Towards a Robust Modeling of Temporal Interest Change Patterns for Behavioral Targeting

Mohamed Aly†, Sandeep Pandey‡, Vanja Josifovski♯, Kunal Punera♮
† Seeloz Inc., Santa Clara, CA, USA
‡ Twitter, 1355 Market St, San Francisco, CA 94103
♯ Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043
♮ RelateIQ, Palo Alto, CA, USA
aly@seeloz.com, spandey@twitter.com, vanjaj@google.com, kunal.punera@utexas.edu*

ABSTRACT

Modern web-scale behavioral targeting platforms leverage historical activity of billions of users to predict user interests and inclinations, and consequently future activities. Future activities of particular interest involve purchases or transactions, and are referred to as *conversions*. Unlike ad-clicks, conversions directly translate to advertiser's revenue, and thus provide a very concrete metric for return on advertising investment. A typical behavioral targeting system faces two main challenges: the web-scale amounts of user histories to process on a daily basis, and the relative sparsity of conversions (compared to clicks in a traditional setting). These challenges call for generation of effective and efficient user profiles. Most existing works use the historical intensity of a user's interest in various topics to model future interest. In this paper we explore how the *change* in user behavior can be used to predict future actions and show how it complements the traditional models of decaying interest and action recency to build a complete picture about the user interests and better predict conversions. Our evaluation over a real-world set of campaigns indicates that the combination of change of interest, decaying intensity, and action recency helps in: 1) scoring significant improvements in optimizing for conversions over traditional baselines, 2) substantially improving the targeting efficiency for campaigns with highly sparse conversions, and 3) highly reducing the overall history sizes used in targeting. Furthermore, our techniques have been deployed to production and scored a substantial improvement in targeting performance while imposing a negligible overhead in terms of overall platform running time.

Categories and Subject Descriptors: H.4.m [Information Systems]: Miscellaneous

Keywords: Display Advertising; Behavioral Targeting; User Modeling; Time-based Features

1. INTRODUCTION

One of the key goals of the display advertising targeting is to identify the relevant audience for a given advertising campaign. This *audience selection* problem has been at the core of advertising from its beginnings in the 19^{th} century to modern day Internet campaigns. Traditionally, audiences have been selected based on user demographics and other attributes. Such attributes have been associated with different advertising channels by surveys. For example fashion magazines contain women related advertising as the majority of the population that consumes those magazines are women. In the beginnings of the online advertising, such demographics or location based marketing was also the mainstream trend. However, the Web setting provides means to track user behavior in greater detail than the offline setting, including capturing user's search queries, page views, clicks on ads and purchases.

This technology brings the ability to target users based on their individual actions, rather than general buckets of demographics or behavior. In addition, the recent trend in behavioral targeting and display advertising has been for the advertisers to instrument their web sites with embedded code to allow third parties to capture the user transactions on their sites. These transactions are commonly called *conversions* and can be used to obtain a sample of the general population that is receptive to a given campaign. The historical activity of users in such a *seed set* allows for modeling response prediction techniques to be applied to learn behavioral patterns that are indicative of the interest in a given campaign. A few recent studies have reported work in such settings [19, 2, 24, 3, 1].

In a real world setting, the task of predicting conversions is made difficult by two, somewhat contradicting, factors. First, the size of the seed set is usually very small compared to the general population. Campaign seed sets can vary from a handful of examples to a few thousands. Generalization to a several billions sized population from such a small set of examples is a difficult task. Second, the overall volume of activity data that comes into the system daily is extremely large composing multiple tera-bytes. These two factors require that we pre-process the raw data before modeling the response prediction. The raw data is first refined and condensed into *user profiles* that are effective and efficient for predicting the conversions. The efficiency requires profiles that are manageable in size, while effectiveness is determined by how much profiles are predictive of conversions.

In this paper we explore different ways to compose user profiles from raw user data. We focus on determining how the temporal patterns of different types of activity impact the predictive power of the profiles. Existing approaches use the *frequency* or *intensity*, and *decay* or *recency* of the activity to determine the weight of the feature. For example,

*All authors contributed to this work while affiliated with Yahoo! Research

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

if a user in general often visits a page that describes scuba diving, he might be interested in diving related products. Furthermore, if the user has recently visited such pages, he would be more likely to engage with such ads. Some forms of such weighting schemes have been showing to be effective in prior work [7]. In this paper we revisit the methods for such feature weighting in a more systematic setting exploring multiple options as decay function, length of user history, etc. to determine which settings do reflect the user interests in the most effective way.

Next, we explore the impact of the *change* in user behavior as indication of interests in a campaign. The intuition here is that many campaigns promote goods and services in which the consumer has an occasional interest. For example when going for a vacation in California, the user might explore different attractions and hotels before the visit. This online search activity will start at certain time before the vacation and seize thereafter. The right time for the user to be exposed to a campaign that suggests tours of San Francisco would be when the change in behavior is detected: the user did not have much activity in this topic before the recent spike. We explore multiple ways to generate profile features based on the *change of interest*, including short-term and long-term changes in interests, and their power to improve the prediction of conversion for display advertising campaigns.

Contributions:. Our contributions in this paper are as follows:

- We concretely formulate decayed-intensity and recency features and present a variety of interest change features to capture temporal patterns in user interests.
- In light of the behavioral targeting platform presented in [3], we empirically show that our temporal features help in improving the targeting performance across a variety of advertising campaigns, both in terms of prediction accuracy of learned models, and in terms of the degree of correlation between features and campaigns.
- We show the important effect of our temporal features in highly reducing the user history lengths and number of features used in targeting. This increases the ability of the behavioral targeting platform to continuously process web-scale user histories in an efficient and periodic fashion.
- Finally, we show the high ability of our temporal features in improving targeting for campaigns with rare conversions.
- In addition to our thorough empirical evaluation, our techniques were deployed to production and showed tangible gains in targeting performance (in terms of online metrics such as eCPA) without imposing any running time overhead on the underlying platform.

Organization of the paper:. This paper is organized as follows. We present our approach in complete detail in Section 2. We begin the section by giving background on the underlying behavioral targeting setting and establishing the terminology and notation used throughout the paper. We present the frequency-based feature weighting technique of the underlying platform in Section 2.4. Then in Sections 2.5 and 2.6 we concretely formulate the different temporal features, including both recency features and interest change features, as well as the rationale behind each of them. In Section 3 we present the results of our empirical evaluation of the impact of these temporal features in a real-world web-scale behavioral targeting setting. We postpone the discussion of related works to Section 4 as it gives us a better opportunity to compare them to our approach. Finally, Section 5 concludes the paper and discusses future research directions.

2. PROPOSED APPROACH

As mentioned before many past works have studied the use of user interest in topics for the behavioral targeting of display ads. In this paper we seek to study the impact that change in user interests can have on targeting accuracy. In order to do so, in this section we define various measures of change of user interest that we will then evaluate in our large scale experiments. We will first define measures that capture the *recency* of users' interests, and then proceed to constructing various ways to model the *change* in these recent interests. However, we begin this section by giving background on display advertising platforms and setting up the terminology and notation used in our description of our proposed approach.

2.1 Background on Display Advertising

In this paper we focus on display ads as graphical ads are presented on the web pages of publishers, both large and small. From its meager beginnings in the 1990's, display advertising has grown to an estimated $12.33 billion industry in 2011, according to emarketer.com. In terms of classical advertising objectives, display advertising covers both *brand* advertising where the advertisers aim to improve users' impression about a product or a brand; and direct advertising (or *performance advertising*), where the ad is intended to incite the user to click on the ad, which will subsequently redirect the user to an advertiser specified *landing page*. Performance advertisers seek a specific action from the user as a result of the advertisement, be it buying a product or service, signing up for an email list, or filling out a form. These advertiser-defined events are called *conversions*.

Traditionally, the display advertising effectiveness is measured using either click-through rate $\frac{\#\text{clicks}}{\#\text{ads shown}}$ or conversion rate $\frac{\#\text{conversions}}{\#\text{ads shown}}$. Publishers such as Google and Yahoo! therefore seek to maximize the click-through rate and the conversion rate of the advertisements shown on its webpages by showing their ads to users who are more likely to click or convert on the advertisement. Identifying such users is based on features extracted from users' past online activities such as page visits, search queries, ad clicks, etc. In this paper we focus on using the co-bidding information to improve the response prediction for a given advertiser and a given ad.

2.2 User Profile Basics

We begin with a brief overview of the underlying behavioral targeting platform (presented in [3]). The platform optimizes various existing campaigns where the advertisers pay per conversion (or action), commonly known as *cost-per-action* (CPA) campaigns. Each campaign is first tuned manually by using traditional demographic and behavioral targeting. The system objective is to refine the targeting

constraints using the user behavioral actions to maximize the number of conversions per ad impression without greatly increasing the number of impressions, which increases the value of our inventory.

We assume that the events prior to the conversion contain an indication of its occurrence and we do not use any events past the conversion event in our prediction framework. As shown in Figure 1, we consider user history as a sequence of events relative to some *target time*, τ, at which time the user is a candidate for targeting. We decompose the user's sequence of events around the target time τ as follows:

$$\begin{aligned} u(\tau) &= (E_F(\tau), E_T(\tau)) \\ \text{where } E_F(\tau) &= \{e \mid e \in E \wedge t(e) < \tau\} \\ \text{and } E_T(\tau) &= \{e \mid e \in E \wedge \tau \leq t(e) \leq \tau + \delta\} \end{aligned}$$

Here $E_F(\tau)$ denotes the events prior to the target time which we call the *feature window* and $E_T(\tau)$ denotes the events that occur between τ and $\tau + \delta$ which we call the *target window* and is set to be of length of a range of hours to a day, whereas $t(e)$ denotes the occurrence time of event e.

Hence, we model behavioral targeting as a machine learning task where each user history is converted into a training example: a user-history is a positive example if it is associated to a conversion in the target window, a negative example otherwise. [1] A user can only be a positive or a negative example, but not both at the same time. Given training users $\{1, \ldots, m\}$ define $T = \langle (x^1, y^1), ..(x^m, y^m) \rangle \in (\mathbb{R}^n \times \{-1, +1\})^m$ to be the training data, where x^i is a feature vector constructed from the events of the user i in the feature window, $y^i = +1$ if the ith example is positive, and $y^i = -1$ otherwise. The test set is defined similarly. Using this problem definition, our platform periodically train per-campaign models using a linear Support Vector Machine (SVM) algorithm.

2.3 User Profile Representation

A user representation method consists of a function $\phi : U \to \mathbb{R}^n$, where $\mathbb{R}^n$ is the Euclidean space of dimension n. A target label function is defined as $\gamma : U \to \{-1, +1\}$. Given this, we extract an appropriate feature and target label set as $(x, y) = (\phi(u), \gamma(u))$. We then select a vector $w \in \mathbb{R}^n$ by solving the following optimization:

$$\arg \min_{w \in R^n} w^2 + C \sum_{i=1}^{m} L(w \cdot x^i)$$

where $L(\hat{y}, y) = \max(1 - \hat{y}y, 0)$ and C is a constant that controls the balance between regularization and minimizing the loss on the training set.

In this work, we investigate different user representation operators $\phi(u)$, basically when converting events to features, through leveraging our temporal feature extraction techniques in computing different weights for each feature.

2.4 Baseline Feature Weighting

We now describe the feature weighting in our baseline system (described in great details in [3]). For all feature types, and for all models, our baseline uses a common feature weighting operator $\phi : E^\star \to \mathbb{R}^n$ which we call relative frequency bag of events (or simply we denote it as the *frequency*). The frequency of an event is defined as the number of days in which the user has performed the event. Events of different types are considered separately, for example, page visits, search queries, etc. We normalize all the values of all features that comprise a feature type such that their L2 norm is 1. This ensures that no one type of feature dominates the feature vector associated with a user's history.

To clarify we consider an example. Consider the feature type *page visits* and denote by event p the "visit to any sports page"; q and r denote visits to pages on other topics. If the user had visited the pages p, q, and r in a sequence over four days as follows:

$$e = (p_1, p_1, p_2, q_2, r_3, q_4)$$

where each event p_i denotes cluster id of the the visit on page p on day i, then the frequency bag of events representation for page visits would be: $F_p(e) = (p : 2, q : 2, r : 1)$ where we use the convention of $p : n$ to mean that the feature p has the value n in the feature vector. Here, the cluster id of page p was visited 3 times but on just 2 distinct dates.

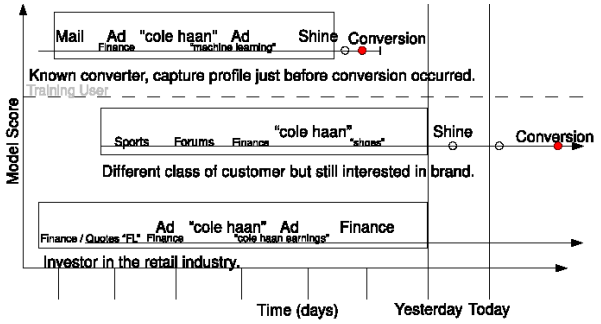

Figure 1: Targeting model is trained on user histories (rectangles) as they existed prior to the start of the conversion process (open circle) that led to the conversion (solid circle). For evaluation, all users are given the same target time (yesterday) and the ad server may choose to show ads at some point in the future to start the conversion process (open circle).

2.5 Features: Recency of User Interests

It has been established by past works that many data mining approaches benefit from exploiting the timing of events along with the data associated with them. For instance, Koren [15] use recency to improve the performance of recommendation systems. Here, we study whether recency of user interest can lend a useful signal for behavioral targeting of display ads.

In our application we intuitively expect that the recent user interests would be more useful than older user interests as a predictor of which advertisements the user is likely to engage with, or *convert on*. This might be due to many reasons. For example, a user's recent interest in "bmw cars" might be more useful for advertising than a user's older interests in "audi cars" because of changing choices, while a recent interest in "skiing equipment" might be more indicative than an older interest in "tennis rackets" because of changing seasons. Finally, recent interests might capture the progression of user's real-life activities – progression from searching

[1] If the user has been to multiple impressions of the same ad, the conversion is *attributed* to the last impression

for mortgage brokers to searching for home renovation companies – and hence might have higher signal for targeting advertisements.

In order to model the recency of user interests to capture the effect outlined above, we exploit the exponential decay formulation, which has been popular in the past in such frameworks. Under this framework, the intensity of a feature is modulated by how long ago it occurred, with the intensity decaying exponentially with increasing time.

In our system, the weight of the decayed intensity feature generated based on a given user behavioral feature p is computed as follows:

$$W_{intensity}(p) = \sum_i d^{(t_{p_i} - \tau))} \quad (1)$$

where, $W_{intensity}(p)$ is the weight of the decayed intensity feature of p, d is the decay factor, t_{p_i} is the date of occurrence i of feature x in the user history, and τ is the target date. Note that the exponent of d is always negative, thus the weight function will always be decaying when the gap between the feature and the target increases. Note that the unbounded sum in the above formula means the sum over all the user history available in a user profile.

While Eq 1 is a principled way to model recency in user interests, it has a deficiency that it relies on a parametric form and parameter values that might be ill-suited for certain users or advertising campaigns. Another way to capture recency of a user's interest in way that does not have this deficiency is through the absolute number of days since the last occurrence of a feature. For example, if a large number of days have passed since the last time a user searched for a "san francisco hotel rates" then it might indicate that though the user was interested in it, it might no longer be the case; the user might have already culminated the travel. In these cases a recency feature which measures the number of days since last activity associated with a feature might be useful.

Formally, this recency feature is computed as follows:

$$W_{recency}(p) = \tau - t_{p_n} \quad (2)$$

where $W_{recency}(p)$ is the weight of the recency feature of p, t_{p_n} is the date of the last occurrence of feature p in the user history.

2.6 Features: Change in User Interests

In Section 2.5, we described how we model the recency aspect of a feature. In many cases, however, it is not the decayed-intensity or last occurrence of feature that matters, but the change in user interests, or even the degree of change that are more important. As an example, consider a user who periodically searches for information on "bmw cars" but recently has ramped up her searches. Compare this user with another user whose activity stream contains constant frequent set of searches on "bmw cars". While both users might have the same number recent on-topic searches, the former one is more suitable for an advertisement from BMW. In this case the change in behavior was considered important after the recent intensities of the users matched. The reduction in feature intensity can also serve as a useful aspect of a feature for the purposes of advertisement targeting.

In order to capture these interest change effects we propose a measure that computes the difference in the activity related to a feature in the most recent *time period* to the activity further in the past. As is clear, to make the above definition concrete we have to fix the notion of a time period, the two time periods being compared, as well as how the activity levels in the time periods are compared.

Size of Time Periods:. The size of the time periods over and across which we compare user activity have to be set carefully so as to avoid any biases. For example, if the size of a time period is not a multiple of a week then care must be taken to deal with the weekend effect; online user behavior differs widely between weekdays and weekends. Moreover, too small a period might not be suitable because of data sparsity issues, while too large a period might smooth out the signal. In order to deal with these issues we compare various choices of size of time periods and report on the results.

Time Periods to be Compared:. In order to detect change the activity levels in the most recent time period we have to employ a notion of baseline user activity in another time period. In our experiments, we compare the user activity in the most recent time period to two different sources of baselines. In one case the baseline activity comes from the entire known user history; we label this as *long-term* change. In the second case the baseline time period is an equal sized period of time just before the most recent time period (this is referred to as *short-term* change).

Absolute vs Relative Change:. The degree of change can be measured in two different ways. In the first case we investigate the use of an absolute increase or decrease in the activity level from the baseline time period to the most recent time period. These features can be formally defined as follows:

- Long Term Absolute Interest Change

$$W_{abs-long-term}(p) = \sum_{p_i : t_{p_i} \in P_{-1}} - \sum_{p_i} \quad (3)$$

$$W_{abs-long-term}(p) = - \sum_{p_i : t_{p_i} \notin P_{-1}} \quad (4)$$

  where $W_{abs-long-term}(x)$ is the long-term component of the interest change feature, P_{-1} is the last period before the target, $\sum_{p_i : t_{p_i} \in P_{-1}}$ is the sum of the occurrences of feature p_i in the last period, and $\sum_{p_i}$ is the sum of occurrences of feature p throughout the user history (where variable i represents a given day i in which the feature p appeared and the sum is over all possible values of i). Note that in Equation 4, the definition of $W_{abs-long-term}(p)$ boils down to the negation of the sum of occurrences of p throughout the user history in case p never occurred in P_{-1}.

- Short Term Absolute Interest Change

$$W_{abs-short-term}(p) = \sum_{p_i : t_{p_i} \in P_{-1}} - \sum_{p_i : t_{p_i} \in P_{-2}} \quad (5)$$

  where P_{-1} and P_{-2} are the last and second-last time periods before the target, respectively. The first term in the Equation 5 is the sum of the occurrences of feature p in the last period, and the second term is the sum of occurrences of feature p in the second-last time period.

Unfortunately, this version of the feature addresses the amount of change but does not capture the change relative to

the baseline value. For example, under the absolute change feature a user who does switch from searching for "bmw cars" on average once a day to on average of 5 times a day is considered the same as a user who goes from searching for the same query from 20 times a day to 25 times. It seems intuitive that the user who has the largest relative jump is most suitable for advertisements from the BMW campaign. In order to test this intuition we encoded it in a feature that measures this relative jump. More formally,

- Long Term Relative Interest Change

$$\#\_periods = \frac{\#\_total\_days}{\#\_days\_in\_a\_period} \quad (6)$$

$$NSW_{P_{-1}}(p) = \frac{\sum_{p_i:t_{p_i}\in P_{-1}} +0.01}{\#\_days\_in\_period\_P_{-1}} \quad (7)$$

$$NSW_H(p) = \frac{\sum_{p_i} +(0.01 * \#\_periods)}{\#\_total\_days} \quad (8)$$

$$W_{rel-long-term-pos}(p) = \frac{NSW_P(p)}{NSW_H(p)} \quad (9)$$

$$W_{rel-long-term-neg}(p) = \frac{NSW_H(p)}{NSW_P(p)} \quad (10)$$

where P_{-1} is the last time period. The annotations relative and long-term are defined above; the positive and negative annotations come from the aspect of increase or decrease in user activity that the feature is attempting to capture. $NSW_H(x)$ is the sum of occurrences of feature x throughout the user history, $NSW_{P_{-1}}(x)$ is the normalized smoothed weight of x for period P_{-1}, while $NSW_H(x)$ is the normalized smoothed weight of x for the whole user history, $\sum_{p_i:t_{p_i}\in P_{-1}}$ is the sum of the occurrences of feature p in the last period, and $\sum_{p_i}$ is the sum of occurrences of feature p throughout the user history (where variable i represents a given day i in which the feature p appeared and the sum is over all possible values of i).

- Short Term Relative Interest Change

$$NSW_{P_{-2}}(p) = \frac{\sum_{p_i:t_{p_i}\in P_{-2}} +0.01}{\#days\_in\_period\_P_{-2}} \quad (11)$$

$$W_{rel-short-term-pos}(p) = \frac{NSW_{P_{-1}}(p)}{NSW_{P_{-2}}(p)} \quad (12)$$

$$W_{rel-short-term-neg}(p) = \frac{NSW_{P_{-2}}(p)}{NSW_{P_{-1}}(p)} \quad (13)$$

As above, here P_{-k} refers to the k^{th} time period before target date, and the positive and negative annotations refer to the increase and decrease in the user activity that is being emphasized in the feature. $W_{shortterm}(x)$ is the short term component of the interest change feature, $length(P_1)$ is length of the last period before the target, $length(P_2)$ is length of the last period before the target, the $S_P(x)$ is the sum of the occurrences of feature x in the last period, $S_H(x)$ is the sum of occurrences of feature x throughout the user history, $NSW_P(x)$ is the normalized smoothed weight of x for period p, while $NSW_H(x)$ is the normalized smoothed weight of$x$ for the whole user history. $\sum_{p_i:t_{p_i}\in P_{-2}}$ is the sum of occurrences of feature p in the before last period, P_2.

# days	# users	# features	# campaigns	dataset size
56	5.2B	834K	200	252GB

Table 1: Basic statistics of the data used.

Note that the period length is ideally the same across P_{-1} and P_{-2}, but in case some user has a shorter history, we account for that by using means like (number of activity per day) as opposed to raw activity counts. Also, note the Laplace smoothing and the normalization added to all four definitions of relative interest change.

3. EMPIRICAL ANALYSIS

In this section we will present the empirical evaluation of the temporal features we proposed in Section 2. We will first describe the experimental methodology and then present results on impact of these features on the performance of behavioral targeting. We will then study the variation in performance due to various parameters. Finally, we will present anecdotal evidence to show that the patterns learned by our features and empirical analysis make intuitive sense.

3.1 Evaluation Methodology

3.1.1 Data

To model the performance of our temporal feature generation techniques, we build targeting models for display advertising campaigns based on the targeting system for conversion-optimization presented in [19, 24, 3]. Table 1 presents some statistics about our data set. We collected 4 weeks of advertising data (i.e., impressions, clicks, and conversions) for 200 campaigns. We note here that users that opt-out of behavioral targeting are not profiled. Each campaign is treated as a separate targeting task. 66% of the data is used for training, while the remaining 34% is used for scoring. The train/test split is performed in a random fashion as described in [3]. As our user profiles span 56 days of user history, each training/scoring example is preceded by at least 4 weeks of user events. This benchmark data set enables us to do rigorous offline experiments. We count users based on the unique number of browser cookies.

We study the performance of our techniques compared to the baseline system developed in [3] [2]. We mainly compare modeling performance in terms of the area under the ROC curve (AUC). Unless otherwise specified, all metrics are measured as conversion-weighted average of AUC across all campaigns in the benchmark set. For simplicity, we denote the conversion-weighted average of AUC as *Weighted AUC*.

3.1.2 Behavioral Activity Types

We now describe the main types of user activities in the user profiles we use. When collecting the user's historical online behavior, we consider both *active* and *passive* activities. Passive activities include viewing ads and visiting pages in which an action is not specifically required upon seeing the page. Active activities include issuing search queries and clicking ads in which users actually perform an action on the page. Browsing activity is somewhat active because the user

[2]This system has shown to be outperforming traditional behavioral targeting techniques in terms of both scale and targeting accuracy

took action to visit the page, but we argue that the activity is less important than specifically typing a search query or clicking on an ad. Previously, the authors of [19, 24] showed that the collection of active and passive user activities is stronger in predicting the user's propensity to convert on a set of advertisements than any type separately.

The activities of the users are a sequence of events collected from server logs. Events are associated with both a timestamp and metadata. For example, an event could be a visit to a finance web page and the metadata associated with the event is the content of the page, which is logged separately and then joined with the event along with the anonymized identifier of the user and the time. We consider several different events, each with a corresponding feature extraction method. Our event types are namely pages visited, search queries (including the actual issued queries and the clicks on the query results), and interactions with graphical advertisements (including views, clicks, and conversions). All possible events belonging to a given type (across all users) are clustered using an existing type-specific hierarchical categorizer. When building the targeting profile of each user, the cluster ids, rather than the raw events, are used to represent the different user activities.

We define the *Profile Density* to be the total number of days of history spanned by a user profile. Figure 2 presents the profile density distribution of the user profiles used for generating our data set. The figure clearly shows that a large percentage of our user profiles have very short history lengths. Given the rarity of conversions for our campaigns, this consequently results in a relatively large number of examples with short history sizes. This further complicates the classification problem and highly increase the importance of strong feature engineering techniques in only keeping discriminative features and hence, improving the targeting accuracy.

3.1.3 Comparing Feature Weighting Techniques

We now describe the comparison technique we follow to compare between the different feature weighting techniques presented in this study (Sections 2.4, 2.5 and 2.6). For a given user behavioral feature, e.g., a page visit p, we generate a new independent feature p_T for every new feature engineering technique T applied. For example, in case we apply both Baseline and Decayed Intensity in a given experiment, we create two separate features, $p_{Baseline}$ and $p_{DecayedIntensity}$. Each of the two features are treated in a completely independent fashion by the later steps of our pipeline, e.g., the feature selection step and the training step. This feature engineering technique allows us to model the performance of applying our feature generation techniques individually or collectively, and hence understanding the exact effect of each of the feature generation techniques on the targeting performance for the different campaigns.

3.1.4 Classification Hyper-Parameters

Our system consists of the application of the linear SVM classifier on a per-campaign basis with ℓ_1 regularization. Given all user profiles, the system builds positive/negative training instances separately for each of the campaigns, then the classifier is run to produce the per-campaign models. Based on some simple experiments, we arrived at the following classification parameters. Because we have a large number of mostly irrelevant features, we choose a strong regularization parameter of $C \in [0.05, 0.7]$ (throughout this paper, we separately tune the C parameter for each of our experiments and show the best performing results in terms of the ROC measure). Next, since we have an highly imbalanced class distribution, we choose for the positive and negative classes for a given campaign a weighting parameter equal to the reverse of the number of examples from that class for that campaign.

3.1.5 Feature Selection

To reduce the dimensionality of our classification problems, we introduce a feature selection step on the feature vectors of all training examples before feeding them to the classifier. This step can be looked at as a pre-processing feature selection step given that we already use ℓ_1 regularization. Our feature selection is carried through considering the mutual information between features and labels and including discriminative features with a large mutual information with labels. We use the following score to evaluate the importance of a feature:

$$I(y, [x]_i) = H(y) - H(y|[x]_i) \tag{14}$$
$$= H(y) + \sum_{[x]_i} \left[\sum_{y} p(y|[x]_i) \log p(y|[x]_i) \right]$$

Whenever $y, [x]_i$ have finite (small) joint support, this can be approximated efficiently by using empirical probability estimates. We use the latter as a criterion to order features, on a per campaign basis, then take the top k features in the ordered list to be included in the feature white list. The final feature white list represents the union of the top-k lists for all campaigns modeled by the platform. It is worth mentioning that, in addition to feature selection, we use the mutual information score of the different features to assess and study the ability of the different feature engineering techniques (both the baseline and the ones developed in this paper) to improve the correlation between user activities and conversions.

3.2 Targeting Accuracy Results

We now describe our experimental results. We start by studying the overall targeting performance of each of our feature generation techniques across the different advertising campaigns. Then, we move on to a more thorough analysis of the results to understand the effect of collectively applying the different techniques, whether the different techniques are complementary or contradicting, and whether the techniques make an improvement for campaigns with sparse conversions.

3.2.1 Comparative Analysis of Temporal Features

We start by comparing the performance of the different feature engineering techniques when applied individually on our system. Table 2 presents the percentage AUC improvement over the baseline of the different techniques. Aside from Decayed Intensity, all the other features types are applied on top of the Baseline (as all these types are not intended to capture feature intensities, thus should not contradict with Baseline). Results show that Decayed Intensity, with ($d = 1.1$), is able to beat the Baseline by 1.58%. With that decay factor, a feature drops to around 50% of its intensity every week (i.e., the feature would completely fade in around 4 weeks). This shows the high importance

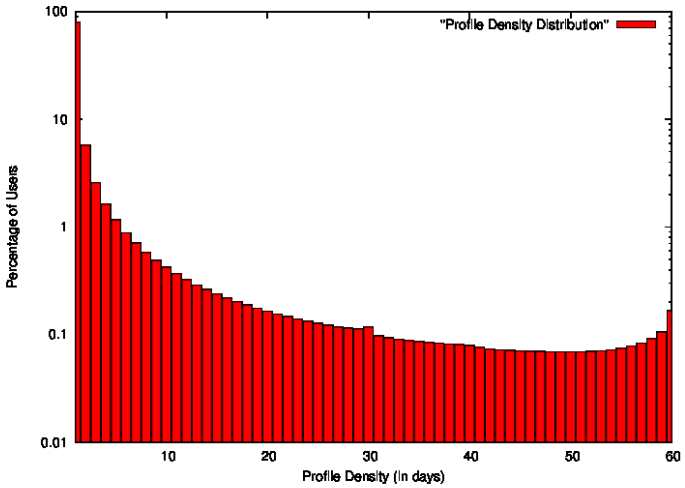

Figure 2: Profile Density Distribution

of shorter term user history when it comes to targeting accuracy. Along the same lines of features favoring recent behavioral actions, the introduction of Recency features to the current Baseline (i.e., to frequency features) further improves the performance with 1.86%. This also shows how the recency features act as complementary to the frequency features as the former favor recent user interests while the latter concentrate more on overall user interests.

Technique	ΔAUC
Baseline	0%
Decayed Intensity(d=1.1)	1.58%
Frequency + Recency	1.86%
Baseline+ Long Term Absolute Interest Change	2.14%
Baseline+ Short Term Absolute Interest Change	3%
Baseline+ Short Term Relative Interest Change	3.43%
Baseline+ Long Term Relative Interest Change	**4.44%**

Table 2: Modeling performance comparison for the different temporal feature engineering techniques. For all interest change techniques, $P = 28$ days.

Moving to the interest change features, we find that in general, the addition of these features on top of Baseline features score even larger performance improvements. This shows how modeling change in user interests can give a signal with relatively high importance, that again complements that given by our current Baseline. In general, absolute interest change features are less performing than relative ones. This can be explained by the high expected variance of these variables, for any given feature type, which relatively limits the classifier ability to draw strong correlations among these features. This problem is mostly solved with the relative features due to the addition of both the Laplace smoothing and the normalization. Among all the interest change features, the Long Term Relative Interest Change is the one that scores the most improvement on top of the current Baseline (namely 4.44% weighted AUC improvement). This can be explained by two reasons. The first is that this technique mainly compares the feature intensities in the last period with the total intensity of the feature. This makes this feature type less prone to fluctuations in behavior, and hence more stable than the Short Term Relative Interest Change that models two consecutive periods of user history. The second reason is that this feature is the one expected to have the lesser number of times where the Laplace smoothing is the actual significant term in the denominator of the negative interest change features. This would result in a smaller overall variance in the values of the negative interest change of any feature, which would increase the ability of the classifier to benefit from such features.

3.2.2 *Mixing Effect of Feature Engineering Techniques*

We now move on to study the effect of applying combinations of our feature engineering techniques. Table 3 shows the collective performance of different mixes of our features. We first try to mix the Decayed Intensity features with the Baseline. Then, we model the mix of the recency feature types, mainly Decayed Intensity and Recency, with the Baseline. For the first mix, targeting accuracy improves by 2.72%, while it improves by 3.43% for the second mix. This result shows the relative importance of each of these feature types. Though one could think that two feature types are redundant, each of the feature types is actually giving a different signal, and thus mixing them together achieves superior performance than applying each of them separately. We then try to mix interest change features with recency features. We first try the mix of absolute interest change features (both short term and long term) together with recency features (in addition to the regular frequency features of the Baseline). Finally, we apply the mix of relative interest change features with both recency and frequency features. The upper hand of relative interest change features shown earlier does not change much when applied with the other feature types and the final mix is the one that shows the superior performance over Baseline by scoring 4.72% improvement in weighted AUC. Note that we did not try further combinations of both absolute and relative interest change feature types to avoid any explosion in dimensionality in the different classification problems.

3.2.3 *Performance on Campaigns with Sparse Conversions*

As mentioned in Section 1, conversion sparsity is an important problem facing any behavioral targeting platform

Technique(s)	ΔAUC
Baseline+ Decayed Intensity	2.72%
Frequency + Recency+ Decayed Intensity	3.43%
Frequency + Recency+ Decayed Intensity+ Long Term Absolute Interest Change+ Short Term Absolute Interest Change	4.44%
Frequency + Recency+ Decayed Intensity+ Long Term Relative Interest Change+ Short Term Relative Interest Change	**4.72%**

Table 3: Modeling performance comparison (in terms of ΔAUC) for the different temporal feature engineering techniques. For Decayed Intensity, $d = 1.1$. For all interest change techniques, $P = 28$ days.

Technique	C<100	$C \in [100, 1000)$
Baseline	0%	0%
Baseline+ Long Term Relative Interest Change	2.58%	4.55%

Table 4: Modeling performance comparison (in terms of ΔAUC) for the different temporal feature engineering techniques on campaigns with the total number of conversions less than 100 and between 100 and 1000, respectively. For all interest change techniques, $P = 28$ days. All ΔAUC values are relative to Baseline

optimizing for conversions. It is interesting to assess the effect of our interest change features on the targeting accuracy when optimizing for campaigns with rare conversions. Table 4 shows a comparison of the performance of Baseline to that of the best performing interest change technique, namely Long Term Relative Interest Change, when applied together with Baseline features, on campaigns with total conversions less than 100 and in the range $[100, 1000)$, respectively. Results show the upper hand of the latter technique over the former by 2.58% and 4.55% for the two sets of campaigns, respectively. The improvements are considered significant, especially for the first set of campaigns, given the high complexity of the classification problems for these campaigns as a result of their extreme conversion sparsity. Additionally, this result has the important consequence of increasing the ability of the underlying behavioral targeting platform in leveraging our new interest change features to efficiently optimize for any advertising campaign, regardless of its conversion count, which highly increases the monetization ability of the platform. Note that applying more exhaustive combinations of temporal feature types did not score additional improvements over the mix of Baseline and Long Term Relative Interest Change, especially for campaigns with less than 100 conversions. This is mainly expected in light of the extremely low conversion counts of these campaigns, which does not allow the underlying classifier in efficiently handling high-dimensionality problems.

3.3 Analysis and Discussion of Results

We now move on to present a more thorough analysis of the performance results of the different feature types proposed in this paper. Our goal here is twofold. The first is to better understand the dynamics of our features and how they interact together when applied for different campaigns. The second is to deduce interesting observations that may improve our understanding of the relationship between the different aspects of user interests and the behavioral targeting classification problem.

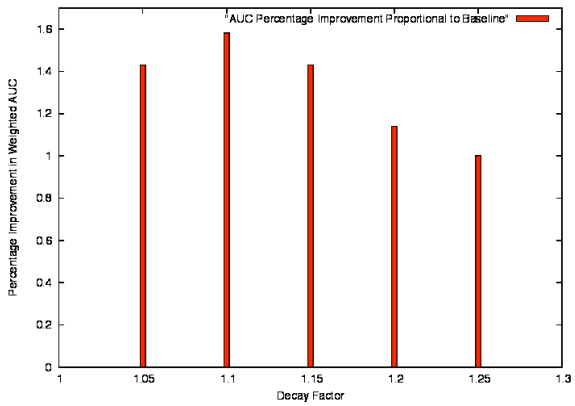

Figure 3: Effect of decay factor on performance of Decayed Intensitytechnique. All ΔAUC are relative to Baseline

3.3.1 *Sweet Spot in User History*

We start by addressing the question of the sweet spot in user history, which can be considered as the hottest and most important period in the user history when it comes to behavioral targeting. To achieve this goal, we run two studies. The first is an analysis of the effect of the decay factor on the performance of the Decayed Intensity features (Figure 3), while the second is the effect of the period length on the performance of the interest change features (Table 5).

In Figure 3, we ran our system only with Decayed Intensity features, while trying different values for the decay factor, d. It is worth mentioning that a value $d = 1$ simply represent our Baseline features. We explored the range of values $[1.02, 2]$ (but only present results for the range $[1.05, 1.25]$ due to its sufficiency as a synopsis of the results). As the Figure shows, performance starts to improve with increasing the value of d till reaching a peak at $d = 1.1$, then it starts to drop till reaching a value of 1% and 0% improvements at $d = 1.25$ and $d = 1.5$, respectively. As pointed earlier, for $d = 1.1$, a feature looses around half of its intensity every week. That is, a feature reaches its *half-life* after the second week as after two weeks, the feature intensity drops down to 1/4 of its original intensity, which in practice results in highly reducing the discriminative effect of the feature. In general, this result preaches for giving an extremely high importance to the most recent two weeks of the user history when it comes to feature intensities. The question is, how about the sweet spot for interest change?

Table 5 compares the performance of our relative interest change techniques, both short term and long term, for different period sizes. The result show a relatively large improvement when comparing $P = 7$ to $P = 14$, especially for the Short Term Relative Interest Change technique. However,

Technique	$P = 7$	$P = 14$	$P = 28$
Long Term Relative Interest Change	4%	4.29%	4.44%
Short Term Relative Interest Change	3%	3.72%	3.43%

Table 5: Effect of period size on performance of interest change technique. Values represent ΔAUC for the different techniques proportional to Baseline

improvements start to decrease with additional increase in the period size. Additionally, we note that an important reason for which the results continue to improve for $P = 28$ is that the increase in the period size results in a lesser effect of the Laplace smoothing, as a larger portion of the user history will actually have both short term and long term components. Again, this result shows the relatively high importance interest changes in the last two weeks of user history (just before the conversion date).

One important significance for these results is related to the web-scale user history to be processed by any behavioral targeting system. As mentioned in Section 1, the large amounts of user history to be continuously maintained and processed is a major challenge for a behavioral targeting system. As noted in [3], maintaining long periods of user history is extremely complicated and costly. This result calls for the need for full maintenance of the recent two weeks of user history. For the rest of the user activities, maintaining sketches representing different functions about the data would be much less costly and as beneficial in terms of targeting accuracy. This is considered as a highly important achievement from the behavioral targeting point of view.

3.3.2 Performance Variance based on Campaign Type

We analyzed the performance of our various temporal feature engineering techniques on different campaigns. As we have explained before in Section 2, we expect different approach to perform well for different types of products being advertised. We observe these in the results (Table 6) as well. As we can see, campaigns related to Travel tend to improve steadily as we add in additional temporal features into our models. This makes sense since travel needs manifest themselves at some specific points in time, and hence exploiting time as features should be useful. Our experiments show that Autos related campaigns exhibit a very large increase in performance when change in behavior is modeled. This makes sense as well since auto purchases are typically long terms affairs with a lot of nascent search/browsing accompanied by a burst in activity just before the purchase times. Finally, we observe that campaigns for fast moving consumer goods like Cosmetics exhibit negative correlation with our recency features. These are items that a user is ready to buy anytime and, as we observe, using recency based features to discard older data actually reduces the performance.

3.3.3 Short Term vs. Long Term Interest Change

One important question consists of understanding the relative value of our different feature types and how they interact with each other, especially for campaigns where the mix of all feature types performs the best. Hence, we present in Table 7 a break-down of the percentages of the different feature types for three advertising campaigns, mainly travel, autos, and cameras. One could see that change of interest is a key for each of the three campaigns, but the question is, what type of change of interest, short term or long term? positive or negative? The results shown in the Table show that the positive change in long term relative interest is, by far, the most important and discriminant feature type among all the interest change feature type. This in fact makes perfect sense, as for each of the three actions, whether searching for a camera, trying to book a travel, or searching for an auto, the user would be starting to do new behavioral actions that she was not used to before. The relative importance of the rest of interest change features is somehow similar for the three campaigns. The important observation for the three campaigns is that, in addition to Positive Long Term Relative Interest Change, Recency features are very important (in fact with a little bit of a higher value). This is clear in the fact that, a recent keyword search for the term "camera" or a recent visit for "orbitz.com" would give clear indication that the user is developing an interest in cameras and travel, respectively. Aside from these two observations, results show a low importance for the frequency features and a moderate to low importance for the Decayed Intensity features.

Technique	Travel	Autos	Cosmetics
Baseline	0%	0%	0%
Decayed Intensity($d = 1.1$)	0.5%	0.47%	-1.9%
Long Term Relative Interest Change	3.79%	14.1%	-1.9%
Frequency + Recency+ Decayed Intensity+ Long Term Relative Interest Change+ Short Term Relative Interest Change	4.09%	14.1%	-4.67

Table 6: Modeling performance comparison for the different temporal feature engineering techniques on 3 advertising campaigns with different types.

Technique	Travel	Autos	Cameras
Baseline	7.33%	5.03%	7.33%
Recency	25.58%	27.3%	26.81%
Decayed Intensity($d = 1.1$)	11.67%	7.44%	9.23%
Positive Long Term Relative Interest Change	21.48%	23.05%	21.28%
Negative Long Term Relative Interest Change	13.65%	14.4%	11.95%
Positive Short Term Relative Interest Change	9.8%%	10.46%	9.78%
Negative Short Term Relative Interest Change	10.4%	12.5%	13.58

Table 7: Feature type distribution in resulting SVM models for campaigns of different types.

4. RELATED WORK

The work presented in this paper is related to recent research in the areas of user profile generation, audience selection and response prediction for display advertising. The

emergence of the web has allowed for collection and processing of user data magnitudes larger than previously possible. Thus has resulted in spike of interest in user data analysis and profile generation as reported in [8, 10, 14, 18]. Profile generation has been reported for a few different applications. In [23] the authors describe profiles for search personalization. Here, the authors build profiles based on the query stream of the user and users similar to it. The authors also report an alternative that is based on the relevance feedback approaches in information retrieval over the documents that the user have perceived as relevant. Both techniques are orthogonal to the work presented in this paper and could be used to produce a potentially richer set of features that will serve as an input to the topical analysis.

User profile generation is also studied in other online settings and also for content recommendation (e.g., [13, 16, 17]). Most of these focus on detecting the user's short term vs long term interest and using these in the proposed application. In our case, we blend the short term and long term interests into a single profile. A survey of user profile generation can be found in [10].

In [11], the authors propose generating user profiles for online services where the user data comes from a variety of heterogeneous sources. Such profiles are sparse for individual service, but the authors show that using the data from multiple services allows for overcoming the sparseness.

In the area of audience selection, Provost et al. [21] have recently shown that user profiles can be built based on covisitation patterns of social network pages. These profiles are used to predict the performance of brad display advertisements over 15 campaigns.

In [7] the authors discuss prediction of clicks and impressions of events (queries, page views) within a set of predefined categories. Supervised linear poisson regression is used to model these counts based on a labeled set of user-class pairs of data instances. Here, exponential decay is suggested as a way to down-weight the past events.

Conversion predicting is a difficult problem where the task is usually combined with click prediction, e.g. predicting first clicks and then predicting the converters in the clickers population [12, 4, 22]. The advantage of this division relates to business logic: the publisher (such as Yahoo! or Google) has data about how likely users are to follow various paths towards clicks on advertisements on their site. On the other hand, advertisers have more information about the paths of users on their website. Therefore, there is a certain cleanliness with regards to data ownership.

On the other hand, paying for conversions has two effects. First of all, there is the maximum level of quality control of the traffic. Problems such as click fraud [9, 20, 26] do not arise. Second, when creating a conversion model, one is aware of the abundance of users who did not click. This plethora of negative data can really help: intuitively, knowing that someone was unlikely to click makes it quite possible that they are unlikely to convert.

In this paper, we focused on building such conversion models. We compared our approach with existing behavioral targeting methods (such as [8, 25]) which optimize for click-through rates and showed how optimizing directly for conversions can lead to improved performance. Compared to previous work on conversion optimization [4, 5, 6, 12, 22, 19, 2, 24, 3, 1], our work focuses on feature definition and weighting that were not explored in the previous approaches.

5. CONCLUSIONS AND FUTURE WORK

Display advertising impacts virtually every Web user and provides the financial backing that allows for the Web's diversity. Audience selection is one of the key technical challenges of display advertising and requires effective and efficient user profiles. In this paper we examined how to compose such profiles from the raw events, by examining different ways to detect patterns of behavior. The work confirms that recent user behavior (especially the most recent two weeks) is more indicative of future ad interaction. We explored a set of interest decay techniques and showed that there is a needed balance between "forgetting" the past completely and diluting the recent activity. The best performance was achieved by balancing both the long term and the short term history. Next, we showed that change in behavior, such as sudden spikes of interest in a particular topic, highly improves the prediction power. Such features cannot be achieved with the simple frequency and decay models. The improvement in performance using these features over a real world dataset is around 5%, which is very significant improvement in this domain, where usually major production improvements measure in 1-2% range.

It is worth mentioning that variations of the techniques presented in this paper were deployed to production as part of the platform presented in [3] and achieved a substantial boost in the platform's targeting accuracy, as computed by online metrics such as eCPA, compared to the old system (we don't share these results for privacy reasons). Furthermore, the addition of the newly-presented temporal feature engineering techniques to the system actually running in production did impose a relatively neligible overhead in terms of overall platform running time. Achieving a boost in the platform's targeting performance without affecting its scalability represents a huge success for a platform intended to effectively mine histories of billions of Internet users on a daily basis.

The behavioral models exploited in this paper expand the state-of-the art in the targeting domain. However, they are far from the final step in this exploration. One can think of many more complex versions of the decayed-intensity features, the recency features and the interest change features. For instance, one may think of modeling the user interest change for more than two time windows and/or for time window of variable sizes. Additionally, more complicated patterns that use temporal and sequencing information can be crafted as a continuation of the work presented in this paper. Such increasingly complex models might bring the display advertising closer to the often quoted goal of the industry to achieve the level of relevance of the search advertising without having explicitly stated intent as a user query.

Acknowledgment

The authorship of this paper reflects the contributions of the authors to the paper. Actually deploying the work presented in this paper to production as part of the platform presented in [3] is a result of a large team effort across multiple organizations at Yahoo!, including the Yahoo! Labs, Yahoo! Advertising Engineering and Product Management groups. We are thankful for the opportunity to have worked with those teams to deploy this state-of-the-art display advertising platform to production.

6. REFERENCES

[1] A. Ahmed, M. Aly, J. Gonzalez, S. M. Narayanamurthy, and A. J. Smola. Scalable inference in latent variable models. In *Proceedings of the International Conference on Web Search and Web Data Mining*, WSDM '12, 2012.

[2] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, 2011.

[3] M. Aly, A. Hatch, V. Josifovski, and V. K. Narayanan. Web-scale user modeling for targeting. In *Proceedings of the World Wide Web Conference*, WWW '12, 2012.

[4] N. Archak, V. S. Mirrokni, and S. Muthukrishnan. Mining advertiser-specific user behavior using adfactors. In *Proceedings of the 19th International World Wide Web Conference*, 2010.

[5] A. Bagherjeiran, A. O. Hatch, and A. Ratnaparkhi. Ranking for the conversion funnel. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 146–153, 2010.

[6] A. Bagherjeiran, A. O. Hatch, A. Ratnaparkhi, and R. Parekh. Large-scale customized models for advertisers. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, 2010.

[7] Y. Chen, D. Pavlov, and J. Canny. Large-scale behavioral targeting. In J. Elder, F. Fogelman-Soulié, P. Flach, and M. J. Zaki, editors, *Proceedings of the 15th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 209–218. ACM, 2009.

[8] Y. Chen, D. Pavlov, and J. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[9] I. Click Forensics. Click fraud index. `http://www.clickforensics.com/resources/click-fraud-index.html`, 2010.

[10] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, pages 54–89. Springer, 2007.

[11] R. Ghosh and M. Dekhil. Discovering user profiles. In *18th International World Wide Web Conference (WWW2009)*, pages 1233–1234, 2009.

[12] Google, Inc. Google analytics. `http://www.google.com/analytics`.

[13] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: User behaviour as a predictor of a successful search. In *WSDM 2010*, pages 221–230, 2010.

[14] H. R. Kim and P. K. Chan. Learning implicit user interest hierarchy for context in personalization. In W. L. Johnson, E. André, and J. Domingue, editors, *Proceedings of the 2003 International Conference on Intelligent User Interfaces (IUI-03)*, pages 101–108, New York, 2003. ACM Press.

[15] Y. Koren. Collaborative filtering with temporal dynamics. *Commun. ACM*, 53(4):89–97, 2010.

[16] R. Kumar and A. Tomkins. A characterization of online search behavior. In *19th International World Wide Web Conference (WWW2010)*, pages 561–570, 2010.

[17] L. Li, W. Chu, J. Langford, and R. Schapire. A contextual bandit approach to personalized news article recommendation. In *19th International World Wide Web Conference (WWW2010)*, pages 661–670, 2010.

[18] L. Li, Z. Yang, B. Wang, and M. Kitsuregawa. Dynamic adaptation strategies for long-term and short-term user profile to personalize search. In G. Dong, X. Lin, W. Wang, Y. Yang, and J. Xu-Yu, editors, *Advances in Data and Web Management*, volume 4505 of *Lecture Notes in Computer Science*, pages 228–240. Springer, 2007.

[19] S. Pandey, M. Aly, A. Bagherjeiran, A. Hatch, P. Ciccolo, A. Ratnaparkhi, and M. Zinkevich. Learning to terget: What works for behavioral targeting. In *Preceedings of the ACM Conference on Information and Knowledge Management*, CIKM '11, 2011.

[20] Y. Peng, L. Zhang, M. Chang, and Y. Guan. An effective method for combating malicious scripts clickbots. In *Proceedings of the 14th European Symposium on Research in Computer Security*, 2009.

[21] F. J. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In *KDD*, pages 707–716, 2009.

[22] B. Rey and A. Kannan. Conversion rate based bid adjustment for sponsored search auctions. In *Proceedings of the 19th International World Wide Web Conference*, 2010.

[23] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *WWW*, pages 675–684, 2004.

[24] S. Tyler, S. Pandey, E. Gabrilovich, and V. Josifovski. Retrival models for audience selection in display advertising. In *Preceedings of the ACM Conference on Information and Knowledge Management*, CIKM '11, 2011.

[25] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web*, 2009.

[26] L. Zhang and Y. Guan. Detecting click fraud in pay-per-click streams of online advertising networks. In *Proceedings of the 28th IEEE International Conference on Distributed Computing Systems*, 2008.