Generalized Utility in Web Search Advertising

Stefan Schroedl, Anandsudhakar Kesari, Anish Nair, Leonardo Neumeyer, Sharath Rao
Sponsored Search Sciences
Yahoo! Labs
4401 Great America Pkwy
Santa Clara, CA 95054
{stefan,anands,anishn,neumeyer,skrao}@yahoo-inc.com

ABSTRACT

In response to *user* queries, commercial web search engines, such as Yahoo!, Google, or Bing, show organic (e.g., web) results on their portals or other *publisher* sites. Most of their revenue originates from showing ads along with these results, by letting *advertisers* bid on possible user keywords. Long-term success depends on delivering and appropriately balancing utility for all of these three groups of participants.

We present some formalizations of utility functions that go beyond instantaneous revenue and include the cost of compromising user experience. We make some simplifying assumptions to implement algorithms for *ranking* (ordering of ads), *filtering* (exclusion of less relevant ads), and *page placement* (deciding how many ads to show on top of the organic results). While utility functions have been proposed in the literature before (e.g. [6]), it was typically applied only to ranking; to the best of our knowledge, the problem of filtering and page placement in web search advertising has not been addressed in depth in the current literature.

Finally, we report experimental results on random live traffic from a commercial search engine, which exhibit significant improvements in behavioral metrics.

1. INTRODUCTION

The prevalent auction model in current web search advertising is based on the *generalized second price (GSP)* principle [5, 17]. Advertisers bid on search terms relevant to their products. They have a chance to get shown on the *search engine result page (SERP)* when users type in these queries, along with competing ads and *organic* (web) results[1]. The ads are ordered according to a function *rankScore* of bid and a proprietary *ad quality score*. A typical implementation is to use a position-normalized estimate of *click-through-rate* as quality score, and define the rank score as the product of this and the bid. When the user clicks on an ad, the advertiser has to pay the minimum bid necessary to retain his rank according to this ordering [12].

While this auction process is geared towards maximizing the search engine's expected instantaneous revenue, its long term success also hinges on its utility for all three groups of participants:

Users visiting the search engine for quick navigation, to find information, or perform transactions.

Publishers are web sites displaying the search results. In addition to the search engine portal itself, third parties can generally enter a revenue-sharing agreement to show the results on their site.

Advertisers try to maximize their return on investment by obtaining a large volume of clicks with lower price than their expected value of conversion.

Some generalized utility frameworks have been proposed earlier [6, 1]. The challenge lies in translating such models into practically feasible adaptations to our current web search advertising system. These changes affect a variety of functions that we summarize under the name *Search Advertising Optimization*:

Ranking orders the ads according to a score of estimated utility.

Pricing determines the cost advertiser have to pay for a click.

Filtering decides *which* ads out of all candidates are eligible to be shown for a given query.

Page Placement determines *where* to show ads on the SERP: above (a.k.a., *"North"* of) the organic results, in a separate column to the right (*"East"*), or below them *("South"*); see Fig. 1.

Out of these functions, most existing literature concentrates on ranking alone. However, it is well known (e.g., from eye-tracking studies [9, 4]) 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. Even if we have given a set of ranked ads, we still have to decide whether they deserve to be placed on top of the web results, or rather should be placed in the less prominent space on the right. 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

[1]For simplicity, in this paper we ignore other types of links like vertical search results, shortcuts, spelling suggestions etc.

ADKDD'10, July 25, 2010, Washington D.C., USA.

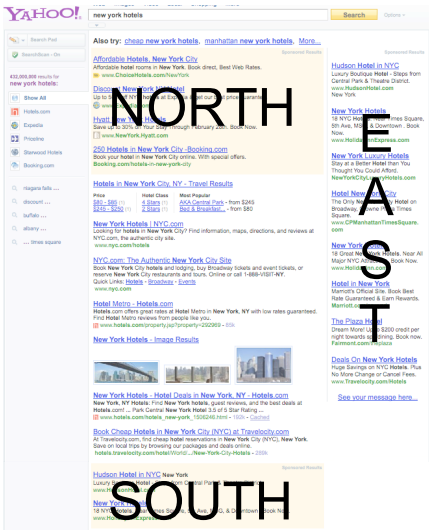

Figure 1: Search Result Page with Ad Placement Regions.

much more severely and might subsequently prevent him from clicking on other ads or from returning to the search engine altogether. To the best of our knowledge, we could not find any publications addressing this problem explicitly and we hope to initiate this discussion here.

The overarching goal of this paper is to explore solutions to *all* sponsored search optimization problems that balance search engine revenue and user experience in a meaningful way. In the next section, we start by briefly reviewing utility from the viewpoints of advertisers, publishers, and users, in turn. Sec. 3 particularly focuses on incorporating *user* utility. For *ranking and filtering*, we cover the necessary position normalization (Sec. 3.1), and then build on Abrams and Schwarz' [1] *Hidden Cost Model* (Sec. 3.2). Sec. 4 formalizes the problem of page placement as a constraint optimization problem. While the exposition up to this point is still general enough to accommodate various concrete methods of estimating utility, we subsequently concentrate on some special cases that are however amenable to implementation within a current commercial search engine. Based on the widely used *DCG* relevance metrics [13], we define the *North Ad Impact* criterion (Sec. 4.2 and 4.3). We also develop a new scheme of separate, cascading auctions for different SERP regions (Sec. 5) that can be used to adapt both ranking and page placement. Finally, in Sec. 6, we summarize various live traffic experiments, and report behavioral metrics results that show significant improvements.

2. UTILITY FUNCTION

Let us briefly review the utilities of the different participants in search advertising.

2.1 Advertiser Utility

Advertisers try to maximize the number and values of *conversions* (e.g., purchases). The relative conversion frequency, times the value per conversion, should exceed the price per click (*PPC*) to be profitable. There is usually a complex and only partially known relationship between the volume of clicks and the bid – higher bids can secure more prominent placement, but will tend to increase cost dependent on competing ads. Also, the composition of queries and search traffic has a significant effect on conversion rate; the user populations visiting different publisher sites can have different characteristics. The actual value per click estimated by advertisers is not directly accessible to the search engine; however, under the assumption of informed advertisers in a symmetric Nash equilibrium, bounds can be derived based on the ranking; moreover, ranking by bid times position-normalized click-through rate guarantees maximization of advertiser value [17].

2.2 Publisher Utility

Ultimately, publisher revenue is a share of the total conversion value. There are two ways to raise it: by taking share away from advertisers (e.g., through modifications of the *rankScore* function), or by striving to create more value for advertisers, and then indirectly profiting. Surely the latter option is the more sustainable one in the long term. Methods can be employed to increase CTR (with no conversion rate change, i.e., excluding 'accidental clicks'), to increase conversion rate (e.g., by better ad targeting to user searches), or both. Since we can safely assume that converting users are satisfied, increasing the conversion rate is good for all three groups.

2.3 User Utility

We have seen that (under some assumptions) giving most of the clicks to the advertisers with highest ranking score improves their utility (and that of the publisher). However, most users visit the search engine primarily for organic results, so a natural conflict in short term utility arises. In the long term, bad ads that tarnish overall user experience and keep them from coming back are detrimental to the search engine as well. We will focus on user utility in the remainder of the paper.

The ultimate measure of user utility is *task completion*: he finds the information or web address he was looking for, or a site to execute an intended transaction. It can only be credited to the entirety of the user's goal-related actions, generally not to an individual search result. Without explicit feedback, task completion is hard to recognize. Nevertheless, a number of implicit measures have been developed that correlate to some degree, e.g., click-through rate, landing page dwell time, scrolling actions, conversions, etc [7, 14].

For these reasons, all frameworks have to make strong simplifications in their user models. Some common assumptions are:

- The user's task-related activity consists of a single search.
- There is at most one click per SERP, and it determines success.
- Results are examined independently of each other.
- The utility of the result list can be decomposed into a sum of individual result utilities.

Suppose that each search result can be clicked; subsequently, the user examines the landing page, which may or

may not be relevant. Conditional on these mutually exclusive cases of user behavior (not seen, not clicked, relevant landing page), we can define the expected user value of an ad as

$$\begin{aligned} U_{user} &= (1 - p(click)) \cdot U_{distr} + \\ &\quad p(click) \cdot [p(rel) \cdot U_{rel} + (1 - p(rel)) \cdot U_{irrel}] . \end{aligned}$$

- U_{rel} is the utility of finding a relevant result - obtaining desired information, navigating to an intended site, or being able to perform a transaction such as a purchase.
- U_{irrel} is the (negative) utility of the user recognizing that a clicked page does not meet his expectations - the click was essentially wasted.
- U_{distr} is the (slightly less negative) utility due to visual distraction and annoyance, e.g. through lowering the probability of noticing other, more relevant results.

The parameters U_{distr}, U_{rel}, and U_{irrel} have orthogonal definitions and can be estimated independently. They could be refined as a function of the ad, position, session, etc: U_{rel} may be higher for more expensive markets; U_{irrel} could be made low for an informational session, and high for commercial sessions; U_{distr} may decrease with rank, and can be a function of the estimated relevance of the shown abstract to the query.

We can define the total value of an ad as the sum of the *declared* advertiser value, and the estimated user value:

$$\begin{aligned} U_{total} &= p(click) \cdot bid + p(click) \cdot p(rel) \cdot U_{rel} + \\ &\quad p(click) \cdot (1 - p(rel)) \cdot U_{irrel} + \\ &\quad (1 - p(click)) \cdot U_{distr}, \end{aligned}$$

which is equivalent to

$$\begin{aligned} U_{total} &= p(click) \left[bid + p(rel)(U_{rel} - U_{irrel}) + \right. \\ &\quad \left. (U_{irrel} - U_{distr}) \right] + U_{distr}. \end{aligned} \quad (1)$$

3. INCORPORATING USER UTILITY FOR RANKING AND FILTERING

3.1 Position Normalization

In an auction for a *single* display slot, we could use a direct estimate according to Eq. 1 and let the result win which has the highest such score. However, this is not directly applicable to the multi-slot case; placing the same ad in a more prominent spot of the SERP will garner more clicks, everything else being equal. We need a *position-normalized* score S so that for all ads a, a' and positions i, $U_{total,i}(a) > U_{total,i}(a')$ whenever $S(a) > S(a')$. It is convenient to assume that the positional effect can be factored both for the click probability and for U_{distr} in the same way:

$$\begin{aligned} p(click|position, ad, \dots) &= f(ad, \dots) \cdot g(position) \\ U_{distr} &= h(ad, \dots) \cdot g(position). \end{aligned}$$

One interpretation is that g is the probability of the user noticing the ad altogether while scanning the page; this probability is assumed to decrease with rank, independently of the actual ads.

Under this assumption, it is feasible to use a ranking score that is similar to Eq. 1, except that the position-independent estimates $p(click|seen)$ and $U_{distr|seen} = U_{distr}/p(seen)$ replace $p(click)$ and U_{distr}, respectively. Engel and Chickering [6] describe a similar framework.

Ranking based on this score results in maximizing the sum of all expected result values. *Filtering* arises as a natural consequence: It is better not to show a result with negative expected value at all.

It should be mentioned that recently there has been a flurry of research to overcome the limitations of the pure rank-based position normalization, and to incorporate externalities between ads (e.g., [10, 2, 15, 3]).

3.2 Hidden Cost Auction Model

Abrams and Schwarz [1] proposed a generalization of the *GSP* auction mechanism that is able to incorporate a *Hidden Cost*. This term models the time-discounted future loss of revenue that results from a user looking at an irrelevant, clicked result: a bad experience can make the user less inclined to click on ads in the future. Hence, when ranking ads according to estimated revenue, this future externality should be incorporated by subtracting it from the short-term declared value. In principle, with a vast amount of historical data, a large number of users, and long lifetimes of ads, it would be possible to precisely determine this cost, by comparing average click-through rates of users before and after having seen a given ad. While such an undertaking is infeasible in practice, because of various issue like infrequent queries and inventory turnover, the framework is still useful to incorporate any estimate of user utility into the second-price auction process.

Recall that ads are ranked by the product of bid and quality score, $rankScore = bid \cdot q$, where q can typically be understood as a position-normalized estimate of click-through-rate. Then, the rationale for *pricing* is to charge the advertiser the minimum bid that would be necessary to maintain his rank (plus a small minimum increment). Ignoring the latter, for all but the last-ranked advertisers i this results in

$$cost_i = \frac{rankScore_{i+1}}{q_i} = \frac{bid_{i+1} \cdot q_{i+1}}{q_i}.$$

Now, in the generalized framework we subtract the hidden cost hc from the declared value, and rank by the net value

$$rankScore' = bid' \cdot q = (bid - hc) \cdot q.$$

To maintain his rank, the advertiser at rank i then has to pay a price per click

$$cost_i = \frac{rankScore_{i+1}}{q_{i+1}} + hc_i;$$

i.e., advertisers are allowed to compete only based on the net bids (actual bid minus hidden cost), and are additionally charged the hidden cost for clicks.

Comparing with Eq. 1, we can equate the hidden cost with the negative expected value of the landing page:

$$hc = p(rel)(U_{irrel} - U_{rel}) + U_{distr} - U_{irrel}. \quad (2)$$

The comparison also shows that the hidden cost is associated with a cost *per click* based on the *landing page*. If q is interpreted as a click-through rate, then *rankScore* would be measured in monetary units *per impression*; and so would

be the additional term in Eq. 1 for SERP annoyance, which is intuitively plausible: the lower the click-through rate, the higher the cost that will be charged for each click.

Although Abrams and Schwarz [1] were mostly concerned with the negative user experience after clicking on an irrelevant ad, it is straightforward to accommodate the hidden cost per-impression:

$$cost_i = \frac{rankScore_{i+1} + hc_{imp,i}}{q_{i+1}} + hc_{click,i},$$

with $hc_{click,i}$ defined as before, and an additional term $hc_{imp,i} = -U_{distr}$.

4. INCORPORATING USER UTILITY FOR PAGE PLACEMENT

Advertisements on top of organic results (rather than in a separate, right-hand column) directly compete with the latter ones for space. For some commercial search terms, ads can be more attractive than web results, but more frequently, they can divert attention and keep users from reaching pages with the requested information. Real-world search engines deliberately risk degradation of user experience in exchange for expected revenue.

Usually, ads and web results cannot be freely mixed, only in North, East, and South Blocks. Ads not shown in the North can still be shown in the East or in the South; however, the bulk of both user impact and revenue stems from the North. Therefore, subsequently we focus on North placement.

Deciding ad placement is a task of integrating two completely separate search engines. In principle, we could find appropriate parameters for Eq. 1 (higher relevance probabilities and utilities for web results should compensate for the lack of revenue) and then optimize this score for each search. Out of a set of possible *slates* of ad and web results (e.g., showing 0, 1, ... North ads), we could choose the one that maximizes the total expected utility. However, these parameters are hard to estimate accurately. Just relying on the utility estimate for page placement would make it hard to influence parameters like e.g. the total number of ads shown directly. Therefore, real-world search engines try to get more fine-grained control over the trade-off between revenue and user utility.

4.1 Page Placement Algorithm

Typically, a target of estimated user utility reduction can be fixed, and then an ad allocation algorithm can be tuned to optimize estimated revenue. In contrast to maximizing a utility function separately for each query, we pose page placement as a *constrained* optimization problem of maximizing overall revenue, given an overall budget of maximum user utility reduction, across all searches.

First, suppose we can run an *offline* simulation based on a sample of N historical user searches from server logs, together with the corresponding ranked lists of web and ad results. For ease of exposition, equate user utility with (negative) average number of ads per page, n_{ad} (also called the *North Footprint, NFP*); so we can allocate $n_{ad} \cdot N$ total ads across the given searches. If we could place any available ad, a simple greedy algorithm would find the maximum expected revenue by choosing the top $n_{ad} \cdot N$ ads in order of decreasing estimated revenue (more precisely, $p(click) \cdot bid$). Of course, in reality there are constraints to be observed:

- Ad placement has to be consistent with ranking.
- An ad at rank i can only be placed if $i = 1$, or ad $(i-1)$ has been allocated already.
- There is a maximum number of ads that can be shown on a given page.

The greedy algorithm can be refined to take care of these dependencies by using a lookup table mapping each search to the next eligible ad. If the ranking score is different from the page placement objective, this strategy is not guaranteed to find the optimal result any more, but in practice we have seen acceptable performance.

In contrast to the offline scenario, the server has to make the page placement decision instantly at the time of a new user request. Note that the placement would have been the same as if it had been done online by the server executing the following algorithm: allocate each placeable ad whose expected revenue is at least as large as that of the $(n_{ad} \cdot N)$-th top ad.

Now, as a generalization, suppose we want to maximize revenue under a constraint of an arbitrary utility function, not only the total number of ads as assumed above. The offline optimization problem turns out to be at least as complex as the *NP*-hard *0-1-knapsack problem* [8] (it is a proper instance if we again drop the ranked allocation constraints):

0-1-Knapsack Problem:
Maximize $\sum_{j=1}^{n} p_j\, x_j$
subject to $\sum_{j=1}^{n} w_j\, x_j \leq W, \quad x_j \in \{0,1\},$

where the weights w_i correspond to user utility, W to the total budget, and the p_i to estimated revenues.

Despite its intractability, many practical instances of this problem can be solved using *dynamic programming* over the space of all searches and ad impressions. However, we are not only interested in offline optimization, but also in a real-time version of it[2]. Therefore, we have to stick to a greedy approximation. One possible heuristic is to order the items by either p_i, $p_i - \alpha * w_i$, or p_i/w_i (in our case, expected revenue over expected user utility), and then allocate eligible ads until the utility budget is exhausted.

4.2 DCG and North Ad Impact

Up to now, we have outlined a formal framework for incorporating user utility into Search Advertising Optimization. However, we have not given details about how to accurately determine the involved parameters, U_{rel}, U_{irrel}, and U_{distr}. In general, this is a complex problem in itself, and a conclusive treatment is beyond the scope of this paper. It could be based on psychological models of user perception; alternatively, we could develop economical prediction models by trying to quantify the impact of seen ads on future revenue from the user. We are planning to investigate such effects in subsequent work. However, for the time being, we resort to some simple but more easily measurable proxies. The remainder of the paper describes three such specializations that lend themselves more readily to live traffic experiments.

[2]The serve-time algorithm can only examine the ads selected for the particular search, whereas the offline dynamic program depends on all ads served over a period of time.

One way of measuring the web search retrieval quality that has become somewhat of an industry standard is the *Discounted Cumulative Gain (DCG)* [13]. It is a weighted sum of the per-item relevance scores (according to human judges) of the top p returned documents, where the weight is a decreasing function of the rank:

$$DCG_p = \sum_{i=1}^{p} w_i \cdot rel_i. \quad (3)$$

This formula is commonly used with a non-linear relevance scale that attempts to capture user satisfaction. The reasoning behind the weights is that according to behavioral studies, users spend a limited amount of effort on scanning the SERP, with most of their attention focused on a top left triangle. A popular choice for the position weight is $w_i = 1/\log_2(1+i)$.

We can view DCG as a special case of Eq. 1 as follows: the position weights w_i correspond to $p(click) \sim const \times p(seen)$, i.e., the user clicks blindly on results with a probability decreasing with the rank. The relevance score rel_i is an aggregate estimate of the expected total post-click value, $(p(rel) \cdot U_{rel} + (1 - p(rel) \cdot U_{irrel})$, and $U_{distr} = 0$.

For a given SERP consisting of ads and organic results, we can use this measure to determine relevance degradation. *North Ad Impact (NAI)* is defined as the difference in DCG_p for the whole page (including top ads), and for the same page with the ads removed. A positive *NAI* corresponds to a decrease of relevance, while a negative *NAI* means that the ads are in fact more relevant.

4.3 Using NAI for Page Placement

Let us define the *incremental NAI* of the ad at rank k as the difference of the DCG_p of the SERP with ads $1, \ldots, k-1$ shown, minus that of the same page with ads $1, \ldots, k$:

$$NAI_k = \sum_{i=1..p-k} (w_{k+i-1} - w_{k+i}) \cdot rel_{web,i} + w_p \cdot rel_{web,p-k} - w_k \cdot rel_{ad,k} \quad (4)$$

That is, all web results get pushed down one rank (and thus suffer a loss in *DCG* weight) except for the last one at rank p, which gets eliminated.

Editorial data is sparse and expensive, so we have to use approximations and predictions for web and ad relevance. We can cheaply make a zero-th order approximation towards constraining North Ad Impact without even having individual relevance judgments as follows. We modify the basic algorithm described in Sec. 4.1 to constrain, rather than the average number of North ads per search, a *weighted* average, with the weights taken from the *DCG* formula (3). This is equivalent of assuming the same (lower) relevance for all ads on the one hand and for all web results on the other hand, and then inserting these averages into Eq. 4. Under these simplifying assumptions, our tuning target would indeed be the North ad impact.

Earlier live traffic experiments showed us that this modified page placement algorithm increased revenue and North footprint; however, an evaluation by human editors confirmed a neutral *NAI*. The distribution of the number of North ads changes, reducing the number of searches with only one or two ads shown in the North, and increasing the percentage of searches with the maximum possible number (four). User experience is less impacted by showing four instead of three ads, as opposed to showing one ad where previously none was shown. A large fraction of overall *NAI* was caused by searches with a single ad in the North - often due to lack of competitors and ensuing unrealistically high bid for infrequent search terms. The *Weighted North Footprint* criterion tends to discourage these cases due to higher first-rank weight.

Beyond *wNFP*, the next more precise approximation to *NAI* is by way of a *relevance model* trained to predict editorial rating. Typically, web ranking already employs a relevance model that is trained to predict editorial ratings, based on features like query word occurrences, link structure etc. If we build a similar model to predict ad relevance on the same target scale, we can use both types of scores for NAI prediction at serve time. Note however, that ad relevance prediction is a more challenging task due to shorter document length and the limited availability of hyperlinks.

We can use the incremental NAI estimate by ranking ads for page placement not only by estimated revenue, but by a *North Ad Placement (NAP)* score that is discounted by the NAI.

5. MULTIPLE CASCADING AUCTIONS

As explained in Sec. 3, practical implementations of ranking and filtering in the second-price auction require the score function to depend on display position only in a very simple, multiplicative way. One way to achieve such flexibility is to have *multiple* auctions using potentially different utility functions. In the same way as advertisers prefer higher ranks on the page, a similar preferential ordering can exist for entire regions of the page (e.g., North placement is generally more desirable than East placement, regardless of the precise rank). In this case, we can auction off the most desirable region first, then run an auction for the next highest value for those advertisers that did not win a placement in the first round, and so on.

The algorithm for the case of two (North and East) regions is given in Fig 2. Note that

- Filtering in step *1(d)* does not permanently remove the candidates, they can still participate in the East auction.
- The cost of the last North ad may not depend on the top East ad.
- The scheme could be generalized for an arbitrary partition of ordered regions; in the extreme case, we could have a separate auction for each single slot.
- In a particular implementation, any of the steps can be trivial (e.g., we currently have no East page placement; the ranking and page placement score could be identical).

The most drastic change is the ability to 'skip' to the North, bypassing other ads in the East that might have a higher ranking *rankScore*, but are not deemed relevant enough. Therefore it can alleviate the "coat tail effect" of the page placement algorithm (Sec. 4.1) that occurs in case of disparity of ranking and NAP score.

One interesting special case of the dual-auction scheme is to use it only to restrict the set of North candidates, but

1. North Auction:
   (a) Rank all candidate ads according to the North ranking score.
   (b) Compute NAP scores for the ordered list (using the same or a different function as ranking).
   (c) Filtering: Remove ads with negative NAP score.
   (d) Compute the costs as in GSP.
   (e) Remember only the winning North ads (if any).
2. East Auction:
   (a) All ads that did not get allocated in the North form the set of East candidates.
   (b) Rank all East candidates according to the East ranking score.
   (c) Filter ads with negative score.
   (d) Compute the costs as in GSP.
3. Merge the ranked North and East ads to form the final ranking.

Figure 2: Algorithm for separate North/East auctions.

otherwise not change the ranking and page placement from the single-auction case. This could be formalized by adding a term to the score that is $-\infty$ for an ad in a North auction that does not meet a given relevance estimate threshold, and zero otherwise.

As an example, consider the following ads in a conventional page placement scheme, where the page placement threshold is 0.9, and the minimum price per click is 0.05:

ad ID	rank	bid	q	rankScore	cost	place
1	1	2.0	0.5	1	1.50	north
2	2	0.5	1.5	0.75	0.17	east
3	3	0.5	1	0.5	0.05	east

Now, imagine a dual auction scheme which requires a minimum q of 1 to be shown in the North (we want to reserve this space for frequently clicked ads; we will refer to it as *dual-coec* in the following section). So only ads 2 and 3 are North-eligible, and participate in the North auction. However, with the same threshold of 0.9, still both would not qualify, and all three ads would end up in the East, in the same order. Often times, we would like to keep the overall north footprint fixed, despite the changed auction mechanism. To this end, suppose that based on the overall distribution, the north threshold is adjusted accordingly to, say, 0.6. Then, the resulting page would look as follows:

ad ID	rank	bid	q	rankScore	cost	place
2	1	0.5	1.5	0.75	0.05	north
1	2	2.0	0.5	1	1.00	east
3	3	0.5	1	0.5	0.05	east

In essence, we have traded higher click through rates in the North for lower prices per click.

The proposed scheme is analogous to the concept of *risk aversion* from social sciences, which is commonly modeled as a non-linear utility function. In our domain, it is reasonable to assume a similar non-linear dependence with regard to user experience as a function of a relevance score: Ads that are only moderately relevant to the user's intent may result in some inconvenience, but a single result that is an egregious mismatch can capture his attention and have a lasting impact on his opinion of the search engine.

6. EXPERIMENTAL RESULTS

In this section, we summarize a number of experiments we ran on a random sample of live US search traffic of a commercial search engine. Each experiment was conducted over a period of one week. To ensure consistent experience, users were assigned to an experiment randomly but fixed based on a hash of their browser cookies. On average, about one million searches per day were issued by 250,000 users.

The resulting metrics are compared to a baseline experiment of equal volume that was run simultaneously. Position normalization of click-through rates is based on the simple yet easily computable model of *COEC (clicks over expected clicks, see e.g. [18] for details)*: position bias is captured in terms of a *reference CTR*, i.e., the mean click-through rate at a given display position on the page (averaging over all ads shown there); the ad-specific term is computed by dividing the ad's observed clicks by the *expected clicks* according to the reference CTR for the position(s) it was shown at. COEC is our implementation of the above-mentioned quality score, q. To cope with sparsity of historical data, a predictive click model [16] was used.

The baseline experiment ranks ads by the product of bid and predicted COEC. By further multiplying this score with the reference CTR for a given rank, we arrive at a proxy for estimated revenue; this is used as the NAP score in the page placement algorithm of Sec. 4.1, with the user utility budget fixed in terms of average weighted north footprint, as described.

Overall, we ran three threads of experiments; all the individual tests are described in Figure 3. In the first thread, we let the *NAP* score be a weighted sum of the revenue estimate and a utility discount based on a *prediction of incremental North Ad Impact* according to Eq. 4, using the top 5 results.

We trained an *ad relevance model* based on n-gram overlap features between the query and the ad text, and on historical click rates to predict editorial ratings; Hillard et al [11] describe details. A separate, independent model is used to score and rank *web results*, in the same way as production web search [19]. Because of engineering requirements, we cached the scores of the top 5 results on the 10 million most frequent queries; for tail queries, we defaulted to average values.

The second thread consisted of experiments with separate North and East auctions, as proposed in Sec. 5. We used different combinations of the COEC and relevance predictions to filter the candidates to be placed in the North.

The third thread used a utility based discount to modify ranking; if the sum of estimated revenue and discount turned negative, the ad was filtered. This is in accordance with the Hidden Cost Model from Sec. 3.2.

To make these experiments comparable to the baseline, in all cases (with the exception of *dual-coec-rel-rev*) we applied page placement thresholds such as to preserve the overall Weighted North Footprint.

Figure 4 gives a summary of the results. All numbers are percentage differences with respect to the baseline experiment. The metrics are defined as follows. When a user issues a query, the search engine responds with a result page. This event is called a *page view*. RPS (revenue per search) – total revenue over total page views; CTR – total num-

Experiment	CTR	CY	PPC	RPS	Cov	Dep	NCTR	Ncov	NFP	wNFP
95% confidence interval	± 0.4	± 0.6	± 1.0	± 1.3	± 0.3	± 0.3	± 0.5	± 0.3	± 0.3	± 0.3
nai-disc	**0.6**	**0.9**	–	–	–	–	**0.7**	–	0.6	0.5
dual-coec	**1.4**	**1.3**	-3.6	-2.4	–	0.6	**3.6**	-1.4	0.7	–
dual-coec-hist	–	–	-6.0	-6.1	–	–	–	-1.2	1.5	0.6
dual-hist	–	–	-1.8	-2.2	–	–	-1.4	–	1.6	1.0
dual-rel	**1.8**	**1.8**	-3.3	-1.6	–	–	**4.1**	-1.4	0.7	–
dual-coec-rel-rev	**7.0**	**6.8**	-9.5	-3.4	–	–	**5.7**	4.9	16.4	13.4
hc-coec	**12.7**	–	-2.3	-1.9	**-10.9**	**-7.6**	-1.4	2.8	–	0.6
hc-coec-rel	**11.7**	–	-4.4	-4.1	**-10.2**	**-4.1**	–	2.0	-0.8	–

Figure 4: Summary of Utility Experiments (all numbers represent relative differences to the control experiment, in percent). For clarity, non-significant results are indicated as '–'.

nai-disc: Set the *NAP* score to a weighted average of estimated revenue and predicted North Ad Impact.

dual-coec: Separate North/East Auctions. Only ads whose *COEC* exceeded a certain threshold were allowed to be shown in the North.

dual-hist: Separate North/East Auctions. Only ads that had any historical click data recorded were allowed in the North. The rationale is to use the availability of history as a proxy for confidence in our COEC prediction.

dual-coec-hist: Separate North/East Auctions. This experiment combines the previous two: the condition for North ads is that the ad has a minimum *COEC* and historical information.

dual-rel: Separate North/East Auctions. North ads were required to pass at least one of a threshold on *COEC* and the estimated editorial relevance.

dual-coec-rel-rev: Separate North/East Auctions. To be eligible for North placement, an ad has to pass at least one of three thresholds: on *COEC*, on relevance score, or on estimated revenue. Different from other experiments, any placeable ad that fulfills this condition is automatically shown, without subsequent page placement procedure.

hc-coec: Page placement identical to baseline. Pricing/Ranking/Filtering applies a Hidden Cost Discount (Sec. 3.2) of the form $param/COEC$.

hc-coec-rel: Same as *hc-coec*, but the discount is based on a weighted average between the relevance score and *COEC*, where the weight depends on the amount of historical data.

Figure 3: Description of live traffic experiments.

ber of clicks over total number of page views with ads; CY (click yield) – total clicks over total page views; PPC – average price per click; NCTR (north CTR) – CTR of all page views with North ads; Ncov (north coverage) – ratio of page views that have North ads; NFP – average number of North ads per page view; wNFP – average DCG-weighted NFP.

In general, we can interpret click yield as a measure of user engagement. If we can maintain RPS, but increase clicks while dropping PPC, advertiser ROI will increase on average, assuming constant conversion rates. Note that a change in price per click does not affect all advertiser equally in the general case; rather, different ranking or placement results in users clicking more on some ads but less on others, whose prices may be different.

For all experiments except *dual-coec-rel-rev*, we retuned the page placement threshold such that the NFP was roughly neutral compared to the baseline. What we want to achieve is to allocate the same ad footprint more efficiently by shifting it from one query to another one.

The NAI-based experiment shows a slight improvements in terms of the click and revenue metrics.

In line with expectations, all of the dual-auction variants (but *dual-hist*) lead to increased (North) CTR and lower RPS due to price drops. Similar to the NAI-based experiments, the *NAP* distribution changes towards lower North coverage and higher North depth. *COEC*-based filtering increases click yield reasonably, but the relevance-model (*dual-rel*) achieves similar click metrics with less revenue loss.

The idea of *dual-coec-hist* was to strengthen the criterion of *dual-coec* by additionally requiring a confidence in the *COEC* estimate, expressed as the availability of historical information. Note that in particular this will exclude newly created ads. Contrary to our expectation, this led to a degradation in both click and revenue metrics. The negative effect is even more pronounced in the case that we use history alone (*dual-hist*). Maybe we can interpret this result in the way that some ads have a high CTR, despite being shown infrequently or being new, and that the click model gives us a decent estimate for them.

The combination of the relevance, *COEC*, and revenue criteria (*dual-coec-rel-rev*) is the most disruptive experiment. Ads have to pass at least one of three thresholds on COEC, predicted relevance, or COEC times bid in order to be shown in the north; in contrast to the previous experiments, once these thresholds are passed, no further page placement conditions based on expected revenue are imposed. Another difference is that we allowed the north footprint to increase, in order to absorb some of the revenue drop. The rationale was that most of the ads were selected based on relevance or high CTR, so despite the increase the user impact should be lower. The CTR and click yield increase are highest for this experiment, and the simultaneous price drop should improve advertiser experience by providing them more clicks at a lower cost. We also conducted a manual evaluation by human raters for this test, which showed a 24% drop in NAI.

While all the experiments mentioned so far do not change the total number of ads shown (coverage and depth), the *Hidden Cost experiments* affect mostly those metrics due to more rigorous *filtering* for negative total scores. The discount in ranking tends to decrease the ranking score and consequently the cost for the next higher ad; this leads to the observed PPC and RPS drop. Despite much lower coverage and depth, there is no loss in total clicks. The variant based on pure CTR estimate, *hc-coec*, shows less of a

price drop than the one incorporating the relevance model score, *hc-coec-rel*; however, the click metrics are very similar. In summary, the Hidden Cost experiments show that it is possible to reduce the number of shown ads greatly while maintaining the overall click volume.

7. CONCLUSION AND FUTURE WORK

In this paper, we developed formalisms to describe general utility functions for web search advertising systems beyond instantaneous revenue, e.g., by taking into account user satisfaction. We defined and outlined problems in sponsored search optimization such as *page placement* and *filtering* that, to the best of our knowledge, have only received limited attention in the current literature. Our overarching goal is to solve these problems in a way that balances search engine revenue and user experience in a meaningful way.

We acknowledge that there is still a gap between general utility formalisms and practically feasible implementations in commercial search engines. The reasons are diverse - actual human behavior is very varied and complex, but we need to simplify user models in order to obtain tractable results; and even for these simplified models, it is hard to estimate hidden parameters (such as U_{distr}, U_{rel}, and U_{irrel}) in a reliable and accurate way. Estimation could be based on psychological models of user perception; alternatively, we could develop economical prediction models by trying to quantify the impact of seen ads on future revenue from the user. These are beyond the scope of this paper, but we will address it in future work. Clearly we have only scratched the surface, but hope to have opened the discussion.

Instead, in this paper, we focused on some special cases of utility that are amenable to current implementation. Based on the widely used *DCG* relevance metrics [13], we defined the *North Ad Impact* criterion. We estimate this online using machine-learned web and ad relevance models. We proposed and implemented three different ways of using relevance information: combining it with the revenue estimates in the page placement score; ranking and filtering according to the *Hidden Cost* model; and breaking up the auction into more or less prominent page regions, and use relevance as an ad eligibility criterion for the former one. We tested these models on live traffic from a commercial search engine. The resulting behavioral metrics show significant improvements, as e.g. measured in terms of total received clicks.

8. REFERENCES

[1] Z. Abrams and M. Schwarz. Ad auction design and user experience. *Applied Economics Research Bulletin*, 2(I), Fall 2008.

[2] G. Aggarwal, J. Feldman, S. Muthukrishnan, and M. Pal. Sponsored search auctions with markovian users, May 2008.

[3] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web*, pages 1–10, New York, NY, USA, 2009. ACM.

[4] 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.

[5] 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.

[6] Y. Engel and D. M. Chickering. Incorporating user utility into sponsored-search auctions. In *AAMAS (3)*, pages 1565–1568, 2008.

[7] 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.

[8] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[9] 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.

[10] F. Guo, C. Liu, A. Kannan, T. Minka, M. J. Taylor, Y. M. Wang, and C. Faloutsos. Click chain model in web search. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *WWW*, pages 11–20. ACM, 2009.

[11] D. Hillard, S. Schroedl, E. Manavoglu, H. Raghavan, and C. Leggetter. Improving ad relevance in sponsored search. In *Third ACM International Conference on Web Search and Data Mining (WSDM) (to appear)*, New York City, February 2010.

[12] B. J. Jansen and T. Mullen. Sponsored search: an overview of the concept, history, and technology. *IJEB*, 6(2):114–131, 2008.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[14] 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.

[15] D. Kempe and M. Mahdian. A cascade model for externalities in sponsored search. In *WINE*, pages 585–596, 2008.

[16] 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.

[17] H. R. Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, December 2007.

[18] 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.

[19] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun. A general boosting method and its application to learning ranking functions for web search. In *NIPS*, 2007.