechanism needs to convert these bids that are in different currencies to a common base, such as a monetary unit. *Expected* price per impression (eCPM) is a natural choice for such a common unit. For CPM ads, the expected price per impression (eCPM) would be the same as their bid for

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.

that impression. For ads that are defined as pay per click (CPC) or pay per conversion (CPA), their eCPM will depend on the expected click or conversion rate of the given impression. More precisely, $eCPM(CPC) = p(click) * bid$ and $eCPM(CPA) = p(conversion) * bid$, where $p(click)$ is the probability that an impression will result in a click, and $p(conversion)$ is the probability that an impression will lead the user to take the actions that constitute a conversion for that advertiser.

Accurate estimations of these probabilities are critical for the efficiency of an exchange [22]. While the problem of estimating click probabilities have been studied extensively in the context of both search [17, 27, 11, 25, 2, 29, 16], contextual [24, 9] and display advertising [10, 14], literature on conversion prediction is much sparser. In this paper we try to address precisely this issue by providing empirical answers to practical questions that arise in conversion prediction and its application to NGD: How different is conversion modeling from click modeling, can we apply lessons learned from clicks directly to conversions? Previous studies suggest that the delay between a click and a conversion event could be quite long, what does the data tell us? How tolerant are our models to temporal changes? How important are user and publisher attributes in predicting conversions?

The focus of our analysis is the conversion probability for clicked ads. While a click is not necessary for a conversion event to happen, most CPA advertisers do not give credit to publishers unless they can trace the conversion event back to an impression on the publisher's page via a click event; thus making the click a prerequisite for a conversion action (hence, this is often called a post-click conversion or PCC).

The rest of the paper is organized as follows. In Section 2 we discuss related work and in Section 3 we describe the data that will be used in the rest of the paper. In Section 4 we present the results of our analysis on the various aspects of post click conversion problem. Section 5 follows by introducing our baseline model and feature selection algorithm. Then in Section 6 we present our experimental results. We finally conclude in Section 7 by summarizing our findings.

2. RELATED WORK

Although the importance of conversion modeling is widely recognized in the computational advertising community [7] the published literature on the subject has only recently begun to accumulate. Becker et al. [4] analyze the impact of landing page on the user's conversion behavior in the context of sponsored search. They first provide a taxonomy for the advertiser landing pages and then show that the conversion rates varies significantly from category to category. Our paper is complementary to this work because we identify what other factors could be playing a role in the user's conversion behavior. In [3], the authors propose a new model that optimizes the conversion funnel even for CPC campaigns. Since the authors evaluate their model for contextual advertising their analysis of conversion prediction is heavily focused on keyword matches. While the auction model proposed in that study could easily be applied to display advertising, their results on conversion modeling would not be directly applicable to graphical ads. In [28] Shi et al. propose the use of Kullback-Leibler divergence between the default distribution of user attributes and the user specific distribution as a feature selection mechanism to build compact user profiles. They show that these compact user profiles improve the performance of conversion models that are used for behavioral targeting.

The work by Agarwal et al. [1] is perhaps the closest to our paper in terms of domain. The authors propose a log-linear model to estimate click and conversion rates, which exploits the correlations in aggregates at multiple resolutions. Their results show that this model performs better than vanilla models that do not make use of the hierarchy that's found naturally in the ad data. While the focus of their paper is purely the modeling aspect, the authors also report observing differences in relative model performances on PCC vs click prediction.

While display advertising has been used from the early days of Internet, data driven research and analysis on the economic impact of it has been relatively sparse. Results published by Reiley and Lewis [26] provide the strongest quantitative analysis to the best of our knowledge. The authors ran a controlled experiment on over a million users who could also be linked with their offline activities and reported that the users who were exposed to the ads showed a statistically significant increase in both in store and online sales. Lewis was also able to show that exposing the same ad to the same user over and over again has diminishing returns [20] and that display advertising has a bigger impact on older users [19].

3. DATA DESCRIPTION

We collected live traffic logs from Yahoo!'s Right Media Exchange (RMX), one of the largest ad exchanges in the world, which serves around ten billion impressions every day. RMX follows a "network-of-networks" model where the connections between advertisers and publishers are facilitated by intermediaries called networks. Networks can have either only publishers or advertisers, or both. Every entity in the exchange (networks, publishers and advertisers) has a unique identifier.

Publishers label their web-pages with site id's. They can also tag different parts of the same page with a different section id. While the exchange does not enforce any rules on the semantics of section id's, publishers often use a different section id for different ad slots on the same page.

An insertion order represents the contractual relationship between a network and an advertiser or a publisher, or between any two entities on the Exchange. A line item is an element of an insertion order. An insertion order can have one or more line items, each with a different pricing type, budget and targeting profile. Advertisers can have multiple ad groups within a line item, each with a custom target profile, and different budget settings. In RMX one ad group can also be associated with multiple line items. In this paper we represent each {line item, ad group} pair as a campaign. In RMX, the relationship between campaigns and ad creatives are not necessarily hierarchical. Advertisers can use the same creative in multiple campaigns.

For CPA campaigns advertisers also need to select the type of conversion (post-click or post-view), and set the time out period for a conversion to be valid after the preceding event. Advertisers can also specify how frequently a conversion is allowed to repeat (*e.g.,* no consecutive conversions can trigger within one hour of the last conversion). Conversions are tracked using pixels, and each distinct conversion action is assigned a pixel id. Advertisers are allowed to reuse existing conversion pixel ids (conversion ids).

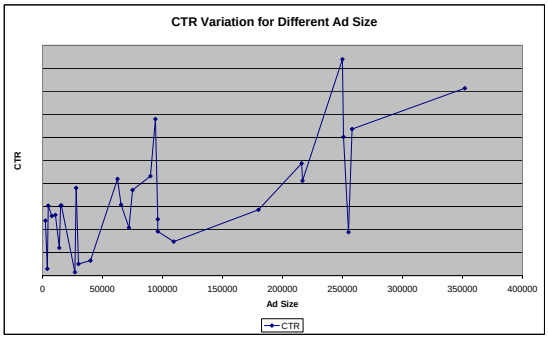

Figure 1: Relative CTR variation for different ad sizes (measured in pixels).

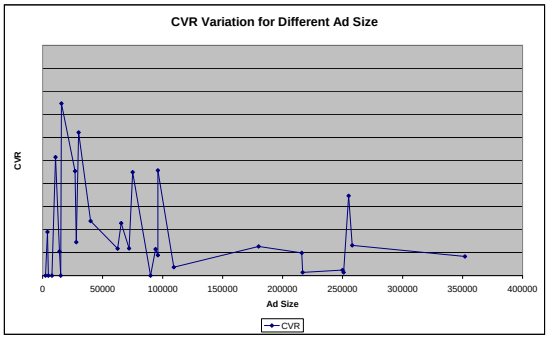

Figure 2: Relative CVR variation for different ad sizes (measured in pixels).

4. POST-CLICK CONVERSION ANALYSIS

In this section we provide an analysis of the post-click conversion problem using a large collection of logged events. The goal of this analysis is to help us uncover special problem properties and patterns of user behavior to guide the modeling tasks discussed later. For this we use data collected for a period of one month containing approximately 1 billion user events involving an ad click.

4.1 Contrasting CTR and CVR

We are interested in uncovering the fundamental properties of the post-click conversion (PCC) event. PCC occurs after a user has clicked on a relevant ad. Thus, a basic question is whether there is evidence that elements of importance in modeling click-through rate (CTR)[1] are also relevant in modeling PCC conversion rate (CVR). We investigate this by exploring three important attributes in click modeling, the user age, the user gender and the ad size.

In click modeling, ad size is strongly correlated with CTR. This makes intuitive sense likely because large ads are on average more attractive and more attention-grabbing to users than small ads, resulting in measurably higher CTR. This has been confirmed in the present data. Figure 1 shows the CTR variation (y-axis) with respect to different ad sizes in terms of number of pixels (x-axis). The absolute CTR value in the Figure is removed to comply with company policy. Due that ad sizes have been predefined (to a large extent) in the on-line advertising industry, only a few points in the x-axis can be measured. Figure 1 clearly indicates that the CTR of ads increase approximately linearly with the ad size. Using the conversion logs for the same period, the post-click conversion rate for different ad sizes is calculated and plotted in Figure 2. In contrast with the trend observed in CTR with regard to the ad size, the CVR for users does not increase when the ad size increases. The size of the ad appears to have no effect on the user's decision on converting or not likely because the user has been directed to a different page (the ad landing page) and has lost the visual perception of the ad on the initial page view. In summary, this provides evidence that the ad size may not be an informative element for PCC modeling, particularly when compared with click modeling.

A similar analysis was extended to user profile attributes such as age and gender. Using the same click logs, Figure 3 shows the variation of click through rates for different age groups and genders (male, female and unknown). There is an unexpected high CTR rate for users with (mostly) unknown age and unknown gender, otherwise we did not see a strong trend in terms of CTR with respect to age groups. However, male users seem to click more often than female users at a younger age, with the trend reversing (to a small extent) for older users. Figure 4 shows the variation of post-click conversion rates for different age groups and genders. We can observe that female users tend to convert more often than male users. Also there is strong linear trend in terms of conversion rate with respect to the age groups. Older users seems to covert more often than younger users. Thus, unlike our observation for ad sizes, the age and gender can clearly play a more important role in post-click conversion modeling.

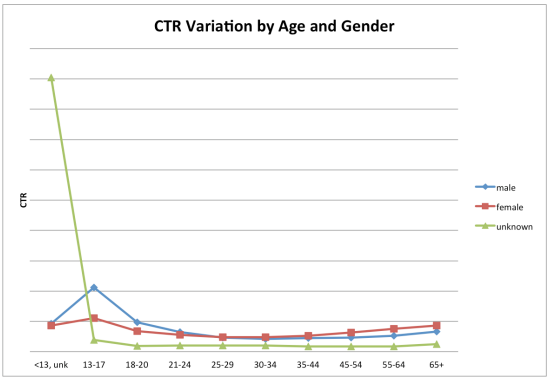

Figure 3: Relative CTR variation for different age and gender categories.

4.2 Analysis of the Click→Conversion Delay

A critical aspect in PCC modeling is the association or attribution of a conversion event to the corresponding click event. In order to build a conversion model, we need to attribute the conversion event to the correct click event as this represents a positive PCC example (vs. a click event without any associated conversion event). A conversion event can happen minutes, hours or even days after a click event. The proper attribution of the conversion event to the right click event, which can done by properly matching the click and conversion event attributes, is essential not only for PCC but also for payment processing.

[1]The probability that a user clicks on a given ad.

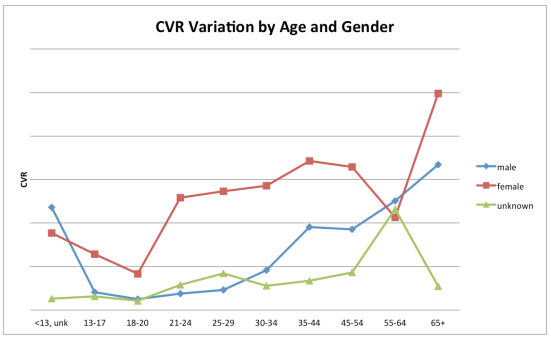

Figure 4: Relative CVR variation for different age and gender categories.

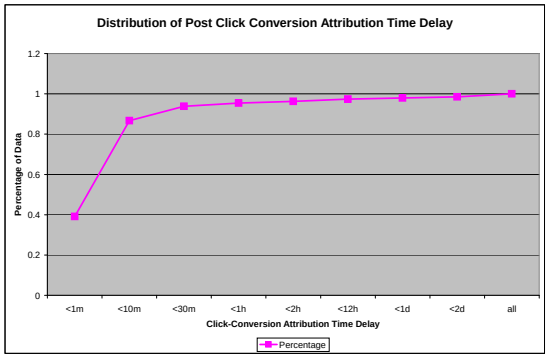

Figure 5: Distribution of click conversion attribution time delay.

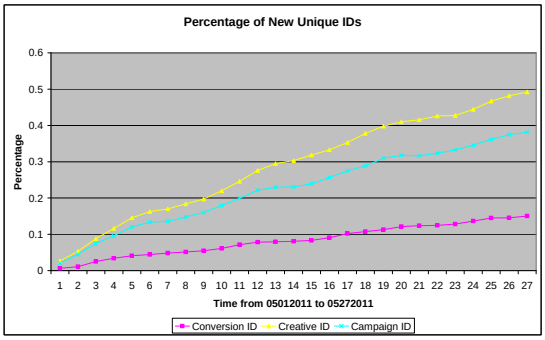

Figure 6: Percentage of new unique conversion identifiers, creatives and campaigns emerging for each day in one month (relative to those existing in the previous month).

In general several conversion events could be associated with the same click, as advertisers may show a sequence of conversion-generating pages to the user after a click on the relevant ad. On the other hand, a conversion event may not necessarily be associated with any ad click[2]. This association process faces certain practical limitations as the longer the time elapsed between click and conversion the more logged events that need to be maintained and matched. In order to get a better picture of the click-conversion process and to answer the question of how much data needs to be utilized for matching conversions with their corresponding click events, we analyzed the properties of the time delay for conversion events.

We calculated the percentage of conversion events with different attribution time intervals as shown in Figure 5. From the graph, we can observe that a large majority (86.7%) of conversion events are triggered within 10 minutes of the click events. Approximately half of these conversion events (39.2%) occur within 1 minute of the corresponding click event. If we look further, we observed that we can match 95.5% of the conversion events within one hour of the clicks. As we are interested in achieving the largest possible recall within practical boundaries, we decided to consider various days of delay. Within two days of the click, 98.5% of the conversions can be recovered. Thus, we would be ignoring approximately 1.5% of the conversion events and as a consequence incorrectly labeling a click event as referring to a negative conversion (no conversion) if the time frame set for post click conversion attribution is limited to 2 days. This was believed to be sufficient given the practical cost of looking further in time; and thus, we utilize this limit throughout the paper.

4.3 Ad Emergence Rate Analysis

Post-click conversion models are built based on data collected in the historical logs, using performance information of ads that have been in the advertising system for some time. When completely new ads are added to the advertising system, models built in the past may not perform well particularly for those new ads (this clearly depends on the generalization power of the attributes utilized for modeling). This property of the system raises some challenges regarding how to keep a conversion prediction model up-to-date. One strategy is to constantly update the model using the latest data[3]. However, this could be too expensive for practical purposes and potentially unnecessary.

In order to design a model updating mechanism (that is able to cope with the dynamic nature of campaign generation) with reasonable trade off between performance and latency, it is important to investigate the ratio of new ads emergence with respect to time. In this section, we calculate the percentage of new ads arising daily for a period of one month relative to ads already existing in the previous month. Three representation levels of the ads are investigated: conversion, creative, and campaign level (these also turned out to be the most informative advertiser-based features in the conversion models studied).

Figure 6 plots the percentage of new unique ads for each day. It is clear from the figure that the percentage of unique new ads in the data is increasing steadily day by day in terms of all the three representations. There is a difference in terms of the emerging rate. Creatives are observed to change most frequently. There are 19.7% new creatives after 10 days and 39.8% after 20 days. Ad campaigns also increase steadily, but at a slower rate, with approximately 16% increase after 10 days and 31% after 20 days. Conversion identifiers are the most stable of the three, with only 5.4% of new unique ids

[2]This can be analyzed in the context of post-view conversion, a process complementary to post-click conversion

[3]There are other phenomena that motivates frequent model updates, including the dynamic nature of on-line information as changes in page-viewing behavior in intrinsically related to ad effectiveness

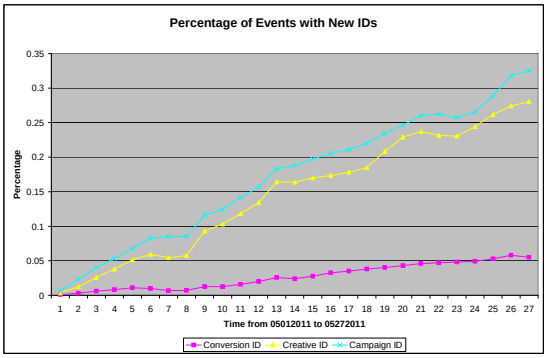

Figure 7: Percentage of events with new conversion identifiers, creatives, and campaigns for each day in one month (relative to those existing in the previous month).

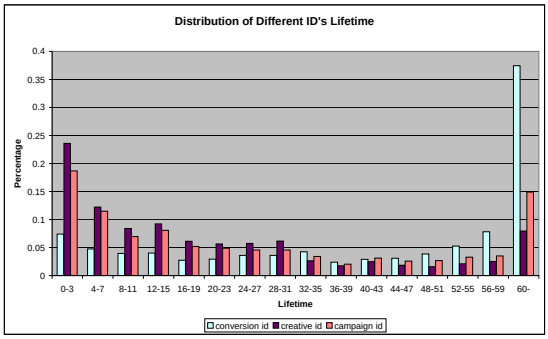

Figure 8: Distribution of lifetime for different IDs.

after 10 days and 11.2% after 20 days. This indicates that advertisers tend to use the same conversion identifiers to track the respective conversion events even after they create new ad creatives and campaigns.

This is beneficial to some extent to PCC modeling since a model relying on conversion identifiers should be relative stable with respect to changes in time. Furthermore, the linear trend of the emerging rate in campaigns and creatives may indicate that the performance dependence of the model with respect to those features is linearly related to its updating frequency.

Similar conclusions can be reached by looking at the percentage of events with new campaign, creatives, and conversion identifiers with regard to the total number of daily events using the same time period as before. This is shown in Figure 7. A proportion of 1.3% of the events contain new conversion identifiers after 10 days and 4.3% after 20 days respectively. This is considered to be a relatively small increase for practical purposes as a model not updated for a very large period (*e.g.,* 10 days) will still consider all the available conversion identifiers in the system.

4.4 Ad Life-time Analysis

The previous section indicates that there are new ads emerging day by day. However, it does not tell their actual lifetime, another relevant property. Figure 8 plots the distribution of the actual lifetime of ads at three levels: conversion id, campaign, and creative levels. 37.4% of conversion ids lives longer than 2 months, which is quite significant compared to 8% for creative ids and 14.9% for campaign ids. 23.6% of creative ids and 18.7% of campaign ids have a lifetime shorter than 3 days, while this number for conversion id is only 7.4%. It is notable that conversion id lives much longer than the creative id and campaign id, which is consistent with the conclusion reached in Section 4.3.

5. POST-CLICK CONVERSION MODELS

In this section we focus on PCC prediction modeling, including various types of feature definitions, and the problem of feature selection given the properties of the attribute distributions in PCC (in particular their non-stationary nature).

5.1 Predicting Conversions after a Click

Non-guaranteed delivery (NGD) advertisement is a large and complex subsystem of the more general on-line display advertisement market. In NGD, multiple market players including publishers, advertisers, ad networks, and ad exchanges, try to optimize ad placement in a distributed manner. Various elements are fundamental in this large ecosystem, including prices, inventories, and (crucially) expected user behavior. The details of this overall system are beyond the scope of this paper. However, the latter element is of particular importance as revenue for a large sub-class of advertisement is tied to user actions as related to ads. One class of such actions is generally referred to as a conversion: a predefined user transaction such as a purchase, registration, download, referral, etc. The focus of this section is on estimating a crucial quantity in this system, that is the probability that an on-line user (*e.g.,* page visitor) will trigger a conversion after clicking on a specific advertisement. We call this the post-click conversion (PCC) prediction problem.

We characterize and formalize PCC prediction as a machine learning problem. For this, we utilize the logs from a large-scale advertising platform to identify the conversions and associate them with their corresponding ad clicks. More specifically, we are interested in modeling the conditional probability of a conversion given that a user clicked on a relevant ad. Two classes of events are of interest in order to approach this problem: 1) events where a user clicked on an ad but did not convert are considered negative conversion instances, whereas 2) events where a user clicked and subsequently converted are considered positive conversion instances.

Given these definitions, let a set of events be represented by $\mathcal{D} = \{\mathbf{x}(b_i, a_i, u_i), y_i\}$ where $\mathbf{x}(b_i, a_i, u_i) \in \Re^d$ is a representation of the information available about the event, normally associated to a publisher b_i, advertisement a_i and user u_i; whereas $y_i \in \{0, +1\}$ is an indicator of whether a conversion was associated to the event. Thus, given some or all event information, the goal is to produce a probability of a conversion $p(y|b, a, u)$, or more succinctly $p(y|\mathbf{x})$.

Throughout this paper we utilize a maximum entropy model of the above distribution. This, also referred to as logistic regression in statistics[13], is a log-linear model that combines the individual contributions of each element of the feature vector representation through a weighting $\mathbf{w} = (w_1, ..., w_d)$. Formally:

Table 1: Summary of basic features considered divided into feature families

Feature family	Feature members
Advertiser	advertiser (id), advertiser network, campaign, creative, conversion id, ad group, ad size, creative type, offer type
Publisher	publisher (id), publisher network, site, section, url, page referrer
User (when avail.)	gender, age, region, network speed, accept cookies, geo, user segments
Other	serve time

$$p(y=1|b,a,u) = \frac{1}{1+\exp(-\sum_{i=1}^{d} w_i \mathbf{x}_i(b,a,u))}, \qquad (1)$$

where $w_i \in \Re$ weighs the feature specific contributions.

We find the optimal values for w_i by employing the maximum-a-posteriori probability criterion: $\mathbf{w}^* = \text{argmax} p(\mathbf{w}|\mathcal{D})$. We attach a Gaussian prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \Sigma)$ to $\mathbf{w}$ and solve:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}\in\Re^d} \sum_{i=i}^{N} \log(p(y_i|b_i,a_i,u_i)) + \log p(\mathbf{w}). \qquad (2)$$

The problem given by formulation (2) is a convex optimization problem and therefore has an unique maximum [6]. It can be solved using a variety of techniques [23]; however, for large-dimensional representations (large d) most approaches face practical limitations related to computing and storing the second-order information of the objective function. Non-linear conjugate methods are a feasible alternative due that they rely on computing the function gradient and only an estimate of its Hessian. We utilize a distributed conjugate gradient algorithm with a pre-conditioner to accelerate convergence [5].

5.2 Features

When building personalized conversion models, it is fundamental to gather the appropriate information about the different axes of personalization. We consider four sets of attributes or features available in most on-line advertisement systems as depicted in Table 1. These attributes are obtained at the event level, that is every time a user clicks and (potentially) makes a conversion on a corresponding ad-publisher pair. However various forms of feature aggregations are utilized and explained below. The first set is represented by advertiser-dependent attributes such as its identifier and ad-specific information for the click event in consideration. The second set correspond to publisher information where the event (click or conversion) took place. The third set includes some pieces of available user information. It is worth mentioning that the availability of these attributes varies across events. The last set includes other information such as the time of the transaction. Because of their *fine-grained nature*, some of the above features carry very little information in their original form (*e.g.,* user age and the event times); thus, these have been quantized appropriately.

The use of feedback features in conversion modeling relies on the simple observation that ads performing better in the past will likely perform better in the future. In PCC modeling, statistics such as the total number of clicks and the conversion rate (conversions/clicks) are often utilized to measure the performance of ads. Feedback features are derived from these statistics by aggregating the historical data for ads at different levels such as advertiser, campaign, creative, and conversion id organized by advertisers as described in Table 1. To capture the variation of ad performance for different publishers, similar features can also be obtained by paring the advertisers at different levels with publishers, *e.g.,* . the publisher-creative pair, the site id-creative pair, etc. Similarly, aggregations such as age- creative, gender-creative, and geo-creative are expected to be good feedback features because they can capture the variation of past ad performance for different users. Feedback features are quantized using a simple k-means clustering algorithm [8] before they are fed to the maximum entropy optimization algorithm. Note that feedback features are often refreshed regularly by updating the statistics with the latest historical information available.

5.3 Automated Feature Analysis

As seen in the analyses of Section 4, post-click conversion (PCC) events have the property that the distribution of certain attributes (*e.g.,* campaign, creative, conversion identifiers) evolve considerably with time. From a machine learning perspective, this raises some difficulties as the underlying attribute distribution is non-stationary. A feature value occurring with some regularity in the training data may not necessarily occur in the same manner on-line (at inference/classification time) and thus a machine learning model would likely learn from values that are not as representative of the test data as expected[4] Two reasons for the above phenomenon include a) the underlying dynamic nature of the attribute distribution (as seen in Figures 6-8) b) the fact that some attribute values are relatively unique to the data point (*i.e.,* logged event) and cannot be used to generalize in a machine learning sense. The latter is an extreme case of the former and includes features such as event identifiers, non-processed time-stamps, etc. Our goal is to automatically discover the relevant/non-relevant attributes with as little human involvement as possible. This is of clear importance when building/updating models regularly in a live system.

This problem can be studied in the context of the mutual information (MI) score, an information theoretic quantity employed to measure the degree of dependence between random variables[18, 12]. In our particular context, let $I(X_j, Y)$ represent the information in feature X_j about Y. A widely used feature selection method consists of selecting the features X_j such that $I(X_j, Y)$ is *high*, with Y the target attribute (*e.g.,* whether a conversion occurred). For example, select the top ranked features [15].

For features with a dynamic distribution as in our case, the MI calculated on a training set fails to provide a good measure of feature relevance. This can be illustrated as follows: Let X_u be a random variable taking unique values (X_u can for instance be an event identifier), then $I(X_u, Y) = H(Y)$ since the values of X_u can fully identify the data point and therefore its label. Formally:

[4]The degree to which this distribution is non-stationary depends on the choice of features.

$$I(X_u, Y) = \sum_{x,y} p(x_u, y) \log p(y|x_u)/p(y) \quad (3)$$

$$= \sum_{y} p(y) \log 1/p(y) = H(Y), \quad (4)$$

since $p(y^*|x_u) = 1$ for some $y = y^*$ (zero otherwise). However, X_u is useless as an attribute for predicting y since its values are unique and not observed in any test set.

In order to address this problem we propose using a related function that explicitly considers a reference distribution. Let the reference distribution be given by $\tilde{p}(x, y)$, then define the MI with respect to the reference distribution by:

$$I_{\tilde{p}}(X_i, Y) = \sum_{x_i,y} \tilde{p}(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}, \quad (5)$$

where the difference compared with the standard mutual information function lies in calculating the expectation with respect to the reference (not training) distribution.

This definition has the problem that the log ratio is undefined for cases when $p(x_i) = 0$. This happens when an attribute value has been seen in the reference distribution $\tilde{p}$ but not in the training set distribution p. Thus, we utilize a smoothing of the training data distribution of the form:

$$p_r(x_i, y) = \frac{Np(x_i, y) + p(y)}{N + |X_i|}, \quad (6)$$

guaranteeing $p_r(x_i) > 0$, where $|X_i|$ is the number of states taken by X_i. It is possible to show that, if $(\forall y)p(y) > 0$, this does not affect the target distribution, that is: $p_r(y) = p(y)$. In the critical case where X_j does not appear in the training data distribution, we can show that $p_r(y|x_i) = p(y)$. In the latter case we have that the log ratio above becomes 0.

The main relevant property of the new information quantity is that as attributes are evaluated on a reference distribution, spurious relationships (such as those seen above) found in a specific data set and that do not generalize to the test data set are mostly ignored. Consider the same illustrative example above: the values of X_u do not appear in the test distribution (as these values are unique to a data point); thus, the test distribution will place no mass on these values, and X_u will have no measured information about the target variable of interest. More formally:

$$I_{\tilde{p}}(X_u, Y) = \sum_{x_u,y} \tilde{p}(x_u, y) \log \frac{p_r(x_u, y)}{p_r(x_u)p_r(y)} = 0. \quad (7)$$

We note that in many instances, any reference distribution $\tilde{p}$ computed on a valid sample (not necessarily the test data) different than the training distribution will allow the proposed measure to avoid spurious relationships particular to the training data. However, a reference distribution closer to the test distribution is preferred as test-specific dependencies will be better captured by the new MI definition.

6. MODELING EXPERIMENTS

In this section we further provide experimental evaluations involving conversion modeling. In particular, we study the importance of publisher information, the impact of model update frequency and methodology prediction accuracy, and the use of more automated feature selection methods in both model interpretability and accuracy. All of these experiments optimize formulation (2) using $\Sigma = \lambda^{-1}\mathbf{I}$, with $\lambda \in \Re$ and employ a Map-Reduce implementation of the pre-conditioned conjugate gradient method referred in Section 5.1.

Except where noted, for all of our modeling experiments we utilize one full month as training data and the following month as test data, as highlighted in Section 4.

Table 2: Impact of publisher information on post-click conversion prediction accuracy (in terms of area under precision-recall curve

Model	AUC (PR)	Lift
Base (No pub info)	0.13432	–
Base + pub	0.14187	5.62%
Base + pub + pub nwrk	0.14239	6.01%
Base + pub + pub nwrk + site	0.14260	6.15%
Base + pub + pub nwrk + site + sect.	0.14298	6.45%

6.1 Relevance of Publisher-Side Information

We now provide an analysis of the relevance of publisher information in the task of post-click conversion prediction. Our motivation for this experiment was to understand whether knowing the publisher would improve prediction at all when the advertiser and the user are known. It seems plausible that once the user clicks on the ad, the publisher's page should not have an impact on that user's conversion behavior. If this were the case, then PCC models could be considerably simplified as no publisher-side information would be needed.

For this we built models that include no publisher information and then models that progressively add more detailed information from the publisher side. We considered the attributes: publisher, publisher network, publisher site, and publisher section. The results are shown in Table 2.

The results indicate that knowledge of the publisher identity benefits conversion predictive accuracy to some extent (5.62% improvement). Additional information helps but to a smaller degree. Thus, while the contribution is not large, the models utilized in the following sections do utilize publisher-side information.

6.2 Model Evaluation with Respect to Update Frequency

The results in Section 4.3 and 4.4 indicate that there is a non-trivial percentage of new ads entering the system every day. In this section, we first evaluate the actual impact of the new ads on post-click conversion modeling and then develop two solutions to address this problem. A single model, called the ID model due to its reliance on identifier attributes (Table 1) is trained using data collected for one entire month, with post-click conversion events as positive examples and clicks without conversions as negative examples (as before). These attributes have been converted into indicator features (one for each attribute-value pair) resulting in approximately $368K$ features. The data collected from the following month is used as test data. The data utilized contained two types of ad campaigns: cost per action (CPA) and dynamic cost-per-mille (dCPM) campaigns. The performance of the model is evaluated using the AUC (area under the receiver operating characteristic or ROC) curve.

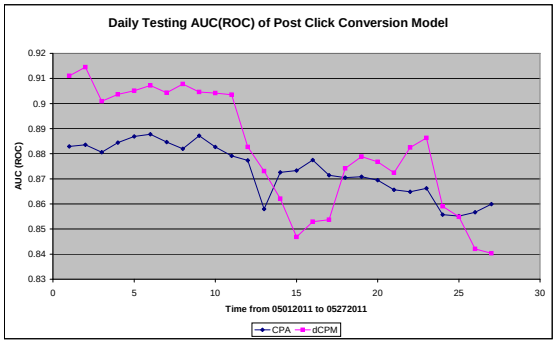

Figure 9: The performance (AUC) of the single ID model degrades with time for both CPA and dCPM campaigns.

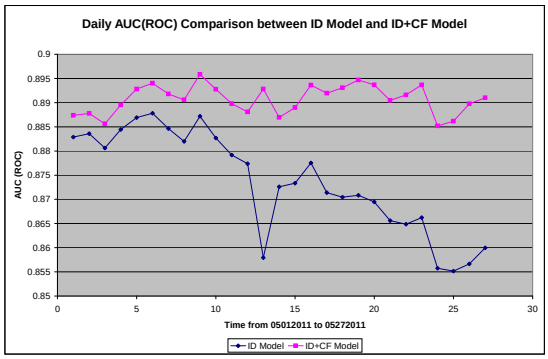

Figure 10: The Performance (AUC) of the ID+CF model is considerably more stable than the that of the ID model.

In order to evaluate the impact of the new ads on the performance of the single model, we divided the test data into separate slices. Each slice of the data represents one day test data of each campaign type. In Figure 9 we plot the AUC calculated based on different slices of the test data using the learned model. The x-axis represents each day in the test data and the y-axis indicates the AUC for each slice of data. We can see from Figure 9 that the performance of the model degrades almost linearly with respect to time for both CPA and dCPM campaigns. It seems clear that the degrading performance is closely related to the new ads emerging in the test data, as elucidated in previous analyses (Sections 4.3-4.4).

We experimented with one way to address the problem of performance degradation above (when using a single model for the entire test period). For this, we use the feedback features introduced in Section 6 and refresh the statistics captured by these features regularly to include information about new ads (on a daily basis). Feedback features capture the current performance of different ads, publishers and users by updating the number of clicks and conversion rates to the reflect their most up-to-date rates. A second approach (not tested) would be to adapt the model parameters to the latest data. In this second approach, the model is simply retrained regularly using Table 1 features, thus including the emerging advertisement-publisher information directly. Note that the first method is considerably less expensive because there is no need to update (re-train) the conversion model *per se*, but only to update the feedback statistics.

Figure 10 shows the performance of a model that utilizes feedback features refreshed every day (ID+CF model) compared with the single ID model. The feedback features included are the number of clicks and the conversion rates for each campaign, conversion identifier, creative, and publisher. In addition we compute the click and conversion rates for every pair of (publisher, creative), (publisher, campaign), (creative, conversion id), (user age, creative), (user age, conversion id), (user gender, creative), (user gender, conversion id), (geo, creative), (geo, campaign) and the triplet (publisher, creative, conversion id) . After quantization, there are approximately $1K$ new features added and refreshed every day, which is quite small compared to the size of the ID features above. From 10, we can see that the ID+CF model clearly outperforms the ID model for every daily slice of testing data. More importantly, the performance of ID+CF is quite stable compared with the ID model, which degrades linearly along time. Thus, we conclude that adding feedback features improves the performance of the post click conversion model and refreshing the features with sufficient regularity (*i.e.,* daily) will maintain these improvements.

6.3 Automated Feature Selection

We utilized the method described in Section 5.3 for determining feature relevance. Our main motivations were not only increasing automation in cases were the underlying data distribution is non-stationary but also decreasing model complexity as the available number of possible features is too large to be used in their entirety during prediction/modeling. Practical considerations, such as memory, latency, and training time constraints, make attribute selection a clear requirement in this task. Non-informative attributes can introduce noise and reduce the predictive accuracy of the system.

We used events for a period of 5 days (training), 1 day (reference), and 2 days (testing). The statistics of this data set were: 125M events for training, 25M for reference, and 50M for testing. The conversion rate varies considerable for different advertisers/publishers/users but rarely exceeds 5%; thus the data set is to a large extent unbalanced.

Our goal is to identify predictive features in the most automated manner possible (reducing time spent by people on this task). Thus, practically all the raw (unprocessed) data attributes available were included in the analysis. These attributes are a super-set of those in Table 1 and include identifiers for the actual (serve/click/conversion) event, advertiser, publisher, campaign, bcookies, timestamps, advertiser/publisher specific attributes, related urls, demographics, user-specific attributes (identifiers, assigned segments), etc. We consider conjunctions of any of these attributes, giving rise to about 5000 possible compound features in practice. Each feature in turn can take from two to millions of possible values.

The important element to consider is that without time-consuming research into attribute definitions, it is extremely tedious to apply most machine learning or data mining/ analysis algorithms. Thus, requiring considerable effort from *e.g.,* machine learning scientists or domain experts. Wrapper methods are not appropriate in this setting as they require training using a large set of variables; this is usually impractical except for some simple models. It is in this set-

Table 3: Top features for conversion prediction along with their mutual information. Top: standard mutual information; Middle and Bottom: modified mutual information. Bottom table contains the top conjunction features.

Single feature	SMI (bits)
event_guid	0.03102
receive_time	0.03059
query_string	0.02963
xcookie	0.02925
user_identifier	0.02923
Single feature	**RMI (bits)**
conversion_id	0.02338
campaign_id	0.02207
ad_group_id	0.02136
line_item_id	0.02090
advertiser_id	0.02082
Conjunction feature	**RMI (bits)**
conversion_id x offer_type_id	0.02379
conversion_id x pop_type_id	0.02369
conversion_id x campaign_id	0.02347
campaign_id x offer_type_id	0.02267
campaign_id x bid_type	0.02235
advertiser_id x pop_type_id	0.02120
advertiser_id x offer_type_id	0.02108
advertiser_network_id x offer_type_id	0.00993
publisher_network_id x advertiser_network_id	0.00770

ting where filter methods, such as the MI methods described here, can be more advantageous.

6.3.1 *Attribute Selection Results*

We applied the standard MI (SMI) ranking algorithm for feature selection. The results, summarized in Table 3(top) reflect our main concern. In the presence of spurious features, or features that are informative about the data point *per se* rank substantially high. The calculated MI score is correct in that it reflects the information content of these features; however, these features are too specific to the training data distribution.

The proposed extension of the MI score utilizing a reference distribution (RMI) provides a more appropriate ranking as shown in Tables 3(mid-bottom). The reason for this is that the information content is calculated with respect to (expectations on) the reference distribution and thus feature values that are not seen in the new distribution are basically considered less important and their impact on the information score is reduced.

More specifically, attributes such as `event_guid` that identifies the data point have maximal information content according to the training distribution (SMI), but near zero information content when calculated with a reference distribution (RMI) (*c.f.*; Section 5.3). A similar effect was observed for other features that have low relevance for prediction such as `query_string` and `receive_time` which unless parsed are too specific, `xcookie` and `user_identifier` which clearly do not generalize across users (but could be quite informative about a small fraction of the test data), and `user_segments` which is encoded as a string with a list of segments. The results for other features are more subtle but follow the same underlying principle where a reference distribution is utilized to avoid spurious dependencies often found when utilizing empirical distributions.

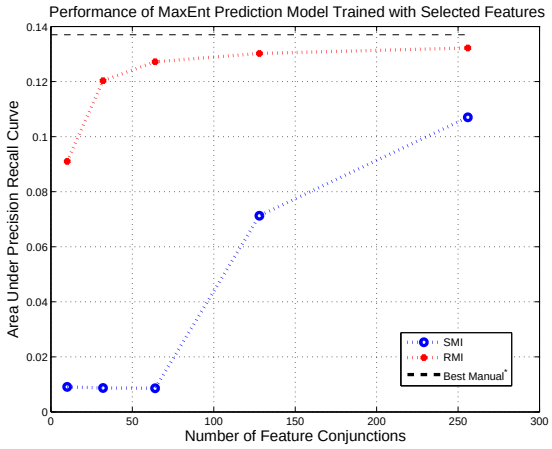

Figure 11: Performance of Logistic Regression Models trained with the top K features given by SMI and RMI

6.3.2 *Learning Performance Results*

Here we explore the question of whether the new feature rankings actually offer any performance gains. For this, we ranked all the feature conjunctions provided by SMI/RMI and trained a Maximum Entropy (Logistic Regression) model using the top K conjunctions given by each method. The results for conversion prediction are shown in Fig. 11 for various values of K.

The graph shows performance in terms of Area Under the Precision-Recall curve (similar results were obtained when measuring the Area Under the ROC curve and percentage of correct predictions). The results indicate that SMI ranks very irrelevant features on top. After more than 100 conjunctions, more relevant features start to be included in the model as evidenced by the performance gains. On the contrary, RMI captures relevant features much earlier, as only the top 10 features are sufficient to perform better than SMI with 128 features. The performance increases considerably at 32 features, and appears to stabilize after 64 features. One interesting question is whether after a large K the performance of both methods will be comparable. As seen at the end of both curves, when 256 features are utilized, still the difference is considerable. We believe the performance of SMI will remain affected by the various irrelevant feature included early on, depending on how much the model is susceptible to the noise introduced by these. Finally, the top line in the graph represents the current best model[5], using the features described in Table 1, after considerable feature engineering (*e.g.*, where features such as `receive_time` and `query_string` have been parsed and transformed into a suitable quantized representation), manual feature selection, grouping, and train/test exploration experiments. Thus, the proposed selection method produced quantitatively comparable results but is much more resource-efficient.

7. CONCLUSIONS

The accurate estimation of click and conversion probabilities is critical for the efficiency of on-line advertisement exchanges. While click-trough rate (CTR) analyses have been

[5]for the train/test period utilized in this experiment

the subject of much attention, post-click conversion (PCC) analysis studies are much more limited in scope and data coverage. As advertisers gain interest in more direct measures of profitability such as conversion rates, PCC becomes more important.

This paper provided a detailed analysis of conversion rates in the context of non-guaranteed delivery (NGD) for display advertising. We provided fundamental properties of the PCC process based on contextual information including a comparison between CTR and CVR for some relevant data attributes, an analysis of click-to-conversion delay, various properties about how rapidly new advertisements appear and stay in the system. We formalized the problem of conversion modeling along with the problem of determining attribute relevance in the specific setting of CVR optimization where the underlying data distribution is non-stationary. We also provided PCC modeling experimental results including measuring relevance of publisher information, measuring the effect of model update frequency (relevant for evolving attribute distributions), and the effect of automated feature analysis in this scenario. All this was done using data from a large-scale advertising platform. We believe this can provide a more thorough understanding of the PCC process and its modeling challenges.

Acknowledgments

We thank our colleagues for their assistance with data collection, model evaluation, and discussions on the RMI idea.

8. REFERENCES

[1] D. Agarwal, R. Agrawal, R. Khanna, and N. Kota. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. In *KDD*, 2010.

[2] A. Ashkan, C. L. A. Clarke, E. Agichtein, and Q. Guo. Estimating ad clickthrough rate through query intent analysis. In *WI-IAT*, Washington, DC, USA, 2009.

[3] A. Bagherjeiran, A. O. Hatch, and A. Ratnaparkhi. Ranking for the conversion funnel. In *SIGIR*, 2010.

[4] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. What happens after an ad click?: quantifying the impact of landing pages in web advertising. In *CIKM*, 2009.

[5] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] A. Broder and V. Josifovski. Introduction to Computational Advertising, 2010.

[8] D. Chakrabarti, D. Agarwal, and V. Josifovski. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, CA, 1967. University of California Press.

[9] D. Chakrabarti, D. Agarwal, and V. Josifovski. Contextual advertising by combining relevance with click feedback. In *WWW*, 2008.

[10] Y. Chen, D. Pavlov, M. Kapralov, and J. F. Canny. Factor modeling for advertisement targeting. In *NIPS*, 2009.

[11] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *WWW*, 2008.

[12] T. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Interscience, New York, NY, USA, 1991.

[13] D. Cox and E. Snell. *The analysis of binary data*. Chapman and Hall, 1970.

[14] N. Gupta, U. Khurana, T. Lee, and S. Nawathe. Optimizing display advertisements based on historic user trails. In *SIGIR Workshop on Internet Advertising*, 2011.

[15] I. Guyon and A. Elisseeff, editors. *JMLR Special Issue on Variable and Feature Selection*. Journal of Machine Learning Research, 2003.

[16] D. Hillard, E. Manavoglu, H. Raghavan, C. Leggetter, E. Cantú-Paz, and R. Iyer. The sum of its parts: reducing sparsity in click estimation with query segments. *Information Retrieval*, pages 1–22, 2011.

[17] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, 2005.

[18] S. Kullback. *Information Theory and Statistics*. Dover, 1968.

[19] R. A. Lewis. Advertising especially influences older users: A yahoo! experiment measuring retail sales. In *Measuring the effects of online advertising on human behavior using natural and field experiments*. Massachusetts Institute of Technology, 2010.

[20] R. A. Lewis. Where is the Wear-Out? Online Display Ads and the Impact of Frequency. In *Measuring the effects of online advertising on human behavior using natural and field experiments*. Massachusetts Institute of Technology, 2010.

[21] M. Mahdian and K. Tomak. Pay-per-action model for online advertising. In *ADKDD*, 2007.

[22] R. McAfee. The design of advertising exchanges. *Review of Industrial Organization*, pages 1–17, 2011.

[23] T. Minka. A comparison of numerical optimizers for logistic regression. http://research.microsoft.com/en-us/um/people/minka/papers/logreg.

[24] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy-channel approach to contextual advertising. In *ADKDD*, 2007.

[25] M. Regelson and D. C. Fain. Predicting click-through rate using keyword clusters. In *In Electronic Commerce (EC)*. ACM, 2007.

[26] D. Reiley and R. A. Lewis. Does Retail Advertising Work? Measuring the Effects of Advertising on Sales via a Controlled Experiment on Yahoo! . Technical report, Yahoo! Labs, 2011.

[27] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW*, 2007.

[28] X. Shi, K. Chang, V. Narayanan, V. Josifovski, and A. Smola. A compression framework for generating user profiles. In *SIGIR Workshop on Feature Generation and Selection for Information Retrieval*, 2010.

[29] W. Xu, E. Manavoglu, and E. Cantú-Paz. Temporal click model for sponsored search. In *SIGIR*, 2010.