Personalized click prediction in sponsored search

Conference Paper · January 2010

DOI: 10.1145/1718487.1718531 · Source: DBLP

2 authors:

Haibin Cheng
Wuhan University of Technology,China, Wuhan
18 PUBLICATIONS **902** CITATIONS

SEE PROFILE

Erick Cantu-Paz
Association for Computing Machinery
94 PUBLICATIONS **6,650** CITATIONS

SEE PROFILE

Personalized Click Prediction in Sponsored Search

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

Erick Cantú-Paz
Yahoo! Labs
4401 Great America Parkway
Santa Clara, CA, U.S.A
erick@yahoo-inc.com

ABSTRACT

Sponsored search is a multi-billion dollar business that generates most of the revenue for search engines. Predicting the probability that users click on ads is crucial to sponsored search because the prediction is used to influence ranking, filtering, placement, and pricing of ads. Ad ranking, filtering and placement have a direct impact on the user experience, as users expect the most useful ads to rank high and be placed in a prominent position on the page. Pricing impacts the advertisers' return on their investment and revenue for the search engine. The objective of this paper is to present a framework for the personalization of click models in sponsored search. We develop user-specific and demographic-based features that reflect the click behavior of individuals and groups. The features are based on observations of search and click behaviors of a large number of users of a commercial search engine. We add these features to a baseline non-personalized click model and perform experiments on offline test sets derived from user logs as well as on live traffic. Our results demonstrate that the personalized models significantly improve the accuracy of click prediction.

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, Measurement, Design, Experimentation, Human Factors

Keywords

Sponsored Search, Click Prediction, Personalization, User Profile, Demographic, Click Feedback, Maximum Entropy Modeling

WSDM'10, February 4–6, 2010, New York City, New York, USA.

1. INTRODUCTION

Sponsored search is an Internet advertising system that generates most of the revenue of search engines by presenting targeted advertisements along with the search results. In the common "pay-per-click" model, advertisers are charged for each click on their ads. To maximize the revenue for a search engine and maintain a desirable user experience, the sponsored search system needs to make decisions related to selection, ranking and placement of the ads. These decisions are based greatly on the probabilities that users will click on ads, and therefore accurate click prediction is an essential problem in sponsored search.

Current state-of-the-art sponsored search systems typically rely on a machine learned model to predict the clickability of ads returned for a user search query. For the experiments in this paper, we will use a system based on numerous user-independent features as a baseline. Some of these features are based on the similarity of the query to the text of the ads and range in complexity from simple word or phrase overlap to more sophisticated semantic similarities between the query and different elements of the ads. Other features are related to the historical performance of ads. In our experience, certain statistics of the past performance of ads are good predictors of the click probability. Yet another group of features gives contextual information, such as the time of day or day of the week. All of these features ignore the users both individually and as parts of groups with similar behaviors and therefore the model will predict the same probability of click for every user. We believe that personalizing the click prediction benefits both the users and the advertisers: The users will be presented ads in the manner that is most relevant to them, and the advertisers will receive clicks from users who are more engaged with the ads.

The objective of this paper is to present the design of personalized click prediction models. An essential part of these models is the development of new user-related features. We base our features on observations over a significant volume of search queries from a large number of users. Our observations suggest that user click behavior varies significantly with regard to their demographic background, such as age or gender. We investigate the click distribution for different users from various backgrounds and design a set of demographic features to model their group clicking patterns. Recognizing that there is still significant variability in demographic groups, we also investigate user-specific features. The new user-related features are integrated with other features in a maximum entropy classification framework and contribute to the final predicted clickability score for each

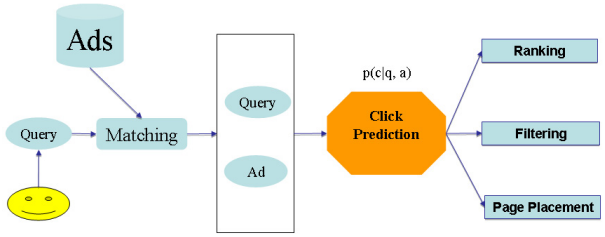

Figure 1: Overview of sponsored search system.

query-ad-user tuple. We tested the personalized models offline on a large test set based on log data of the Yahoo! search engine. The results show that the personalized models are significantly more accurate than the non-personalized baseline. In addition, we report results of a test on live traffic of a personalized model that confirm the offline evaluations.

The rest of this paper is organized as follows. Section 2 briefly outlines one approach to click prediction in sponsored search. Next, we present a study on user click distribution in Section 3. The personalized click prediction framework is proposed in Section 4. The user related features developed in our work are introduced in Section 5. The experimental setup and results are presented in Section 6. We discuss some related work in Section 7 and conclude the paper with a summary of our findings and proposals for future work in Section 8.

2. CLICK PREDICTION

Sponsored search is a complex advertising system that presents ads to users of search engines. It involves several processes as illustrated in Figure 1. The input query from the user is used to retrieve a list of candidate ads. The exact mechanisms of query parsing, query expansion, and ad retrieval used in our system are beyond the scope of this paper. For our purposes we assume that we receive a set of candidate ads that need to be scored by the click model to estimate the probability they will be clicked. This estimate is an essential component of the sponsored search system as it influences user experience and revenue for the search engine: The click probability is a factor to rank the ads in appropriate order, filter out uninteresting ads, place the ads in different sections of the page, and to determine the price that will be charged to the advertiser if a click occurs.

We formulate click prediction as a supervised learning problem. We collected click and non-click events from logs as training samples, where each sample represents a query-ad pair presented to a user. Assume there is a set of n training samples, $\mathcal{D} = \{(\mathbf{f}(q_j, a_j), c_j)\}_{j=1}^{n}$, where $\mathbf{f}(q_j, a_j) \in \Re^d$ represents the d-dimensional feature space for query-ad pair j and $c_j \in \{-1, +1\}$ is the corresponding class label ($+1$: click or -1 : non-click). Given a query q and ad a, the problem is to calculate the probability of click $p(c|q, a)$. The maximum entropy model (**ME**) [4] is well suited for this task because of its strength in combining diverse forms of contextual information, and formulates the click probability for a query-ad pair as follows:

$$p(c|q, a) = \frac{1}{1 + \exp(\sum_{i=1}^{d} w_i f_i(q, a))} \quad (1)$$

where $f_i(q, a)$ is the i-th feature derived for query-ad pair (q, a) and $w_i \in \mathbf{w}$ is the associated weight. Given the training set $\mathcal{D}$, the maximum entropy model learns the weight vector $\mathbf{w}$ by maximizing the likelihood of exponential models as:

$$\mathbf{w} = \max(\sum_{i=1}^{n} \log(p(c_i|q_i, a_i)) + \log(p(\mathbf{w}))) \quad (2)$$

where the first part represents the likelihood function and the second part utilizes a Gaussian prior on the weight vector $\mathbf{w}$ to smooth the maximum entropy model [7]. There are many approaches available in the literature [15] to solve this kind of optimization problems including iterative scaling and its variants, quasi-Newton algorithms, and conjugate gradient ascent. Given the large collection of samples and high dimensional feature space, we use a nonlinear conjugate gradient algorithm [16].

2.1 Features

An accurate maximum entropy model relies greatly on the design of features $\mathbf{f}$. There are many possible features that can be derived for the purpose of predicting click probabilities. One class of features explores the lexical similarity between the query and ads by calculating word or phrase overlap of the query to different elements of the ads. These features rely on a simple assumption that users tend to click on ads that appear to be relevant to their query and that query-ad overlap is correlated with perceived relevance. We have found some usefulness in these features, but it is clear that the discrimination power of lexical features is limited due to the typically short queries and simple ads.

Another set of features is derived from the historical performance of ads. In our experience, these features are good estimators of the future performance of ads. It is well known [9] that the click-through rate (CTR) of search results or advertisements decreases significantly depending on the position of the results. To account for this position bias, we use a position-normalized statistic known as clicks over expected clicks (COEC):

$$\text{COEC} = \frac{\sum_{r=1}^{R} c_r}{\sum_{r=1}^{R} i_r * CTR_r}, \quad (3)$$

where the numerator is the total number of clicks received by a query-ad pair; the denominator can be interpreted as the expected clicks (ECs) that an average ad would receive after being impressed i_r times at rank r, and CTR_r is the average CTR for each position in the result page (up to R), computed over all queries and ads.

We can obtain COEC statistics for specific query-ad pairs, and we have found these features to be good predictors of click probabilities. However, many impressions are needed for these statistics to be reliable and therefore data for specific query-ad pairs can be sparse and noisy. To ameliorate this problem, we can obtain additional COEC statistics

by counting clicks and expected clicks over aggregations of queries or ads. The full details of these aggregations are outside the scope of this paper, but briefly we note that the advertisers organize their ads in ad groups, campaigns, and accounts. We can exploit this organization and count clicks and expected clicks for different combinations of query-ad groups, campaigns, and accounts. Of course, other aggregations using query or ad clusters are possible and can be used in practice. The COECs and expected clicks computed over each of the aggregation levels are used as inputs to the maximum entropy framework.

There are other features that provide additional context that we have observed are helpful in predicting the click probabilities. For example, features such as time of day, day of week, and the position on the page where ads have been displayed.

Note that all the features described so far ignore differences among users. We will use the model described in this section as a baseline for our personalized models with user information.

2.2 Feature Quantization, Conjunctions, and Selection

The click feedback (CF) features are expected to be skewed, and this may cause problems to many learning algorithms, including maximum entropy. In our work, the features are transformed into the log form and then quantized using a simple K-means [5] clustering algorithm with objective function:

$$\arg\min \sum_{j=1}^{k} \sum_{v_i \in C_j} ||v_i - u_j||^2 \quad (4)$$

where v_i is the feature value after log transform, u_j is the centroid of cluster C_j and k is the number of clusters. Each cluster of feature values represents a quantized segment. We introduce binary indicator features for each segment, and use these binary features as inputs to the ME model. We also introduce a binary indicator feature to indicate that a certain value is missing, which is common for click feedback features (for new ads or new queries, for example).

To model relationships among features, we create feature conjunctions by taking the cross product of the binary indicators for pairs of features. We select the features to be conjoined using domain knowledge. For example, we have conjunctions between COECs and expected clicks for each of the levels over which we aggregate click feedback statistics. If the COECs and ECs are quantized into ten segments each, we will add 100 new binary features corresponding to all the possible combinations of the segments.

The feature set will grow exponentially after quantization and adding conjunctions. To limit the growth, we eliminate binary features and conjunctions that appear less than 10 thousand times in the training data. After quantization, conjunction and selection, the features are used as inputs to the maximum entropy click model.

3. USER CLICK ANALYSIS

As a first step towards our goal for personalized click models, we conduct a preliminary analysis of the log data to unveil some patterns of user behavior in sponsored search.

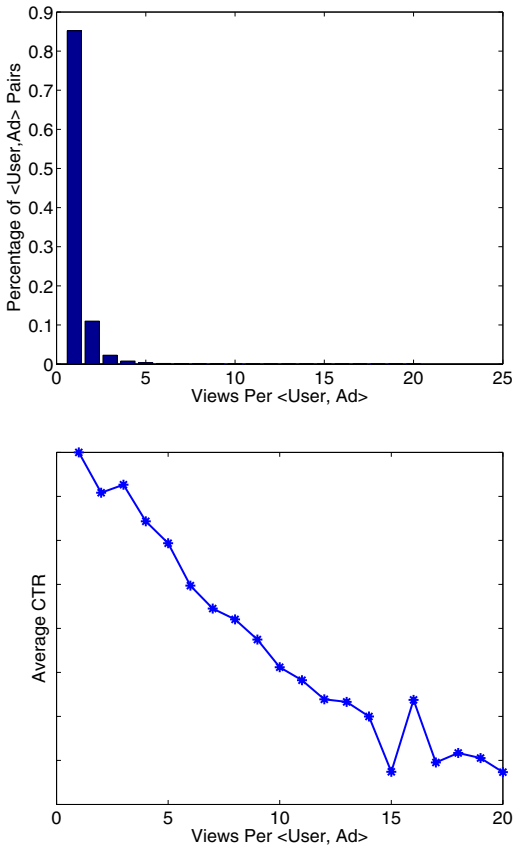

Figure 2: The upper graph shows the distribution of user-ad views. The bottom graph shows that CTR of user-ad pairs is inversely proportional to the number of views.

3.1 User-Ad View and Click Distributions

For the initial part of the analysis, we consider one day of data (2008/08/01) sampled from the logs of the Yahoo! search engine. Our initial objective is to examine the distribution of user-ad views and clicks. The upper graph of Figure 2 shows that the distribution of user-ad views follows a power law distribution. In the figure, the X-axis represents the number of daily observations for each user-ad pair and the Y-axis represents the percentage of unique user-ad pairs. The data show that approximately 85% of user-ad pairs are observed only once and 10% are observed twice, which means that only 5% of the user-ad pairs are observed more than twice. The bottom graph of Figure 2 plots the average CTR for each group of user-ad pairs with the same number of views (note the sudden jump at $x = 16$ may be due to limited data size). The average CTR is clearly decreasing linearly when the views are increasing. This trend motivates one of the features that we describe in Section 5, in which we compare the past click behavior of each user to a group of users with similar searching patterns.

3.2 View, Click and Bid Distributions per User Demographics

In marketing, consumers are segmented into groups with similar needs or shopping behaviors. This segmentation can

occur along different dimensions, such as behavioral, demographic or geographic variables. In the work reported here, we segment users into demographic groups using data disclosed by them as part of the registration process to obtain an account with Yahoo!. These demographic data are not available for all users of the search engine and are not guaranteed to be accurate.

For this part of the study, we use the entire training data that will be described in detail in Section 6, but it is sufficient to say for now that these data span two months of activity of more than 400 million unique users. We partition users with demographic information into groups according to their age, job, marriage status, interests and their occupation as shown from the top to the bottom panels in Figure 3. The X-axes of the graphs in Figure 3 represent the demographic groups. For example, age is partitioned into 8 segments and there are four possible values for marriage status. The Y-axes show the number of views, average CTR, and average bids for each demographic category. We omit the scales to protect proprietary information. The average bids in the right panel indicate how the ads shown to different segments of users differ in terms of the bid advertisers make and are suggestive of query differences among the groups.

Here we list a few of our observations from Figure 3:

- From all the graphs, we can clearly see that female users always view and click more ads than males and that those ads have higher bids, except a few rare and noisy cases like in age group 0–13.
- Click-through rates have a very strong positive correlation with age.
- Users who are single are more likely to click than married and divorced users.
- Users with travel, shopping and business interests are more likely to click ads, whose bids seem to be higher too. Users with music, sport entertainment interests are less likely to click ads.
- Users whose occupations are in financial, engineering, and travel industries are more likely to click than those in entertainment and education industries.

These data show strong correlations between the users' demographic background and their searching and clicking behaviors, which could used to improve the accuracy of click prediction.

4. PERSONALIZED CLICK PREDICTION

The baseline prediction model described in Section 2 estimates the click probabilities using user-independent features and therefore will produce the same estimates for every user. In this section, we propose a family of models that incorporate user information.

Our vision of personalized models is a direct extension of the maximum entropy model introduced in Section 2. Let $\mathcal{D} = \{(\mathbf{f}(q_j, a_j, u_j), c_j)\}_{j=1}^{n}$ represent the new training set where each sample j represents a click or non-click event when ad a_j is presented to user u_j for query q_j. Instead of assigning a click probability for each new query-ad pair as shown in Equation 1, we develop a new click prediction function $p(c|q, a, u)$ for each query-ad-user tuple as:

$$p(c|q, a, u) = \frac{1}{1 + \exp(\sum_{i=1}^{d} w_i f_i(q, a, u))} \tag{5}$$

where $f_i(q, a, u)$ is a feature derived from the query (q), ad (a) or user (u) or a combination thereof, and w_i is the weight associated with each feature. Similarly, the weight vector $\mathbf{w}$ is learned by maximizing the likelihood of exponential models with Gaussian prior as:

$$\mathbf{w} = \max(\sum_{i=1}^{n} \log(p(c_i|q_i, a_i, u_i)) + \log(p(w))). \tag{6}$$

Again, the nonlinear conjugate gradient method is used to solve this optimization problem.

In the personalized click model, user information is integrated into the ME model directly as features and combined with features derived from the query and the ads. The learning algorithm will automatically tune the weights as before. The next section describes the user-related features we use.

5. USER FEATURES

We consider two sets of user-related features. The first set is composed of demographic group features, which capture the behaviors of group of users segmented according to their demographic background. The second set of features is composed of features associated with each particular user.

For both sets of features, we accumulate historical information over multiple days in the past. We remove the users with fewer than two ad views during a day. We found that this was a convenient way to smooth the data. Then, for each item, we consider a time window that is long enough to accumulate 500 ad impressions or reach a maximum of 90 days. This variable length window is a good compromise between accumulating enough data to compute reliable statistics for items with low daily activity and capturing temporal changes in behavior for items with high activities.

5.1 Demographic Features

In our work, we partition users into demographic segments based on age, gender, marriage status, interests, job status, and occupation. To incorporate this information into the maximum entropy framework, we introduce binary features for each possible value of the demographic variables. For example, there are eight binary features indicating each of the possible age groups, and only one of these eight features triggers for each user.

Besides using the demographic information as direct inputs to the click model, we develop click feedback features to capture the historical behavior of user groups. We can segment the users based on one or multiple of the demographic variables available to us. Specifically, in our experiments we take the cross product of these gender and age groups to partition the users, resulting in 16 groups composed of users with the same gender and of the same age group. Once the demographic groups are formed, we envision numerous combinations with other factors for which we already compute click feedback features. For example, we already accumulate historical information for all the ads of specific advertiser accounts and, along those lines, we can conceive of accumulating data for combinations of accounts and demographic groups. Such a combination would capture the relative preference of different demographic groups to specific brands or merchants. Table 1 lists some of the combinations we conceive as being useful in click prediction.

As for specific features, our experiments use the COEC and expected click features described earlier, but we can

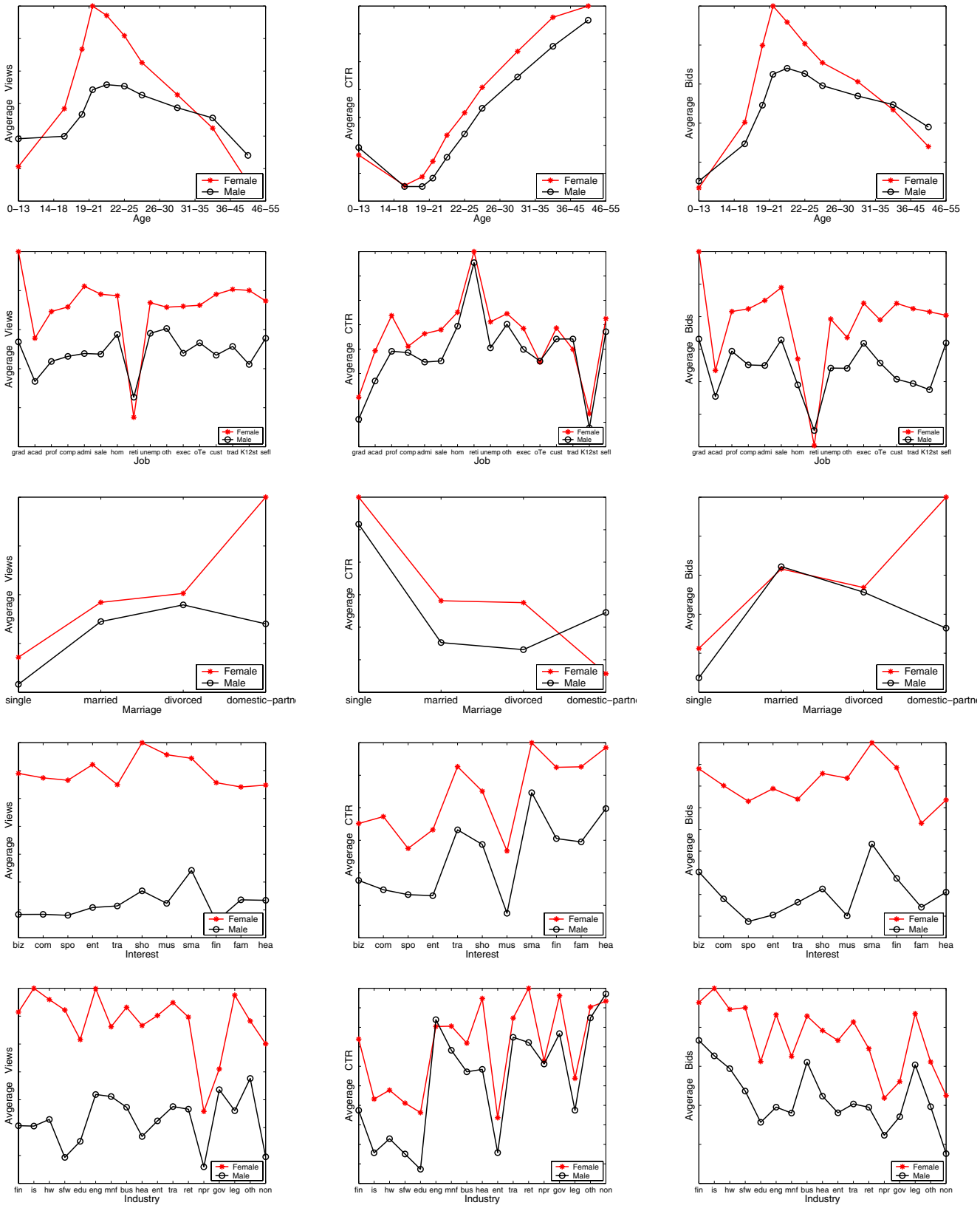

Figure 3: User click, view and bid distributions (from left to right) with regard to demographic background: age group, job occupation, marriage status, user interest, industry (from top to bottom).

Table 1: Examples of User Demographic Features. The items in *italic* case denote data used in the experiments reported in this paper. "Demo" stands for any partitioning of the users based on demographic variables.

Feature class	Feature examples
demographic profile	*gender, age group, marriage status, interests, job, occupation*
demo CF	demo
demo-ad CF	demo-domain, demo-account, *demo-ad group*, demo-campaign, demo-creative, demo-term, demo-phrase (in title, description, etc.)
demo-query CF	*demo-query*, demo-phrase(in query)
demo-location CF	*demo-most-specific-location-available*, demo-zip, demo-state

envision numerous other features such as the total number of ad views (**NV**), total number of ad clicks (**NCLI**), and total number of unique queries (**NQ**). We hypothesize that adding demographic features to the click model will improve the accuracy of click prediction and finally improve the ranking, filtering and placing ads in sponsored search. For example, 20-year old female users that click frequently on ads of a certain retailer will be presented with more of these ads and in more prominent positions than the average user.

5.2 User-Specific Features

User-specific features capture individual user's interactions with the ads shown in the search result page. This is attractive as there could be significant variability between members of demographic groups. However, deriving user specific features from log data is challenging for several reasons. First, user-specific data are very noisy because of the intrinsic randomness in users' actions. For example, a user may click an ad accidentally. Second, the user specific data are very sparse since a large portion of the search users may not click on any ads in a given time period.

A simple user-specific feature could be the user CTR measured over all queries and all ads. However, it is probably not fair to compare the CTR of users seeing more ads (e.g., commercial queries) to the CTR of users seeing fewer ads (e.g., academic queries). To ameliorate this bias, we design a new feature called user activity normalized COEC (**UCOEC**). This feature is inspired by the observations in Section 3.1, where we observed that the average CTR decreases as the number of views of user-ad pairs increased. To compute the UCOEC feature, users are grouped together based on the total number of ads they have seen. The average CTR is calculated for each user group. Then each user's CTR is normalized by dividing it by the group average CTR. The calculation of UCOEC is denoted as:

$$\text{UCOEC}_u = \frac{\text{CTR of user } u}{\text{Average CTR of } u\text{'s group}} \tag{7}$$

The intuition behind UCOEC is to indicate a user's click activity compared to users with similar searching behavior.

We can also derive other user-specific click feedback features at the user, user-query, and user-ad levels. These features would be similar to the demographic features described in the previous section, but using individual users instead of aggregating historical information over groups of users. Specific features such as EC, COEC, NV, NCLI, and UCOEC can be derived over the different user-specific levels of aggregations. A list of user-specific click feedback features are described in Table 2.

The user-level click feedback features measure the propensity of individual users to click on ads in general. User-query click feedback features are a bit more specialized and capture the propensity of users to click on certain queries or groups of queries. Finally user-ad features capture the user preferences on certain ads or advertisers (represented by their domain or account).

6. EXPERIMENTS

This section first describes the models and data sets used in the evaluations. Next, we present offline and online evaluations of the accuracy of the personalized models compared to a non-personalized baseline model.

6.1 Experimental Methods

We are interested in comparing models with different sets of features to investigate their relative usefulness in click prediction. Our evaluation methodology is two-fold: first we compare the different models using a test set, and then we evaluate the most promising models on live user traffic. In both cases, the query-ad pairs are ordered by the predicted clickability score and traditional methods such as precision-recall (**P-R**) curves as well as area under curve metric (**AUC**) are used to measure the accuracy of the models. Here precision is defined as the number of query ad pairs clicked on by users divided by the total number of query ad pairs labeled as click by the model, and recall is defined as the number of query ad pairs labeled as click by the model divided by the total number of actually clicked query ad pairs.

We evaluated six personalized models with different combinations of features against one non-personalized baseline model. All the models used the same lexical and contextual features and a basic set of user-independent click feedback features. The variations in feature sets were related to user-dependent features. Specifically, we tested models with the following combinations of user-dependent features:

1. user COEC and ECs
2. user and user-account COEC and ECs
3. user and user-query COEC and ECs
4. user, user-account, and user-query COEC and ECs
5. user, user-account COEC and ECs, and demographic features
6. demographic features

The demographic features above refer to the categorical demographic features as well as click feedback COECs and ECs aggregated over (gender-age group-query), (gender-age group-ad group) and (gender-age group-location).

The training and testing data used in our experiments were sampled from the Yahoo! sponsored search traffic logs

Table 2: Examples of User-Specific Features. The items in *italic* case denote data used in the experiments reported in this paper. "Demo" stands for any partitioning of the users based on demographic variables.

Feature class	Feature examples
user level CF	*user*
user-query CF	user-query, user-phrase(in query)
user-ad CF	user-domain, *user-account*, user-ad group, user-campaign, user-creative, user-term, user-phrase (in title, description, landing page etc.)

for a period of 2 months. Each sample of the data is a query-ad view event labeled as click or non-click by a user of the Yahoo! search engine. In the data there are approximately 467 million unique users, 100 million unique queries and 20 million unique ads. Each user is identified by a unique browser cookie in the Yahoo! network. We split the data to create training and testing sets that are disjoint with regard to both time and users. Specifically, samples of the first 51 days and associated with 3/4 of randomly selected users were used as training data, while samples associated with the remaining 1/4 of users and occurring in the last 11 days were used as testing data. There were approximately 2.44 billion samples in the training data.

To better illustrate the results, we divide the test data into two disjoint sets based on the matching technology used to retrieve the ads. In all major commercial sponsored search systems, candidate ads are retrieved either by directly matching the (possibly normalized) query to the keywords bidded by the advertiser or by some "broad" match technologies that seek to improve recall, such as query expansion. Yahoo! calls the former matches "exact" and the latter "advanced."

For each click-view event, we constructed a vector of features to be used in the click models. The lexical and contextual features were extracted in an ad-hoc fashion. The click feedback features were calculated by processing log data starting three months prior to the beginning of the period in which the training set was collected. As explained before, when computing the user-related click feedback features we eliminated events by users viewing no more than 2 ads during a day. All features were quantized and we applied feature selection using the methods described in Section 2.2.

6.2 Offline Performance Comparisons

We tested the first four personalized models with user-specific features using the offline test data sets described above. The results on the exact and advanced slices are reported with P-R curves plotted in Figure 4 and the corresponding AUCs are recorded in Table 3. In terms of AUC, adding only the user-specific click feedback feature improved over the baseline by 11.41% in the testing data with exact-matched ads and 7.41% in the testing data with advanced-matched ads. Adding the user-account and user-query click feedback features shows additional significant performance improvements: AUC increased by 16.51% with user and user-account click feedback features and 17.46% with user and user-query features in the exact-matched slice. The model with all the three sets of user related features performs the best with an 18.59% AUC increase in the exact-match slice and 12.30% increase in advanced-match. These results suggest that the user-specific click feedback features are very helpful in generating more accurate prediction score in sponsored search click modeling.

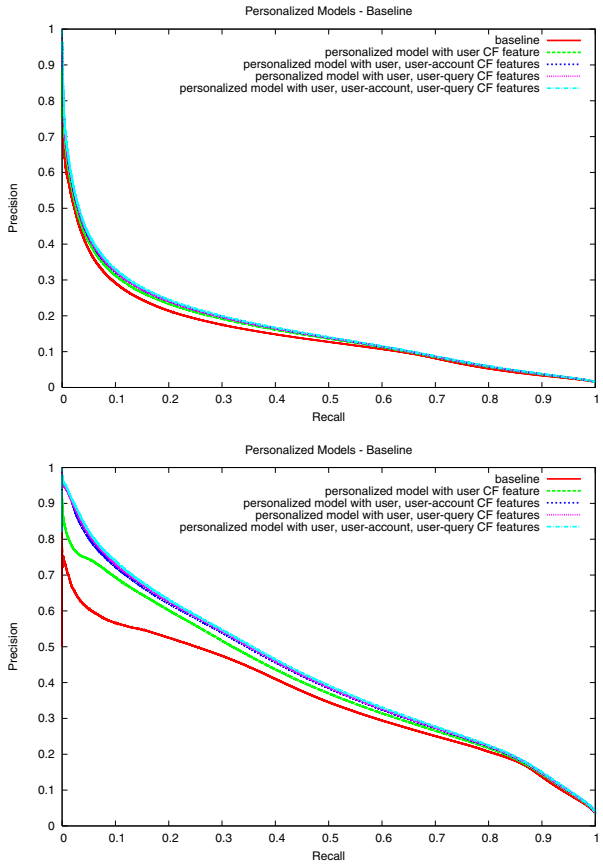

Figure 4: Offline performance of personalized models using user specific features on two slices of testing data: exact-match and advanced match.

Next, we investigate the performance of personalized models with demographic features. In addition to the comparison to the non-personalized baseline, it is interesting to see how the demographic features perform on top of user-specific click feedback features. Three models are compared in addition to the baseline: with demographic features; with demographic, user, and user-account; and with user and user-account features (these are the personalized models 4, 5, and 6). The P-R curves and corresponding AUC are reported in Figure 5 and Table 4. Adding the demographic features to the baseline resulted in a very small increase of 0.15% in the AUC on the exact-matched test data. The improvement became even less significant when the demographic features were added to the user-specific features. This is probably because the user's background information has already been captured by the user-specific click feedback features. However, the user specific data are very sparse, so we inves-

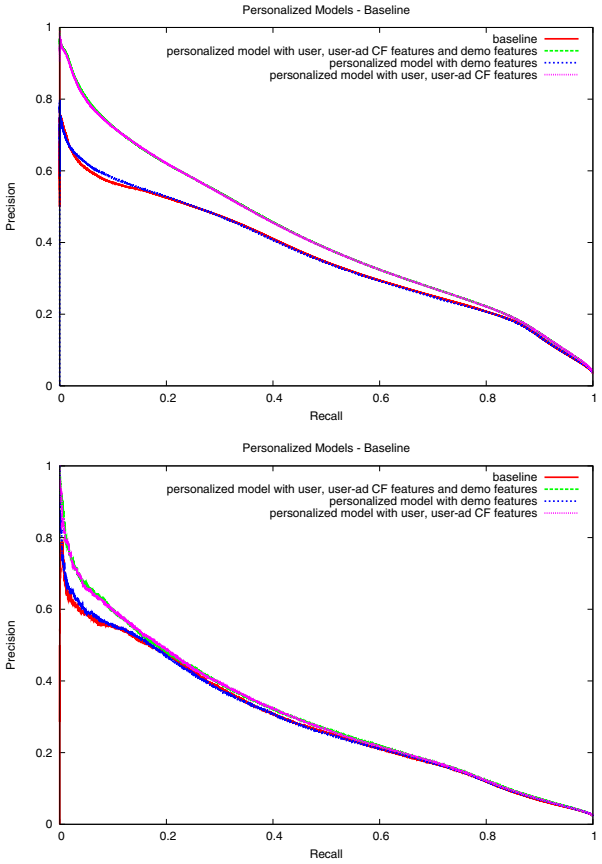

Figure 5: Offline performance of personalized models using demographic features on two slices of testing data: exact-match and data with demographic features, but no user-specific feedback (6%).

tigated whether the demographic features are useful when the user-specific features are not available. As shown in the second panel of Figure 5 and second column of Table 4, we obtained more improvement by using demographic features when user click information is not present, which accounts for approximately 6% of the testing data.

6.3 Online Performance Comparison

Since the personalized model with user-specific features shows significant improvement over the baseline, it is further tested using real time traffic. Although the model with user, user-ad and user-query features seems to work best as shown in Table 3 and Figure 4, the user-query click feature table was too large to fit into the memory available in the serving system because of the large number of user-query pairs. Thus, for online testing we selected the model with user and user-ad features.

We ran the personalized model and the non-personalized baseline for several weeks using a sufficient portion of Yahoo! search traffic and plot the P-R curve for all the logged samples during one of the days in Figure 6, where the baseline has the exact same settings except the lack of user related features. Comparing Figure 4 with Figure 6, it is obvious that the online performance is consistent with offline results with significant performance boosting in personalized model over the non-personalized baseline model.

Table 3: AUC of personalized models using user-specific features.

model	exact	advanced
baseline	0.35778	0.15043
user	0.39870 (+11.44%)	0.16157 (+7.41%)
user,ad	0.41686 (+16.51%)	0.16605 (+10.38%)
user,query	0.42025 (+17.46%)	0.16618 (+10.47%)
user,query, ad	0.42428 (+18.59%)	0.16894 (+12.30%)

Table 4: AUC of personalized models using demographic features.

model	exact	demo, but no user
baseline	0.35778	0.28735
demo	0.35832 (+0.15%)	0.28840 (+0.37%)
user,ad	0.41686 (+16.51%)	0.30462 (+6.01%)
(user,ad),demo	0.41704 (+16.56%)	0.30512 (+6.18%)

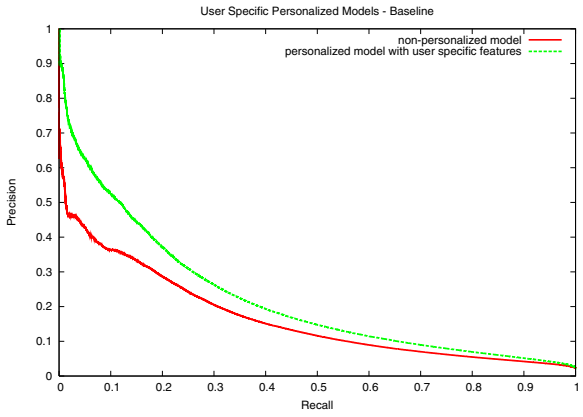

Figure 6: Online performance of the personalized model.

7. RELATED WORK

Sponsored search has received significant attention from both industry and academic research in the last decade [13][11]. To target search engine users with the most interesting ads, sponsored search needs to solve several important problems, such as query to ad matching [1], click prediction for ranking, filtering and placement of candidate ads [12], and pricing of the final presented ads [10].

As a core component in sponsored search system, click prediction uses a machine learning model such as maximum entropy [4] using various features extracted from multiple sources like user, query and ad. Certain lexical (or syntactic) features model the relevance between query and ad by treating the ad's text as a short document and building a language model as in classic information retrieval [14][13]. Besides text features, click feedback features based on aggregating historical click data [8] have been shown to be very effective in predicting the clickability of ads [6]. Furthermore, previous research shows that factors such as the position in which ads are displayed on the page affect the clicks the ads receive [9].

These previous works consider information from queries and ads, but do not capture the user behavior completely. Although there has been work in web search area to incorporate user behavior information directly to improve the web search ranking [2], there is limited research on integrating user information to improve ads ranking in sponsored search. The only work we found is the recent study on post-result user behavior in sponsored search [3], however, its goal is to predict the user's action after clicking the ads instead of improving the click prediction.

8. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a personalized click modeling framework to improve the accuracy of click prediction in sponsored search. Specifically, we derive user-specific click feedback features and demographic features from logged user data. We evaluated the personalized models offline on a large scale data set based on logs from the Yahoo! search engine. We observed significant improvements when user-specific click feedback features were added to a baseline click model without user features. We also saw notable improvements for the personalized model with demographic features compared to the baseline. Another interesting point is that adding demographic features to the model that already included user-specific features helped improve accuracy, especially for users without enough user-specific click feedback. We further tested one personalized model with user-specific features by running it on live traffic and the performance boost was consistent with the offline evaluations.

There are several promising directions for future research. One of these directions is to extend the personalized feature set. In the paper we presented numerous features that have not been completely evaluated yet, and many more features can be developed. Another direction for future research is to take into account session-based information. Events such as query reformulations, dwell time on landing pages after clicking on links, and the sequence of clicks have additional information that can be aggregated and exploited in personalized click models. Another area of research is to group users in different ways. We have seen that using self-disclosed demographic information has some value, but automatic clustering of users based on their behavior can give rise to smoothed personalized statistics that may prove useful in click prediction.

Acknowledgments

We thank our colleagues Eren Manavoglu, Divy Kothiwal, Ozgur Cetin, Anand Murugappan, Kannan Achan and Vadim Von Brzeski for their assistance with data collection and model evaluation.

9. REFERENCES

[1] V. Abhishek and K. Hosanagar. Keyword generation for search engine advertising using semantic similarity between terms. In *Proceedings of the ninth international conference on Electronic commerce*, pages 89–94, New York, NY, USA, 2007. ACM.

[2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2006. ACM.

[3] J. Attenberg, S. Pandey, and T. Suel. Modeling and predicting user behavior in sponsored search. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1067–1076, New York, NY, USA, 2009. ACM.

[4] A. L. Berger and V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71, 1996.

[5] 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.

[6] D. Chakrabarti, D. Agarwal, and V. Josifovski. Contextual advertising by combining relevance with click feedback. In *Proceeding of the 17th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2008. ACM.

[7] S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, 1999.

[8] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *Proceeding of the 17th international conference on World Wide Web*, pages 227–236, New York, NY, USA, 2008. ACM.

[9] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the international conference on Web search and web data mining*, pages 87–94, New York, NY, USA, 2008. ACM.

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

[11] D. C. Fain and J. O. Pedersen. Sponsored search: a brief history. In *Proceedings of 2nd Workshop on Sponsored Search Auctions*, 2006.

[12] Google. How are ads ranked? `http://www.google.com/support/grants/bin/answer.py?hl=en&answer=98917`.

[13] B. J. Jansen and T. Mullen. Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business*, 6:114–131, 2008.

[14] C. Liu, H. Wang, S. Mcclean, J. Liu, and S. Wu. Syntactic information retrieval. In *Proceedings of the 2007 IEEE International Conference on Granular Computing*, page 703, Washington, DC, USA, 2007. IEEE Computer Society.

[15] T. P. Minka. A comparison of numerical optimizers for logistic regression. Technical report, Microsoft, 2003.

[16] A. Mordecai. *Nonlinear Programming: Analysis and Methods.* Dover Publishing, Mineola, NY, 2003.