Personalized Ad Placement in Web Search

Stefan Schroedl, Anand Kesari, Leo Neumeyer
Sponsored Search Sciences
Yahoo! Labs
4401 Great America Pkwy
Santa Clara, CA 95054
{stefan,anands,neumeyer}@yahoo-inc.com

ABSTRACT

Commercial search engines typically provide web result links (a.k.a. "organic" results) along with advertisements in response to user queries. Well-targeted ads can be quite useful to shoppers, but there is always a risk that less relevant ads can affect the search engine experience. In the long term, excessive advertising may result in "ad blindness" (customarily skipping over the ad section); or it might even prevent them from returning to the search engine altogether. Because advertising is typically the main revenue source of search engines, we propose to balance ad impact on a per-user basis.

In this paper, we use long– and short-term historical user behavior to infer the user's relative preferences between organic results and ads. This information is used to adapt the number of ads shown (*filtering*) and their location on the search results page (*page placement*).

We trained and evaluated several machine-learned click prediction models. The offline results show significant gains in predictive accuracy. In addition, we tested our models on live traffic from a commercial search engine. Our real-time implementation makes adaptation directly perceptible for users. Behavioral metrics show significant improvements over a non-personalized base line.

Keywords

Web search advertising, personalization, page placement, ad filtering, click prediction, machine learning

1. INTRODUCTION

Current commercial search engines typically provide organic web results in response to user queries and then supplement with ads that provide revenue based on a "cost-per-click" billing model [8, 17]. Ads are selected from a database populated by advertisers that bid to have their ads shown on the search engine result page (*SERP*). The search engine typically uses an estimated probability of a click on an ad, together with its bid in order to decide which ads to show and in which order.

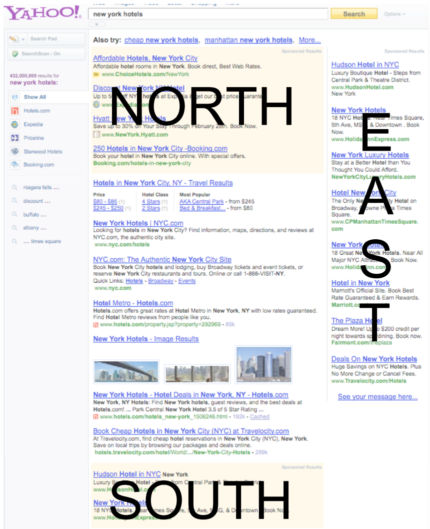

Figure 1: Search Result Page with Ad Placement Regions.

In addition to selecting and ranking candidate ads, it must decide how many ads to show (*"filtering"*) and how prominently (*"page placement"*): above the search results (which we will call *"North"*), or in a separate column to the right (*"East"*); see Fig. 1. It is well known [11, 6]) that the real estate at the top of the SERP is much more valuable than the rest of the page due to highly selective user attention. The stakes are high, since ads receive significantly more clicks in the North; however, on the flip side, showing an irrelevant result here can hurt user experience much more severely, e.g., by distracting him from more relevant web results. The user might subsequently lose confidence to click on *any* ads, or stop using the search engine altogether in the worst case.

The typical terseness and ambiguity of search queries has led to a large body of research to refine web search results based on the context of an individual user; see [19] for a general overview. Besides *explicit* relevance feedback, one way of personalizing search is by means of taking the history of interaction with the search engine as *implicit* feedback [1]. From past queries and clicks, a user model can be built either in terms of a topic/concept classification hierarchy, or in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ADKDD'10, July 25, 2010, Washington D.C., USA.
Copyright 2010 ACM 978-1-4503-0221-0 ...$10.00.

terms of a vector space model (bag of keywords). By building similar representations for result documents, individual preferences can be predicted. Another line of research tries to form user groups for collaborative filtering approaches [7]. All these predictions can then be used for re-ranking or filtering the result list.

In principle, many of these techniques are also applicable to match *advertisements* to users that are more attractive based on their preferences. However, there are also aspects that are fundamentally different from information retrieval perspective predominant in web search. For example, ads in the North region directly compete and displace the organic links. Therefore, it is useful to know the user's relative preference between organic results and ads, or the degree of his commercial intent. This can be a dimension of user preference orthogonal to the aspect of actual contents, and to the best of our knowledge it has not been addressed in depth in the current literature.

In this paper, we explore approaches to improve user experience by adapting to an individual's relative preferences between web and ad results. We want to show less ads to purely information-seeking, ad-averse users and conversely more to shoppers, thus improving overall satisfaction. Moreover, we believe that personalization benefits advertisers as well, as they receive more clicks from users more engaged with the ads.

The rest of the paper is organized as follows. In the following section, we set the stage by briefly summarizing our sponsored search terminology, particularly by defining the problems of page placement and ad filtering. Sec. 3 explains our approach of modifying these algorithms based on a factor called *user click propensity (ucp)*, and how the latter one is derived from historical click log data. Sec. 4 refines this factor by means of more granular machine-learned prediction models. For validation, we present results about offline click prediction accuracy (Sec. 5), and report on live traffic experiments to confirm that this indeed translates into favorable behavioral metrics.

2. REVIEW OF SPONSORED SEARCH

Search engines typically take a four-stage approach to the sponsored search problem:

1. Find relevant ads for a query, by retrieving them from a database whose bid terms are either identical (*"exact match"*) to the query, or are presumed to be related according to various rewriting techniques (*"advanced match"*).

2. Estimate an ad-specific click-through rate *(CTR)*. Historical click rate is a good predictor, if plenty of data is available. To cope with sparseness of data, search engines typically rely on a machine-learned *click prediction model* that considers a variety of additional features, such as syntactic and semantic similarity between the query and the ad snippet [15].

   Since display position has a dominant influence on CTR, regardless of ad quality, we need a *position-normalized* measure. One simple yet easily computable model is *COEC (clicks over expected clicks, see e.g. [21] for details)*: position bias is captured in terms of a *reference CTR* $\overline{ctr}_i$, i.e., the mean click-through rate at a given page position i (averaging over all ads that have been shown at position i); the ad-specific click is computed by dividing the total number of obtained clicks by the sum of *expected clicks* (according to the reference CTR) for the position(s) the ad was shown at.

3. Rank, price, and possibly filter ads. It is typical to order the ad candidates by $coec \cdot bid$ [8, 17]; this product is often also called *eCPM*. The cost per click is determined as the minimum amount an advertiser would have to bid to maintain his rank; this results in a cost of $eCPM_{i+1}/coec_i$ for the ad at rank i, or a *minimum reserve price* in case of the last ad. *Filtering* can be implemented based on the same score by imposing a minimum threshold; ads with lower eCPM are not shown. The rationale for filtering is to improve user experience by showing a less 'cluttered' results page, while sacrificing only a small fraction of search engine revenue.

4. Decide how to display the ads on the search page (most importantly, how many ads to show in the north section). A simplified procedure for page placement might work as follows. For each ad at rank i, we determine a *page placement score*. One implementation is to estimate expected revenue as

   $$\overline{ctr}_i \cdot coec_i \cdot bid_i.$$

   Assume a fixed, global threshold θ_{north}. Starting with the top-ranked ad, compare the estimated revenue with this threshold. If it is larger, allocate the ad to the North; continue with the following ads, until either an ad doesn't qualify, or a maximum available number of display slots have been filled. Place the remaining ads (if any) into the East[1]. This procedure attempts to maximize expected revenue, given a constrained budget of *North Footprint* (*NFP*, average number of North ads per search). The tuning of θ_{north} determines the balance between revenue and user perception, and is a business decision.

Search engines customarily use *web search logs*, e.g., to estimate ad click rates. In the following, we assume two types of events are tracked, searches and clicks. A search is followed by zero or more clicks on SERP links, organic results or ads; for our purposes we subsequently ignore web clicks. Corresponding searches and clicks are associated by means of a unique search identifier. All records contain the timestamp, browser cookie, and associated queries; we equate browser cookies with users. We refer to the set of searches and clicks issued by some browser cookie, within a given time period, as its *history*. In general, when using data from web search logs, we require that apparent spam and robot traffic has been filtered using standard detection algorithms.

3. PERSONALIZATION

We have seen in the previous section that click prediction is a central ingredient for many functions in sponsored search; therefore, this problem has been addressed in several articles, [15, 14, 5]. However, in contrast to the web

[1] For simplicity, in the following we ignore South ads.

search problem, personalization has not received a lot of attention in this context. If we somehow knew that a user was generally disinclined to click on an ad, independently of the relevance, we could improve his experience without losing revenue by showing him less ads. More generally, by showing less ads to users that are less likely to click, and more to those with commercial intent, we can keep the North footprint fixed and still receive more and better targeted clicks (or alternatively, depending on our choice of θ_{north}, reduce the North footprint and thus improve user experience without losing clicks). The theme of this paper is to utilize a user's search history for personalized click prediction.

The approach we are going to describe models users' propensity to click ads, while assuming that relative preferences of ads and of positions on the page remain unchanged across the users. More precisely, we refine the rank-normalized estimate of CTR for ad i as $coec_i \cdot ucp$, where the *user click propensity*, *ucp*, is a personalized factor determining how likely a user is to click on any ad, compared to an average user.

Note that *ucp* doesn't affect ranking or pricing, since all scores are scaled proportionally. However, it can very well be used to personalize the number of ads shown to a user, as well as the placement of these ads. For filtering, the necessary condition for being shown,

$$coec \cdot bid > eCPM_{min}$$

can be generalized to

$$coec \cdot ucp \cdot bid > eCPM_{min};$$

for North placement, the condition

$$\overline{ctr}_i \cdot coec_i \cdot bid_i > \theta_i$$

becomes

$$\overline{ctr}_i \cdot coec_i \cdot ucp \cdot bid_i > \theta_i.$$

Thus, for habitually low clickers the scores will be reduced, and might drop below the threshold as a consequence.

There are different ways to determine the personalization factor. We distinguish between long-term and short-term personalization; these aspects were dubbed "personalization" and "adjustment" in [18]. They are complementary to each other: the former one can e.g. capture the fact that some users pay more attention to ads in general, while others customarily skip to the web results right away. On the other hand, while someone is shopping for a digital camera, he might click on a number of ads; but once he actually bought one, his click rate could drop back to his lower, long-term average. After some preliminary exploration, we chose the period of the last 24 hours as "short-term" (we will call the corresponding factor ucp_{st}), and the last 28 days as "long-term" (ucp_{lt}); this seemed to be a reasonable trade-off between the amount of data and computational demand.

More precisely, we computed historical click propensity factors as follows. For a given user, we collect all of his search and click events going back up to a maximum fixed time period. For each viewed page, we can compute the *total* predicted probability of a click on any ad as

$$p(click) = \sum_{i=1}^{N} \overline{ctr}_i \cdot coec_i,$$

where there are N ads shown at positions $1, .., N$, and $coec_i$ is the prediction of the baseline, non-personalized click model for the i-th ad. By dividing the actually obtained clicks (within some time period) by the sum of these predictions, we can obtain an average click propensity for this user (for a well-tuned click prediction model, the average ratio should be one). To distinguish this ratio from the concept of COEC, which only refers to page position normalization, we will speak of *clicks over predicted clicks*, or *COPC*. In order to avoid large deviations in the case of sparse data, it is reasonable to smooth the *ucp* as

$$ucp = \frac{\sum_i click_i + click_0}{\sum_i p(click)_i + click_0},$$

where i runs over all search events for the user during the time period, $click_i$ and $p(click)_i$ are the observed and predicted clicks for search p, respectively. The constant $click_0$ corresponds to the weight of a prior, with a prior *copc* of one. Hence, a new user without history, or one with very little history, will have a *ucp* at or near one.

4. MACHINE-LEARNED PERSONALIZATION MODEL

The short-term clickability is still a very crude model: it only generates an overall average click propensity, but doesn't use available information about the exact timing within the history window, nor of the relationship between previous queries and the current one. Intuitively, if the user issues a query that is similar to one he issued before and on which he clicked an ad, he might be more likely to click on the current page, too.

To exploit relationships like this, we trained a prediction model with user-specific session features based on view and click events within the last 24 hours.

We attempt to capture query similarity using a number of syntactic overlap features. Let q^* denote the current query, and q_i an earlier query. We count the number of common words, $|q^* \cap q_i|$; the *word cosine distance* is defined as

$$wcos(q^*, q_i) = \frac{|q^* \cap q_i|}{\sqrt{|q^*| \cdot |q_i|}},$$

and the *word overlap* is

$$w\_overlap(q^*, q_i) = \frac{|q^* \cap q_i|}{|q^*|}$$

(note the asymmetrical normalization). We also define the measures *wprefcos* and *wpref_overlap* analogously by counting common *prefix* words (i.e., the maximum common words that occur in both queries in the same order, starting at the first word). Finally, we add features that count characters instead of words.

These measures are applied to the most recent query, the most recent clicked query, and the most recent non-clicked query (if existing). Moreover, we form *weighted click propensity factors* over all (clicked, non-clicked) previous queries in the history. For example, we define

$$wcos\_copc = \frac{\sum_i click_i \cdot wcos(q^*, q_i)}{\sum_i p(click)_i \cdot wcos(q^*, q_i)}.$$

This score is analogous to the overall click propensity factor of Sec. 3, but gives proportionally higher weight to more similar queries.

Other features we included are: average and current word and query lengths; total number of searches and clicks in history; number of searches, clicks, total $p(click)$ and *coec* for repeat queries; elapsed time (in seconds) since the last search and click. A non-obvious feature is *query session clickability, QSCB.* This is a hash table of relative click propensities that was computed computed offline over a period of a month. It is indexed by the current query, the total $p(click)$, and clicks in the preceding 24-hour window. To cope with data sparsity, the latter two were quantized into roughly equal-sized bins. For example, we found that if the current query is "att", the total *p(click)* over the last 24 hours is between 0.2 and 0.6, and no ads were clicked, the probability of a current ad click is reduced by a factor of 0.73.

Each instance of the training data consists of a page view, and contains all the input features described above. Our aim is to predict clicks, and find a click propensity factor that acts on top of the baseline (non-personalized) prediction of the click model *and* long-term user click propensity, $p(click)_i \cdot ucp_{lt,i}$. Under a squared loss function, we attempt to minimize

$$L = \sum_i (p(click)_i \cdot ucp_{lt,i} \cdot \hat{y}_i - click_i)^2, \qquad (1)$$

where $\hat{y}$ denotes the model output. This is equivalent to setting the target function to

$$t_i = \frac{click_i}{p(click)_i \cdot ucp_{lt,i}}$$

and then minimizing

$$L = \sum_i w_i \cdot (\hat{y}_i - t_i)^2,$$

with an instance weight of $w_i = (p(click)_i \cdot ucp_{lt,i})^2$ for example i. For convenience, we will refer to this model as ml_{st}.

Alternatively, we can also derive a *combined* model of long-term and short-term click propensity by feeding long term *ucp* into the model as an input feature. In that case, the baseline prediction consists only of the click model, and we minimize

$$L = \sum_i w_i \cdot \left(\hat{y}_i - \frac{click_i}{p(click)_i} \right)^2,$$

with an instance weight of $w_i = p(click)_i^2$. This model will be called ml_{slt}.

The modeling technique of *stochastic gradient-descent boosted trees* was chosen due to its robustness and good experience in web search modeling [10, 20]. In a preprocessing stage, we explored the space of learning parameters using five-fold cross validation; finally we chose the following parameters: number of trees: 60; shrinkage: 0.15; minimum samples: 500; sample rate: 0.5; BFS nodes: 8.

We generated the session features (described above) from web logs for all searches conducted by a fixed set of users over a period of one week. From this, we randomly sampled 1.5 million records for training data. This represented 1.1 million unique users [2]. The entire data set was used for training the model. In addition to this, we performed five-fold cross validation to evaluate model performance: five models were trained, on each subset of 4 folds. Predictions on a record were obtained from the model which was trained excluding the corresponding fold.

Model	**AUC**	**Rel. [%]**
$p(click)$ (no personalization)	0.4668	0.0
$p(click) \cdot ucp_{st}$	0.4795	2.7
$p(click) \cdot ucp_{lt}$	0.4895	4.9
$p(click) \cdot ucp_{slt}$	0.4942	5.9
$p(click) \cdot ucp_{lt} \cdot ml_{st}$	0.5130	9.9
$p(click) \cdot ml_{slt}$	0.5262	12.7

Figure 2: Area under Precision-Recall Curve for the Different User Prediction Models.

5. OFFLINE MODEL EVALUATION

Model development data was collected by sampling 1.5 million user actions (page views and corresponding clicks, if any) from the web logs, randomly split into five subsets. The GBDT model was trained using the entire data set. Model performance was evaluated using five-fold cross validation.

Fig. 3 compares the precision/recall curves for short-term and long-term click propensities; Fig. 4 also includes the two machine-learned models. (as usual, recall means the percentage of views with clicks that are correctly recognized by the model; precision is the percentage of clicks over all instances classified positive by the model). The points on the curve can be obtained by varying a threshold applied to the output score. Training and test sets were separated according to five-fold cross validation. The area under the precision-recall curve is summarized in Fig. 2.

Long-term user click propensity, ucp_{lt} clearly improves prediction over the baseline, particularly in the low-recall/high-precision range. The short-term click propensity[3], ucp_{st}, is significantly weaker than the long-term click propensity for prediction in the low-recall region, but slightly outperforms in the region of high recall. A combination of the two,

$$ucp_{slt} = \frac{\alpha_1 \cdot \sum_{i \in \mathbb{S}} click_i \; + \; \alpha_0 \cdot ucp_{lt}}{\sum_{i \in \mathbb{S}} p(click)_i \; + \; \alpha_2},$$

where $\mathbb{S}$ is the set of searches in the *short-term*, outperforms the long term click propensity across the entire range of recall[4].

Applying the short-term GBDT model together with ucp_{lt} fares much better, with about double the improvement compared to long-term effects alone (10% boost). The difference is most pronounced in the highest-precision region around 5% recall. Finally, the *combined* short-term/long-term GBDT model improves the accuracy the most (12%), however mostly in the low– and mid-range recall region.

Fig. 5 illustrates the score distributions. Roughly a third of all long-term user click propensities are close to one, due

[2] The average number of actions per user was reduced due to sampling. Features were generated before sampling to make the sampled data set representative of the general population of searches.

[3] To remedy data sparsity, we smoothed with a prior of $click_0 = 1$.

[4] The parameters $\alpha_0, \alpha_1, \alpha_2$ were set by minimizing a loss function similar to Eq. 1 on an independent set of the sampled records using `MATLAB`'s `nlinfit` procedure.

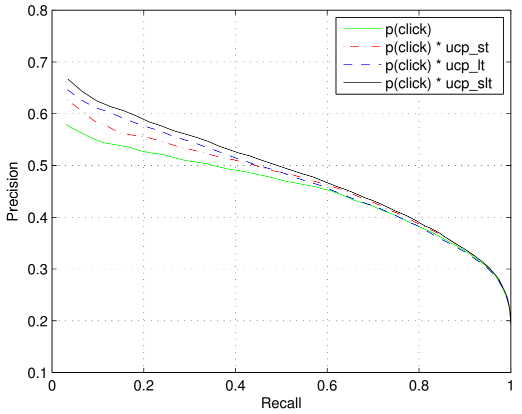

Figure 3: Precision-Recall Curves for Short-Term and Long-Term Click Propensity.

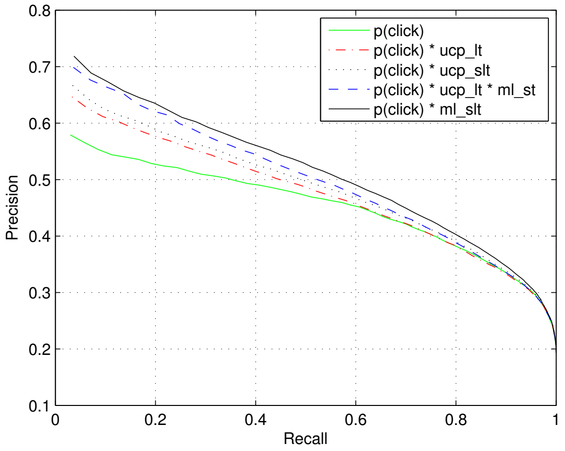

Figure 4: Precision-Recall Curves for Click Prediction.

to missing or little recorded history. Apart from this sharp peak, the distribution spreads out smoothly, with a thin, long tail. The distribution including the short-term GBDT model ml_{st} has a similar shape, except that it differentiates more for users with less history; the peak is spread out. Curiously, the combined model has a much more compressed score distribution, extremely small or large click propensities are much rarer. Fig. 6 examines this relation more closely: at an operating point around one, the two are more or less proportional, however the influence flattens out at around ucp_{lt}=1.5. This might tell us that just applying a proportional factor is too simplistic – there could be other effects interfering, e.g. query selection bias.

An instructive feature of the gradient-boosted tree technique is that it allows estimation of relative variable importance; it is based on the normalized sum of the reduction in squared error, from every split point across all features [10]. The 20 most important variables for the two models are listed in Figs. 8 and 9.

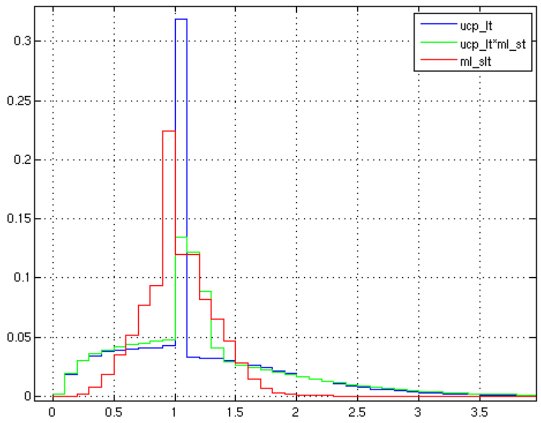

Figure 5: Frequency of Click Propensities, by number of searches.

In line with previous studies [16, 13], we found that during the 24-hour period about 27% of queries were repeated at least once. Multiple reasons have been cited for that, such as "bookmarking" and navigational behavior, but also page refreshes and browser back-button clicks. Therefore, it is not a surprise that the features derived from user behavior on the same query (such as total *p(click)*, *copc*, etc) rank very prominently.

For the short-term model ml_{st}, the second most important feature is the time since last click, suggesting that once a user is clicking on an ad, he is more likely to click on other ones in short succession, as one might expect for shopping/browsing behavior. LATEST_WPREFCOS and LATEST_C_OVERLAP measure the similarity to the previous query, helping to decide relatedness.

Not surprisingly, for the combined model ml_{slt}, long-term ucp is the most important feature, by a large margin. The three next important features UCP_ST, WCOS_COPC, and SAME_COPC denote the relative click propensities over all searches in the most recent 24-hour window (short-term user click propensity), over all those searches but weighted with the cosine-distance, and only over identical queries, respectively.

We could summarize our observations as follows. Long term attitude of users towards ads in general have large influence on click behavior; this can be easily captured by keeping track of the relationship between predicted and observed clicks. Most recent history (say, within a day) is an equally important predictor. However, to exploit that, we need to adopt a more granular view of the specific queries and timing; for example, a large fraction of queries are repeated, and previous behavior can be projected into the future. Due to interactions between these two components, such that a model using both together as input features performs the best.

6. LIVE TRAFFIC EXPERIMENTS

In this section, we summarize a number of experiments we ran on a random sample of US live search traffic of a commercial search engine. Each experiment was conducted

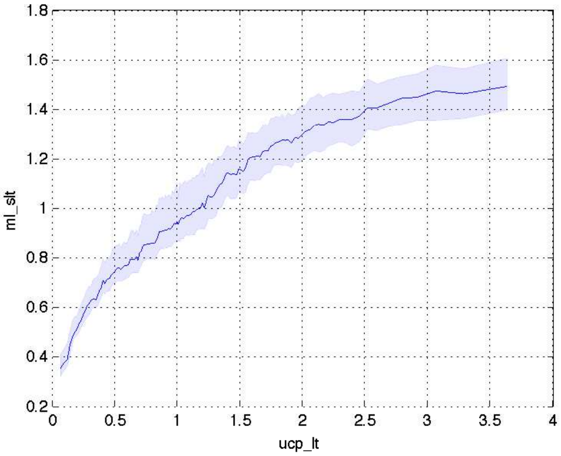

Figure 6: Median and Inter-Quartile Range of Combined Model Score, as a function of long-term click propensity.

over a period of one week. Users were assigned to an experiment randomly based on a hash of their browser cookie. To ensure a consistent experience, this assignment remained fixed for the duration of the experiments. On average, about one million searches per day were issued by 250,000 users for each experiment. Result metrics are compared to a baseline experiment of equal volume that was run simultaneously.

Long-term click propensities were computed offline from historical web logs, and then supplied to the serving system through a lookup table. To record user-specific short-term features, we relied on a novel, experimental stream processing framework called *S4* (we will give details about *S4* in a separate, forthcoming paper). This system allows for user feedback in almost real time, and thus makes adaptation perceptible: for example, a user clicking on ads for a query could see more (North) ads when he repeats the same query a few seconds later.

Figure 10 gives a summary of the results. All numbers in the table are percentage differences with respect to the corresponding baseline experiment. The metrics are defined as follows: *FP (footprint)* – average number of shown ads per page view; *NFP (North footprint)* – average number of shown North ads per page view; *CY (click yield)* – total clicks over total page views; *PPC* – average price per click; *RPS (revenue per search)* – total revenue over total page views.

Clearly, any change in the way we determine page placement scores will result in a different *NFP*. We can either measure this difference, or else retune θ_{north} to match the number of ads shown in the baseline, and compare the difference in click metrics. We did the first in the case of Experiment 1, and the latter for the remaining experiments, which use Experiment 1 as a baseline to reveal incremental changes.

Experiment 1 uses the long-term user click propensity factor, ucp_{lt}, placement in the way explained in Sec. 3. With a fixed threshold, this reduces the North footprint by about seven percent; however, despite showing significantly less North ads and thus improving user experience, no clicks or revenue were lost (in fact, there is even a small, though not statistically significant gain).

Description	Variable Name(s)
Syntactic features of current query	C_LEN, W_LEN
User click propensity (short/long term)	UCP_ST, UCP_LT
Temporal features and count of historic user actions	TIME_SINCE_LAST_CLICK, TIME_SINCE_LAST_SEARCH, H_CLICK, H_COUNT
Query Session Clickability lookup	QSCB
Predicted clicks on past searches	AVG_PCLICK
Similarity of current query to past queries. Number of times current query was searched in the past.	LATEST_WPREFCOS, LATEST_C_OVERLAP, MAX_W_PREF_OVERLAP, SUM_CPREFCOS, COUNT_SAME
Predicted clicks from previous identical or similar queries	SAME_PCLICK, C_OVERLAP_WEIGHTED_PCLICK, WPREFCOS_WEIGHTED_PCLICK, WCOS_WEIGHTED_PCLICK
Observed clicks and non-clicks on previous searches, and similarity of current query to queries which were clicked or not clicked. Number of times same query was clicked.	LAST_CLICKED, LATEST_NONCLICK_C_OVERLAP, MAX_C_OVERLAP_CLICK, C_OVERLAP_WEIGHTED_CLICK, LATEST_NONCLICK_WPREFCOS, LATEST_CLICK_WCOS, SAME_CLICK, LATEST_NONCLICK_WCOS, WCOS_WEIGHTED_CLICK, CPREFCOS_WEIGHTED_CLICK, MAX_W_PREF_OVERLAP_CLICK
Ratio of observed clicks to predicted clicks (COPC) for entire history, latest search, and similar searches.	SAME_COPC, WCOS_COPC, C_OVERLAP_COPC, CPREFCOS_COPC, WPREFCOS_COPC

Figure 7: Description of Input Features.

Experiment 2 employs ucp_{lt} to personalize both placement and *filtering* of ads. We reduce the average number of shown ads by seven percent, with neutral click yield and a small penalty in revenue (despite out attempt to fix *NFP*, we observed a small unintended increase, so that the real revenue drop might be slightly larger).

In Experiment 3, we apply a weighted combination of the short-term and long-term *ucp*, where the weight depends on the number of searches in the most recent 24-hour window. This allows us to roughly double the footprint reduction (15%) of Experiment 2, with similar click and revenue metrics.

Experiments 4 and 6 shows that the increased prediction accuracy of the GBDT models ml_{st} and ml_{slt}, as demonstrated in Sec. 5, indeed translates into better live traffic metrics. While showing roughly the same number of total

Exp.	Description	Personalized Placement	Personalized Filtering	Baseline	FP	NFP	CY	PPC	RPS
95% confidence interval					±0.34%	±0.28%	±0.57%	±0.95%	±1.29%
1	Long-term ucp	yes	no	No Pers.	-0.60%	-7.43%	0.49%	-0.18%	0.31%
2	Long-term ucp	yes	yes	Exp. 1	-7.21%	1.22%	0.22%	-0.92%	-0.70%
3	Weighted average of long-term and short-term ucp	yes	yes	Exp. 1	-15.22%	0.20%	-0.25%	-0.76%	-1.01%
4	model ml_{st}	yes	no	Exp. 1	-0.21%	0.02%	1.02%	-0.66%	0.35%
5	model ml_{st}	yes	yes	Exp. 4	-8.72%	0.55%	1.18%	-0.17%	1.01%
6	model ml_{slt}	yes	no	Exp. 1	-0.62%	0.88%	1.81%	-1.07%	0.73%

Figure 10: Summary of live traffic results.

rank	feature	importance
0	SAME_PCLICK	100
1	TIME_SINCE_LAST_CLICK	68.7462
2	LATEST_WPREFCOS	67.4443
3	SAME_COPC	66.2287
4	LATEST_C_OVERLAP	64.7187
5	QSCB	63.268
6	WCOS_COPC	60.7432
7	W_LEN	56.7256
8	TIME_SINCE_LAST_SEARCH	56.4712
9	WPREFCOS_WEIGHTED_PCLICK	54.9567
10	C_OVERLAP_COPC	48.2477
11	SUM_WCOS_NONCLICK	46.4606
12	UCP_ST	45.8821
13	MAX_W_PREF_OVERLAP	38.0269
14	MAX_WCOS_NONCLICK	37.9318
15	AVG_PCLICK	32.878
16	C_LEN	32.7651
17	MAX_CPREFCOS_NONCLICK	30.2747
18	LAST_CLICKED	28.0465
19	LATEST_NONCLICK_C_OVERLAP	26.6513

Figure 8: Relative Variable Importance of ml_{st} model (top 20).

rank	feature	importance
0	UCP_LT	100
1	UCP_ST	47.9298
2	SAME_COPC	38.3466
3	WCOS_COPC	34.6456
4	QSCB	30.3276
5	SAME_PCLICK	30.3177
6	MAX_WCOS_NONCLICK	27.2795
7	TIME_SINCE_LAST_CLICK	26.7955
8	W_LEN	24.5511
9	LATEST_C_OVERLAP	23.5178
10	C_OVERLAP_COPC	23.2873
11	TIME_SINCE_LAST_SEARCH	22.592
12	SAME_PCLICK_RATIO	22.4506
13	H_CLICK	19.0222
14	WPREFCOS_WEIGHTED_PCLICK	18.338
15	LATEST_W_PREF_OVERLAP	17.8811
16	SUM_WCOS_NONCLICK	16.3379
17	C_LEN	15.3651
18	LATEST_CPREFCOS	13.3816
19	CPREFCOS_COPC	12.9754

Figure 9: Relative Variable Importance of ml_{slt} model (top 20).

and North ads, we obtain one resp. 1.8 percent more clicks, without revenue loss. Note that the drop in price, together with the increase in clicks, is overall beneficial for advertisers too in terms of cost-per-acquisition (assuming constant conversion rates).

In line with the accuracy predictions of Sec. 5, the live traffic results also confirm that explicitly using the long-term user click propensity as an input feature (ml_{slt}) outperforms the factorized model.

Experiment 5 is equivalent to experiment 4 with model ml_{st} for page placement; however, it additionally performs personalized *filtering*. Analogously to the pair of Experiments 1 and 2, we can achieve a nine percent reduction of shown ads, without significant revenue loss.

7. RELATED WORK

Approaches to incorporate user behavior as implicit feedback for web search ranking have been reported earlier, e.g., [7, 1, 2, 12, 9, 18]. On the other hand, similar applications in *sponsored search* have received limited attention. One exception is a recent study that aims to shed some light on *post-click* user behavior [3].

The most related work to ours is that of Chen and Cantú-Paz [4], who aim at extending the baseline, *per-ad* click model by incorporating user features. A number of different user–, user-group–, and demographic properties are proposed, as well as their conjunction with queries and ads. A wealth of interesting findings are reported. For example, there are strong correlations between search/click behavior and demographics. However, if behavioral metrics are already used as features in the click prediction model, there is little gain in accuracy by using demographics *in addition*; the former category captures most of the relevant information of the latter.

There are two main differences between this work and ours. First, we use a coarser abstraction level, not distinguishing between individual ads (therefore, the AUC numbers are also not directly comparable). While this aggregation necessarily sacrifices some possible accuracy, in practice the increase in confidence through data density, and the practical considerations for storage and computation, seem to outweigh this disadvantage. In fact, the comparison of models including different types of features in [4] suggests that the inclusion of user-query features alone into the baseline model lead to only slightly less improvement than user-query-ad features, and slightly better than user-ad features alone.

Secondly, Chen and Cantú-Paz only evaluated user features derived from a period of 2 months of data; short-term or real-time adaptation is not considered.

8. CONCLUSION

In this paper, we described a novel approach to adapt the sponsored search functions of *page placement* and *filtering* to an individual user's preferences between organic search results and advertisements. While a host of previous research has addressed personalization for web result ranking and filtering, personalization of ad presentation has received little attention, and we hope to open up this discussion. The idea is to show less ads to ad-averse users and more to shoppers, thus improving overall satisfaction with the search engine for both groups. At the same time, advertisers benefit by receiving more clicks from users that are more engaged with the ads, and hence more likely to conduct a conversion on their site.

We proposed and compared several approaches based on user history: long-term and short term user click propensity, and machine-learned models that take into account user actions on a more granular level. We have shown that these models significantly increase the accuracy of click prediction on historical search log data. Our live traffic results indicate that personalization allows us to show significantly less ads in less prominent places, without losing clicks or revenue; and to gain more and better targeted ad clicks.

There are several promising directions for further research. More refined query similarity measures could capture reformulations beyond simple word overlap, e.g., use of synonyms. Dwell time features can indicate a user's satisfaction with landing pages. The textual contents of previously clicked results could be utilized, in addition to the short query string. While clicks on web results were not accessible to us in our implementation, these actions could provide additional, valuable cues to user intentions.

9. REFERENCES

[1] 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. ACM Press, 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10. ACM Press, 2006.

[3] J. Attenberg, S. Pandey, and T. Suel. Modeling and predicting user behavior in sponsored search. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1067–1076, New York, NY, USA, 2009. ACM.

[4] H. Chen and E. Cantú-Paz. Personalized click prediction in sponsored search. In *In Proceedings of the 15th International World Wide Web Conference (WWW-10*, April 2010.

[5] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 227–236, New York, NY, USA, 2008. ACM.

[6] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 87–94, New York, NY, USA, 2008. ACM.

[7] Z. Dou, R. Song, J.-R. Wen, and X. Yuan. Evaluating the effectiveness of personalized web search. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1178–1190, 2009.

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

[9] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. In *ACM Transactions on Information Systems*, number 23, 2005.

[10] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[11] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 417–420, New York, NY, USA, 2007. ACM.

[12] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7, 2007.

[13] B. Piwowarski and H. Zaragoza. Predictive user click models based on click-through history. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 175–182, New York, NY, USA, 2007. ACM.

[14] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors, *WWW*, pages 521–530. ACM, 2007.

[15] B. Shaparenko, O. Çetin, and R. Iyer. Data-driven text features for sponsored search click prediction. In *ADKDD '09: Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 46–54, New York, NY, USA, 2009. ACM.

[16] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR*, pages 151–158. ACM, 2007.

[17] H. R. Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, December

2007.

[18] S. Wedig and O. Madani. A large-scale analysis of query logs for assessing personalization opportunities. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 742–747, New York, NY, USA, 2006. ACM.

[19] J.-R. Wen, Z. Dou, and R. Song. *Encyclopedia of Database Systems*, chapter Personalized Web Search. Springer-Verlag, New York, September 2009.

[20] J. Ye, J.-H. Chow, J. Chen, and Z. Zheng. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 8th ACM Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, November 2009.

[21] W. V. Zhang and R. Jones. Comparing click logs and editorial labels for training query rewriting. In E. Amitay, C. G. Murray, and J. Teevan, editors, *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*, May 2007.