User Browsing Models: Relevance versus Examination

Ramakrishnan Srikant Sugato Basu Ni Wang Daryl Pregibon

Google Research
{srikant, sugato, niwang, daryl}@google.com

ABSTRACT

There has been considerable work on user browsing models for search engine results, both organic and sponsored. The click-through rate (CTR) of a result is the product of the probability of examination (will the user look at the result) times the perceived relevance of the result (probability of a click given examination). Past papers have assumed that when the CTR of a result varies based on the pattern of clicks in prior positions, this variation is solely due to changes in the probability of examination.

We show that, for sponsored search results, a substantial portion of the change in CTR when conditioned on prior clicks is in fact due to a change in the relevance of results for that query instance, not just due to a change in the probability of examination. We then propose three new user browsing models, which attribute CTR changes solely to changes in relevance, solely to changes in examination (with an enhanced model of user behavior), or to both changes in relevance and examination. The model that attributes all the CTR change to relevance yields substantially better predictors of CTR than models that attribute all the change to examination, and does only slightly worse than the model that attributes CTR change to both relevance and examination. For predicting relevance, the model that attributes all the CTR change to relevance again does better than the model that attributes the change to examination. Surprisingly, we also find that one model might do better than another in predicting CTR, but worse in predicting relevance. Thus it is essential to evaluate user browsing models with respect to accuracy in predicting relevance, not just CTR.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Human Factors

KDD'10, July 25–28, 2010, Washington, DC, USA.

1. INTRODUCTION

Web search engines have become an essential tool for navigating the vast amounts of information on the internet. Implicit user feedback, specifically, click-through data, is valuable for optimizing search engine results [2, 10, 14]. Click-through data plays an equally important role in estimating the quality of sponsored search results [15].

Any attempt at using click-through data for search or sponsored search runs into the following issue: Eye-tracking studies show that people tend to scan the search results in order [3, 11]. However, they are likely to click on a result as soon as they find one that they consider helpful, and if that result provides a sufficiently helpful answer, may not look at other results. This causes *position bias*: the same result will get a higher click-through rate (CTR) if it is positioned towards the top of the page (versus the bottom).

Thus algorithms that use click-through data have to take position bias into account. Initial work on estimating position bias modeled CTR as the product of perceived relevance (probability of a click given that the user examined the result) times the probability of examination (probability that the user would examine this specific position) [15]. The examination probability was estimated by looking at the CTR of the same result in different positions.

Subsequently, there have been many papers on better estimating the probability of examination by using the pattern of clicks on prior results [5, 7, 9, 12], or using both prior clicks and the relevance of prior results [4, 8, 18]. We discuss this work in detail in Section 2. The key point is that all of these papers assume that if CTR changes when conditioned on the pattern of clicks on prior results, the change in CTR is *solely* due to changes in the probability of examination.

Consider a query such as "Canon S90". The user could be planning to buy the camera immediately, in which case the sponsored results are highly relevant. On the other hand, if the user is just starting to learn about the camera, the sponsored results will be much less relevant. Thus a click on the first sponsored result is a signal that the other sponsored results are also relevant.

Say we now partition the query instances corresponding to the query "Canon S90" into two sets based on whether the first result got a click: the "click" and "no-click" sets. The second result will be highly relevant for the query instances in the "click" set, and less relevant for the "no-click" set, even though the query and result are the same. Thus the second result will have higher CTR in the "click" set than in the "no-click" set. However, current user models assume that the relevance is the same in both the "click" and "no-click"

sets, and that all the difference in CTR for the second result is because users in the "click" set were more likely to examine the second result than users in the "no-click" set.

Contributions.

In this paper, we examine the implications of the above insight. In Section 3, we show that, for sponsored search results, an increase in relevance is indeed responsible for a substantial portion of the increase in CTR when conditioned on prior clicks. We then propose three new user browsing models in Section 4, which attribute CTR changes solely to changes in relevance, solely to changes in examination (with an enhanced model of user behavior), or to both changes in relevance and examination. We evaluate the accuracy of these models when predicting CTR in Section 5, and the accuracy when predicting relevance in Section 6. Our results show that, surprisingly, one model might do better than another in predicting CTR but worse in predicting relevance. We conclude with a summary of our results and directions for future work in Section 7.

2. BACKGROUND

Prior user browsing models for web search results can be partitioned into three groups based on how they estimate the probability that the user examines a specific position:

- Models that assume examination is independent of the other results for the query.
- Models that assume examination depends on the pattern of clicks on prior results.
- Models that assume examination depends on both the pattern of clicks on prior results, and the relevance of prior results.

Some of the models were originally targeted at sponsored search, while others were targeted at organic search results. However, while the parameter values might differ, all of these models are general enough to apply to both organic search and sponsored search.

2.1 Examination independent of other results

We use the following notation:

- Let $\phi(i)$ denote the result at position i in a ranked list of results (whether organic results or sponsored results).
- Let C_i denote a binary random variable that captures the event that a user clicks on $\phi(i)$.
- Let E_i denote the event that the user examines $\phi(i)$.

The **examination hypothesis**, originally proposed by Richardson et al. [15] and formalized by Craswell et al. [5], observes that to be clicked, a result must be both examined and relevant:

$$\Pr(C_i = 1) = \Pr(C_i = 1|E_i = 1)\Pr(E_i = 1). \quad (1)$$

Richardson et al. [15] assume that the probability a result is viewed depends solely on its position, and is independent of other results.

We call the statistical model that derives from the examination hypothesis the **baseline model**:

$$\Pr(C_i = 1) = r_{\phi(i)}\,\alpha_i, \quad (2)$$

where

- $r_{\phi(i)} = \Pr(C_i = 1|E_i = 1)$ represents the relevance of the result in position i, and
- $\alpha_i = \Pr(E_i = 1)$ models the position bias.

Richardson et al. [15] proposed estimating the α_i parameters by presenting users with the same result at different positions and observing the change in CTR.

2.2 Examination depends on prior clicks

An implicit assumption in the above formulation is that the probability of examining the result in position i does not depend on click events in other result positions. A plethora of recent papers explore models that incorporate this information into the examination probabilities.

The **cascade hypothesis** [5] assumes that users scan each result sequentially without any skips:

$$\Pr(E_1 = 1) = 1,$$
$$\Pr(E_i = 1|E_{i-1} = 0) = 0.$$

The **cascade model** [5] further constrains that the user continues examining results until she clicks on a result, and does not examine any additional results after the click:

$$\Pr(E_i = 1|E_{i-1} = 1) = 1 - C_{i-1} \quad (3)$$

This model is quite restrictive since it allows at most one click per query instance.

The **dependent click model** (DCM) [9] generalizes the cascade model to instances with multiple clicks:

$$\Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 1) = \lambda_i,$$
$$\Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 0) = 1.$$

The authors suggest estimating the position effects λ_i using maximum likelihood.

The **user browsing model (UBM)** [7] is also based on the examination hypothesis, but unlike the cascade model and DCM, does not force $\Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 0)$ to be 1. In other words, it allows users to stop browsing the current results and instead reformulate the query (or perhaps give up). UBM assumes that the examination probability is determined by the preceding click position $p(i) = argmax_{l<i}\{C_l = 1\}$:

$$\Pr(E_1) = \alpha_1$$
$$\Pr(E_i = 1|C_{1:i-1}) = \alpha_i\,\beta_{i,p(i)}, \quad (4)$$

where $\alpha_i = \Pr(E_i = 1)$ is the examination probability of position i without taking other click information into account (just as in the baseline model), and $\beta_{i,p(i)}$ denotes the correction factor over α_i given $p(i) = argmax_{l<i}\{C_l = 1\}$.[1] To avoid confusion between the above "user browsing model" in Equation 4, and the category of user browsing models, we will refer to this specific model as UBM.

The **bayesian browsing model** (BBM) [12] uses exactly the same browsing model as UBM. However, BBM adopts a Bayesian paradigm for relevance, i.e., BBM considers relevance to be a random variable with a probability distribution, rather than a fixed (but unknown) value to be estimated. In the context of this paper, where we are focused on the user browsing model, UBM and BBM can be considered equivalent.

[1]Note that our notation is slightly different than that in [7]. We consider β to be a correction factor on α, while they used β for the product of our definitions of α and β.

2.3 Examination depends on prior clicks and prior relevance

Next, we summarize models that take into account both clicks on prior results, and the relevance of those results, to predict the probability of examination.

The **click chain model** (CCM) [8] is a generalization of DCM obtained by parameterizing λ_i and by allowing the user to abandon examination of more results:

$$\begin{aligned}\Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 0) &= \alpha_1 \\ \Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 1) &= \alpha_2(1 - r_{\phi(i-1)}) + \alpha_3 r_{\phi(i-1)}.\end{aligned}$$

Thus if a user clicks on the previous result, the probability that they go on to examine more results ranges between α_2 and α_3 depending on the relevance of the previous result.

The **general click model** (GCM) [18] treats all relevance and examination effects in the model as random variables:

$$\begin{aligned}\Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 0) &= \Pi(A_i > 0) \\ \Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 1) &= \Pi(B_i > 0) \\ \Pr(C_i = 1|E_i) &= \Pi(r_{\phi(i)} > 0).\end{aligned}$$

This allows online inference within the cascade family. These authors show that all previous models are special cases by suitable choice of the random variables A_i, B_i, and $r_{\phi(i)}$.

2.4 Post-click models

In our discussion so far, relevance referred to "perceived" relevance – whether the user considers the result relevant before she clicks on the result. Post-click relevance is a measure of whether the user was satisfied with their experience after clicking on the result. Perceived relevance is positively correlated with post-click relevance [16]. However, there are cases where perceived relevance is high and post-click relevance is low (e.g., snippet or creative is inaccurate), or vice versa (e.g., only a small fraction of people searching "yahoo" want answers.yahoo.com – but for those people, it's perfect). Thus both perceived and post-click relevance are equally important for user satisfaction.

The **dynamic bayesian model** (DBM) [4] uses the "user satisfaction" (post-click relevance) of the preceding click to predict whether the user will continue examining additional results:

$$\begin{aligned}\Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 0) &= \gamma \\ \Pr(E_i = 1|E_{i-1} = 1, C_{i-1} = 1) &= \gamma(1 - s_{\phi(i-1)}),\end{aligned}$$

where $s_{\phi(i-1)}$ is the satisfaction of the user in the previous clicked result. They propose an EM-type estimation method to estimate γ and the user satisfaction variables.

The **session utility model** (SUM) [6] proposes a user browsing model based on the "intrinsic" (post-click) relevance of the sequence of clicked results in a user session. However, it does not model examination or pre-click relevance.

Our focus in this paper is on correctly estimating examination and perceived relevance. Thus in the rest of the paper, we will use "relevance" as shorthand for "perceived relevance", and focus on the models in Sections 2.1 and 2.2. We will briefly revisit the other models when we discuss future work in Section 7.

3. CONSTANT RELEVANCE ASSUMPTION

3.1 Instance relevance

There is an implicit assumption underlying the models in Sections 2.1, 2.2 and 2.3:

> **Constant Relevance Assumption**: The relevance of a result to a query is constant across query instances.

More formally, these models assume that the perceived relevance, $\Pr(C_i = 1|E_i = 1)$, of a result is independent of the pattern of clicks on prior results:

$$\Pr(C_i = 1|E_i = 1, C_{1:i-1}) = \Pr(C_i = 1|E_i = 1)$$

The constant relevance assumption may appear quite reasonable at first glance: Isn't the relevance of the result simply dependent on the query and the result? However, "relevance" can have two very different meanings:[2]

- **aggregate relevance**: the relevance of a result for a query, averaged over all instances of the query, or
- **instance relevance**: the relevance of a result for the current query instance.

Aggregate relevance depends only on the query and result – precisely the intuition behind the constant relevance assumption. However, $\Pr(C_i = 1|E_i = 1, C_{1:i-1})$ corresponds to instance relevance (i.e., relevance for the current query instance), not aggregate relevance. It is easy to make a case that $C_{1:i-1}$ is in fact a predictor of instance relevance, especially for sponsored search results.

The key intuition is that the query string does not fully capture *user intent*. Consider a query "Canon T2i". Some subset of users who issue this query will be interested in buying a camera at the time they issued the query, and sponsored search results will be highly relevant to them. Other users may be potentially interested in buying a camera at some point in the future, but are currently primarily interested in learning more about the camera. The sponsored results will be much less relevant to these users. Thus for queries with multiple user intents, often only one (or a subset) of these intents is represented by sponsored search results. For such queries, $\Pr(C_i = 1|E_i = 1)$ for a query instance will be strongly correlated with $\Pr(C_j = 1|E_j = 1)$ for the same query instance, where positions i and j both correspond to sponsored results.[3]

3.2 Testing the constant relevance assumption

Assume that for some position, we can get a set of query instances where we know the probability of examination is close to 1, i.e., the user examined the position with high probability. Given such a set, it would be easy to test whether the constant relevance assumption is valid. If CTR is independent of the pattern of clicks on other positions, then the assumption is true. If CTR increases as the number of clicks on other positions increases, then the assumption is false, since the CTR change must be solely due to change in instance relevance.

[2]If one also considers post-click relevance, "relevance" has four distinct meanings: post-click versus pre-click, and aggregate versus instance relevance.

[3]The fact that query reformulations are common suggests that similar correlations may also exist among organic search results.

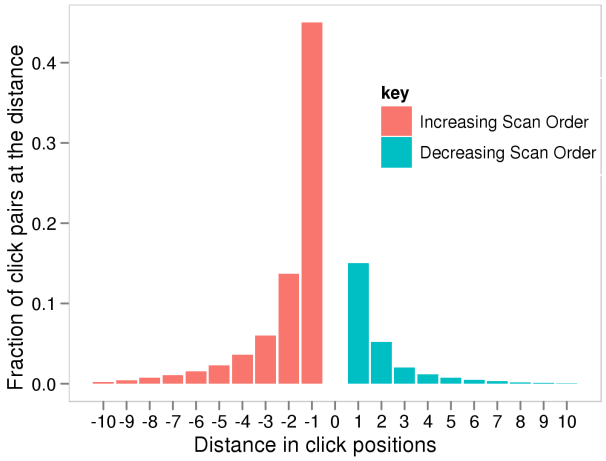


Figure 1: Distribution of the difference in positions between temporally adjacent clicks.

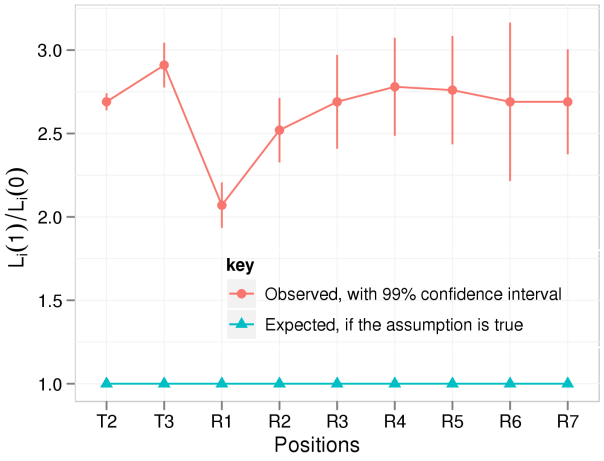


Figure 2: Testing the constant relevance assumption.

We now show that there is indeed a way get such a subset of query instances. If users scan results linearly from top to bottom, and there is a click at position i, then the user must have examined the results at positions 1 through $i-1$. We show next that users indeed scan linearly from top to bottom.

3.2.1 Linear scan from top to bottom

There is evidence from eye-tracking studies [3, 11] that users scan results linearly from top to bottom. The user browsing models in Sections 2.2 and 2.3 also assume linear scan.

However, there have also been eye-tracking studies showing that rather than a simple linear scan, users do page chunking. They partition the page into chunks, select the chunk they want to examine, and then scan items in that chunk in a linear fashion [1]. Hence we first do some due diligence to see whether the data supports the linear scan assumption.

Formally, we would like to assume that a click at position i implies that the user examined all preceding positions with probability close to 1:

$$C_i = 1 \implies \Pr(E_j = 1) \geq (1-\epsilon), \forall j < i. \quad (5)$$

Figure 1 shows the distribution of temporally adjacent pairs of clicks as a function of their positional distance, over all sponsored search results. A negative distance corresponds to pairs of adjacent clicks where the user clicks are linear scan order (i.e., top to bottom), while a positive distance corresponds to bottom to top order. The gap at 0 simply means that users typically do not have consecutive clicks on the same result.

Only around 5% of temporally adjacent click pairs are both not in linear scan order and have a gap greater than 2. Even in these cases, it's possible that the user scanned linearly and then went back to an earlier result, e.g., when comparison shopping. Thus Equation 5 appears to be a reasonable basis for testing the constant relevance assumption.

3.2.2 The data speaks

Now that we have a set of queries and positions where $\Pr(E_i = 1)$ is close to 1, we can look at whether relevance changes when conditioned on the pattern of other clicks for that query instance.

We first introduce some notation. Let $s(i)$ equal 1 iff there was a click below position i, i.e., the set $\{j | j > i, C_j = 1\}$ is non-empty. Let $p'(i)$ equal 1 if there was a click above position i, and 0 otherwise. Equation 5 implies that if the constant relevance assumption is true, we would expect:

$$\Pr(C_i = 1 | p'(i) = 1, s(i) = 1) = \Pr(C_i = 1 | p'(i) = 0, s(i) = 1)$$

This follows from the fact when $s(i) = 1$, $\Pr(E_i = 1) \approx 1$, and the constant relevance assumption implies that relevance is also fixed, so

$$\begin{aligned} \Pr(C_i = 1 | p'(i) = 1, E_i = 1) &= \Pr(C_i = 1 | E_i = 1) \\ &= \Pr(C_i = 1 | p'(i) = 0, E_i = 1) \end{aligned}$$

Let $Q_i(j)$ denote, for some specific query and result, the subset of query instances where $s(i) = 1$ and $p'(i) = j$. If the CTR is different when conditioned on $p'(i)$, i.e.,

$$\frac{\sum_{q \in Q_i(1)} C_i}{\sum_{q \in Q_i(1)} 1} \neq \frac{\sum_{q \in Q_i(0)} C_i}{\sum_{q \in Q_i(0)} 1}$$

and the difference is statistically significant, we would have disproved the constant relevance assumption.

To increase the power of the statistical test, we extend the above test to multiple queries and results, within a single configuration and position. Let $T_i(j)$ denote, for a specific configuration, the set of query instances where $s(i) = 1$ and $p'(i) = j$. Since there may be small differences in the mix of queries and results between $T_i(0)$ and $T_i(1)$, we change the denominator to be the expected clicks, rather than impressions. Let $L_i(j)$ be defined as:

$$L_i(j) = \frac{\sum_{q \in T_i(j)} C_i}{\sum_{q \in T_i(j)} \Pr(C_i = 1 | E_i = 1)}$$

The numerator is the observed number of clicks, i.e., the sum of the observed relevance. The denominator is the sum of the expected relevance. The constant relevance assumption implies that the ratio of the observed to the expected relevance should be independent of $p'(i)$. Hence if the constant

relevance assumption is true, and any errors in relevance estimates are approximately equal in both sets, $L_i(1)$ should be roughly equal to $L_i(0)$.

We present results for the 3-8 configuration (3 top results, 8 rhs results) in Figure 2. Here $T2$ refers to the second top result, $R3$ refers to the third rhs result, etc. The red line shows the observed value of $L_i(1)/L_i(0)$ for different positions, with a 99% confidence interval ($\pm$ 2.58 standard deviations). The confidence intervals were estimated by partitioning the data into 10 sets and computing variance. The ratio is quite consistent across positions. The dip at R1 is because users typically scan the top sponsored results first, then the organic search results, and then the rhs sponsored results. $L_i(1)/L_i(0)$ is much greater than 1 for all positions, and the results are statistically significant. This disproves the constant relevance assumption.

The results are consistent across other configurations. The weighted average of $L_i(j)/L_i(0)$ over all configurations and positions is 2.69, with a 99% confidence interval of $\pm$ 0.05.[4]

4. NEW USER BROWSING MODELS

Having shown that changes in CTR when conditioned on other clicks are at least partly due to changes in instance relevance, we propose new user browsing models that take advantage of this insight. Our first model, *pure relevance*, is a strawman that assumes that CTR changes are solely due to changes in instance relevance, and not due to changes in examination. (However, as we will see later, this strawman does surprisingly well.) Our second model, *max-examination*, assumes (like prior work) that CTR changes are solely due to changes in examination – but uses additional information to better predict whether the user examined the result. Our third model, *JRE* generalizes both these models, and allows CTR changes to be caused by both changes in examination and instance relevance.

4.1 Pure relevance model

The *pure relevance model* assumes that any changes in the probability of a click due to conditioning on prior clicks is caused solely by change in the expected instance relevance of the result. Thus the probability of examination is assumed to be independent of clicks on other results:

$$\Pr(E_i = 1|C_{\neq i}) = \Pr(E_i = 1) = \alpha_i,$$

where $C_{\neq i}$ is the pattern of clicks in all positions except i, i.e., $C_{\neq i} = C_{1:i-1,i+1:m}$ where m is the number of positions. The model then assumes that the changes in instance relevance can be estimated using the total number of clicks in other positions:

$$\Pr(C_i = 1|C_{\neq i}, E_i = 1) = r_{\phi(i)}\,\delta_{\eta(i)} \tag{6}$$

where

- $\eta(i) = \sum_{k\neq i} C_k$ is the total number of clicks in positions other than i,
- $\delta_{\eta(i)}$ is the correction factor to get the expected instance relevance (for the result in position i) when conditioned on $C_{\neq i}$.

[4]The weight for a given configuration and position is $\min(\sum_{q\in T_i(0)} C_i, \sum_{q\in T_i(1)} C_i)$.

Thus $r_{\phi(i)}$ is the aggregate relevance, while $r_{\phi(i)}\delta_{\eta(i)}$ is the estimated instance relevance after conditioning on $C_{\neq i}$. The pure relevance model is then defined by

$$\begin{aligned}\Pr(C_i = 1|C_{\neq i}) &= \Pr(E_i = 1)\Pr(C_i = 1|C_{\neq i}, E_i = 1)\\ &= \alpha_i\, r_{\phi(i)}\,\delta_{\eta(i)}.\end{aligned} \tag{7}$$

Pure Relevance versus Baseline.

While the pure relevance and baseline models will yield different CTR estimates for any query instance, they will in fact yield *identical* relevance estimates $r_{\phi(i)}$. For any result, both pure relevance and baseline have the same probability of examination. Since $r_{\phi(i)} = \sum C_{\phi(i)}/\sum E_{\phi(i)}$, and clicks and examination are the same, the relevance estimate must also be the same.

4.2 Max-examination

Recall from Equation 5 that if there is a click on a position below i, then there is a high probability that position i was examined. So, if we include information about clicks below position i while estimating the probability of examination, the model should have significantly more information than models (like UBM/BBM) that only consider clicks above position i.

With the above intuition in mind, we propose the *max-examination* model. As before, let $p(i)$ be the position of the preceding click. Let $s(i)$ be 0 if there is no click below position i and 1 if there is a click below i. We define $e(i)$:

$$e(i) = \begin{cases} p(i) & \text{if } s(i) = 0\\ i+1 & \text{if } s(i) = 1.\end{cases}$$

We then replace the $p(i)$ used in the UBM model (Equation 4) with $e(i)$, and thereby incorporate the case when the click occurred below i. To avoid confusion, we also change the notation to $\gamma_{i,e(i)}$ instead of $\beta_{i,p(i)}$. This yields the following equations for the max-examination model:

$$\Pr(E_i = 1|C_{\neq i}) = \alpha_i\gamma_{i,e(i)} \tag{8}$$

$$\Pr(C_i = 1|C_{\neq i}) = \alpha_i\gamma_{i,e(i)}r_{\phi(i)} \tag{9}$$

4.3 Joint relevance examination model (JRE)

A natural generalization of the pure relevance and max-examination models is to combine their features, and allow CTR changes to be caused by both changes in examination and changes in instance relevance. We call this the *joint relevance examination (JRE) model*. We combine the relevance component from the pure relevance model (Equation 6), with the examination component from the max-examination model (Equation 8), to get

$$\begin{aligned}&\Pr(C_i = 1|C_{\neq i})\\ &\quad= \Pr(E_i = 1|C_{\neq i})\Pr(C_i = 1|C_{\neq i}, E_i = 1)\\ &\quad= \alpha_i\,\gamma_{i,e(i)}\,r_{\phi(i)}\,\delta_{\eta(i)}.\end{aligned} \tag{10}$$

Note that an estimate of $\gamma_{i,e(i)}$ in the JRE model will not be the same as the corresponding value in the max-examination model, since the credit is shared between γ and δ. For the same reason, an estimate of $\delta_{\eta(i)}$ will be different in the JRE model and the pure relevance model.

Conceptually, there is a single "true" value of aggregate relevance, $r_{\phi(i)}$. However, different models may yield different estimates of $r_{\phi(i)}$ – we will explore this issue further in Section 6.

Diversity of results.

In the pure relevance and JRE models, we implicitly assume that the set of results are homogeneous, and therefore a click on one result would likely be a good predictor of clicks on other results. However, it is easy to come up with scenarios where the results are diverse, e.g., for the query "jaguar", one would expect a positive correlation between clicks on results about the car, or between clicks on results about the animal, but a negative correlation between clicks on a car result and clicks on an animal result.

Instance relevance may also be different between top and rhs sponsored search results, even when both sets of results are on the same topic. Users typically scan top sponsored links before organic results, while they scan rhs sponsored links after organic results. Therefore the top results may be relevant, while the rhs results may be less relevant if the top or organic results were sufficient to answer the query.

Fortunately, given a partitioning of the results into (approximately) homogeneous groups, it is trivial to update the pure relevance and JRE models. The only change is that $\eta(i)$, rather than being the number of clicks on other results, becomes the number of clicks on other results in the same group.

5. PREDICTING CTR

We showed in Section 3.2.2 that the constant relevance assumption was incorrect. We now evaluate which user browsing models best predict CTR in offline analysis: models that attribute CTR changes solely to examination, solely to instance relevance, or to both examination and instance relevance?

We compare the three models proposed in Section 4 to two of the user browsing models from prior work:

- UBM [7] (or equivalently, BBM), which is the most general (powerful) of the models in Section 2.2,
- the baseline which does not use co-click information (Equation 2), but yields identical relevance estimates as the pure relevance model.

We describe how we fit the parameters in the models in 5.1, followed by experimental results in Section 5.2.

5.1 Methodology

Baseline.

To evaluate the user browsing models for predicting CTR, we need to combine them with a machine learning system for predicting relevance. We used Google's production system for predicting relevance of sponsored links for both the machine learning and user browsing components of the baseline model. The user browsing model in Google's system does not make use of co-click information, and is thus similar to the baseline model (Equation 2). The machine learning model for predicting relevance uses the query, the position bias for the position in which the result appeared, whether the result was clicked, and various features of the query and the sponsored result to predict relevance.

Other Models.

Our baseline model is sufficiently complex that, if we directly add a new feature (such as a different user browsing model), it's not possible to isolate the accuracy gains from the new feature, versus the new feature shifting the model to a different local optimum. To get a fair comparison, we use the output of the baseline model ($\alpha_i r_{\phi(i)}$) as a given, and separately optimize the co-click dependent parameters.

The input data for these models was a 10% sample, over a week, of the logged predicted CTR ($\alpha_i r_{\phi(i)}$) from the baseline model, along with co-click information. We used a 50-50 split of the data into training and testing.

For each model, we fit parameters separately for each configuration (number of top and rhs sponsored results) and position. To keep the notation simple, we will express the model for position j without reference to the configuration. We next describe how we estimate the parameters for each model.

UBM/BBM.

We estimate the $\beta_{i,p(i)}$ parameters in UBM/BBM (Equation 4), where $p(i)$ represents the position of the preceding click (if any) above position i, using:

$$\beta_{i,k} = \frac{\sum_{p(i)=k} C_i}{\sum_{p(i)=k} \alpha_i r_{\phi(i)}}$$

The numerator corresponds to the total number of clicks at position i for query instances where $p(i) = k$, while the denominator corresponds to the expected number of clicks (without including $\beta_{i,k}$). Thus $\beta_{i,k}$ is set to the value where the number of clicks predicted by UBM ($\sum_{p(i)=k} \alpha_i \beta_{i,k} r_{\phi(i)}$) equals the observed number of clicks.

Max Examination.

The methodology for estimating $\gamma_{i,e(i)}$ is similar, we again equalize the predicted and observed number of clicks:

$$\gamma_{i,k} = \frac{\sum_{e(i)=k} C_i}{\sum_{e(i)=k} \alpha_i r_{\phi(i)}}$$

Pure Relevance.

For the pure relevance model from Section 4.1, we similarly estimate $\delta_{i,\eta i}$ using:

$$\delta_{i,k} = \frac{\sum_{\eta(i)=k} C_i}{\sum_{\eta(i)=k} \alpha_i r_{\phi(i)}}$$

As we discussed in Section 4.3, the top and rhs sponsored results are sufficiently diverse that clicks on the top are not necessarily a signal of instance relevance for the rhs (and vice versa). Thus we restrict $\eta(i)$ to be the number of clicks at other position in the same slot, i.e., the number of other top clicks for top sponsored results, and the number of other rhs clicks for rhs sponsored results.

JRE.

For JRE, we need to estimate two parameters, $\gamma_{i,e(i)}$ and $\delta_{i,\eta(i)}$, hence we use iterative fitting. We initialize all the γ and δ parameters to 1.0. We then repeatedly re-estimate the parameters using:

$$\gamma_{i,k} = \frac{\sum_{e(i)=k} C_i}{\sum_{e(i)=k} \alpha_i r_{\phi(i)} \delta_{i,\eta(i)}}$$

$$\delta_{i,k} = \frac{\sum_{\eta(i)=k} C_i}{\sum_{\eta(i)=k} \alpha_i r_{\phi(i)} \gamma_{i,e(i)}}$$

The results are not sensitive to the algorithm for fitting the

parameters – the relative accuracy of the models was similar when we tried logistic regression (after mapping the models to odds space).[5]

5.2 Results

Figure 3 shows the results for the 3-8 configuration (3 top sponsored results, 8 rhs sponsored results), for three different accuracy metrics: log-likelihood, squared error, and absolute error. The y-axis shows improvements in the metric relative to the baseline model for the various positions in this configuration (shown on the x-axis). As before, $T1$ represents the first top result, $R3$ the third rhs result, etc.

First, note that there is a clear ordering, consistent across all metrics, between the models: JRE $\geq$ pure relevance $>$ max-examination $>$ UBM/BBM $>$ baseline.

Consistent with prior work, UBM/BBM does significant better than baseline, by leveraging co-click information. Interestingly, max-examination does significantly better than UBM.[6] The difference is highest for the first rhs position $R1$. Top clicks are not a strong predictor of examination for the rhs, while a click below $R1$ is a strong predictor of examination.

Pure relevance does slightly better than max-examination wrt log-likelihood, and substantially better wrt the other metrics. Surprisingly, JRE does only slightly better than pure relevance. Together, these results suggest that changes in instance relevance are probably responsible the majority of the changes in CTR (when conditioned on other clicks), with the caveat that the examination and relevance features are correlated (not clearly separable).

Figure 4 shows the overall percentage improvement across all configurations and positions, with 99% confidence intervals (2.58 standard deviations) computed by partitioning the test data into 10 groups. The relative performance of the models is consistent with Figure 3. The separations are all statistically significant, except between relevance and JRE for the first two metrics.

6. PREDICTING RELEVANCE

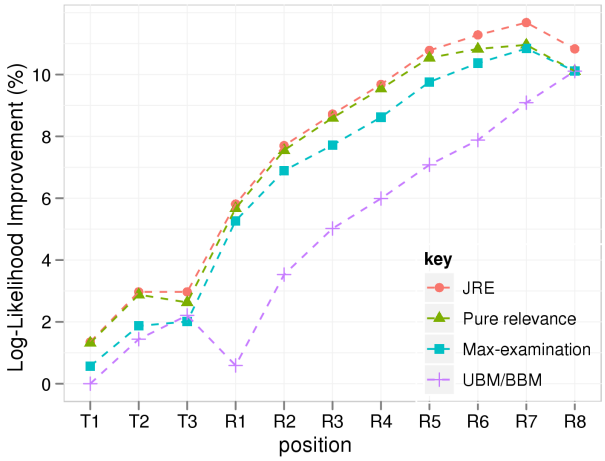
Many papers on user browsing models for web search evaluate their models solely based on the accuracy of the model for predicting CTR in offline analysis [4, 7, 8, 9, 12, 18]. Exceptions include [5], who ran live experiments on organic search results, and [6], who used human ratings of relevance.

Intuitively, one might expect that the model that does best at predicting CTR ($\Pr(C_i = 1)$) offline will also do best at predicting relevance ($\Pr(C_i = 1|E_i = 1)$). However, this is not the case. Consider the results from Sec-

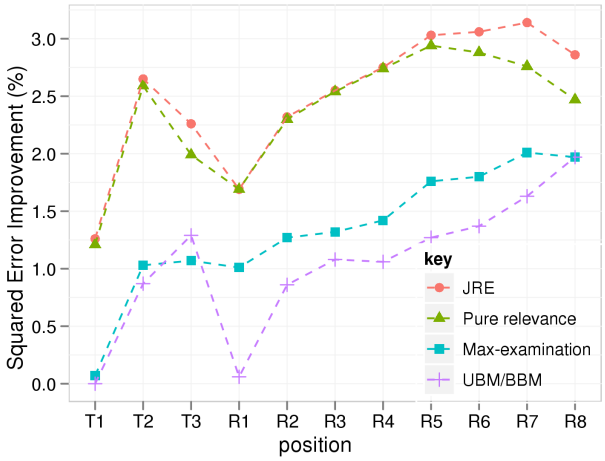
[5]In the initial version of the paper, we had mapped each model into the corresponding model in odds space, and used logistic regression – glm() in package:stats in R – to fit the parameters. Based on reviewer feedback, we decided to directly fit the parameters. The results in the initial version of the paper were almost identical to those presented below, since CTR for sponsored links is sufficiently less than 1, and with a small enough range, that effectively odds is a linear function of probability.

[6]The only exception is at T3, where max-examination does slightly worse. Whether there was a click on $T1$ or $T2$ is not purely a predictor of examination, it's also a predictor of instance relevance of $T3$. Our guess is that, for $T3$, the instance relevance component of this signal is more important than the increase in probability of examination due to the user clicking on a rhs result.

(a) Log likelihood, improvement over baseline.



(b) Squared error, improvement over baseline.



(c) Absolute error, improvement over baseline.

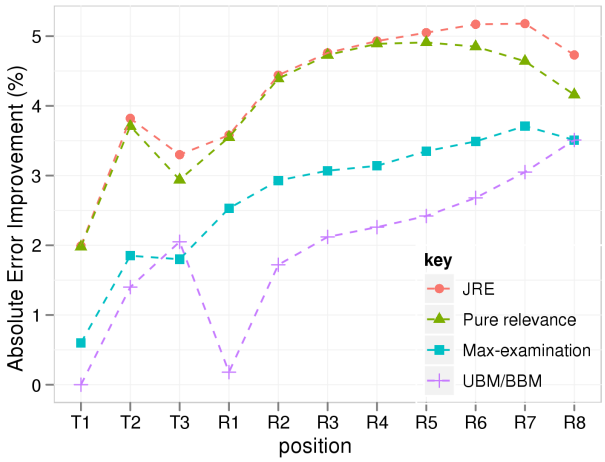


Figure 3: Accuracy for the 3-8 configuration.

	Log Likelihood	Squared Error	Absolute Error
UBM/BBM	1.82 ± 0.05	0.44 ± 0.03	0.75 ± 0.03
Max-Exam.	2.82 ± 0.07	0.52 ± 0.05	1.11 ± 0.03
Relevance	3.22 ± 0.08	1.16 ± 0.07	1.88 ± 0.04
JRE	3.34 ± 0.09	1.21 ± 0.07	1.98 ± 0.04

Figure 4: Percentage improvement over all configurations, with 99% confidence intervals.

tion 5.2. For predicting CTR, pure relevance dominated max-examination which dominated baseline. However, baseline and pure relevance yield identical relevance estimates, they only differ in the CTR estimates. So if max-examination is worse than baseline and pure relevance at predicting relevance, then max-examination and baseline are a pair of models where one is better at CTR and the other at relevance. If max-examination is better than baseline and pure relevance in predicting relevance, then max-examination and pure relevance form a similar pair, with one better at CTR and the other at relevance.

To get intuition on how a model might do better at CTR but worse at relevance, consider the following example. Assume an oracle that outputs 1 if the user clicked, and 0 else. Consider a browsing model that uses the oracle's output as $\Pr(E_i = 1)$, the probability of examination. For any result, $\Pr(E_i = 1)$ equals 1 if the user clicked and 0 otherwise, so we get $\Pr(C_i = 1|E_i = 1) = 1$. Every result has the same relevance. This model will have perfect accuracy in predicting CTR, but terrible accuracy for relevance. We give a more realistic example in Appendix A, that shows treating clicks on other results as predicting examination, when they are (partly) predicting instance relevance, can be similarly problematic.

We now present results using live experiments to determine which of the two models, baseline/pure relevance, and max-examination are better at predicting relevance. Sponsored results are ranked by expected revenue per impression, i.e, bid × relevance; results with higher expected revenue are shown in higher positions where they are more likely to be examined by users. A model that yields more accurate relevance estimates should result in a more accurate ranking of results, and therefore higher revenue and CTR.

6.1 Methodology

As in Section 5.1, we used Google's production system for predicting relevance of sponsored results as the baseline model. Since we are now evaluating relevance, the production system also serves as the pure relevance model (since pure relevance and baseline have identical relevance estimates).

For the max-examination model, we estimated $\gamma_{i,e(i)}$ from the logs (separately for each configuration and position), using the same methodology as in Section 5.1. We can now use the logged probability of examination (from the baseline model) and $\gamma_{i,e(i)}$ to get the probability of examination with the max-examination model. We used Sawzall [13] to compute, for each sponsored result, the cumulative probability of examination with both the baseline model and the max-examination model, over a week of data (without sampling). Let these values be E_b and E_m respectively. For each result, we multiplied the relevance scores from the baseline model by E_b/E_m to approximate the relevance scores we would have obtained had we trained a machine learning model to predict relevance directly using the max-examination browsing model.[7] We selected the 2 million most significant changes, where significance was defined as the number of clicks for that result times the change in relevance. This subset covered a substantial majority of the total significance of all the changes.

We applied these 2 million adjustments to the baseline model to get a reasonable proxy to the max-examination model. In particular, the direction of the difference in accuracy between this model and the baseline should be the same as between a trained-from-scratch max-examination model and the baseline – and what we care about (in this paper) is not the exact magnitude of the difference in accuracy, but only about understanding which one is better at predicting relevance, max-examination or the baseline?

6.2 Results

We ran a live experiment [17] on a small fraction of the google.com sponsored search traffic, and compared the metrics of the baseline and max-examination models. We found that baseline/pure relevance had better revenue and CTR than max-examination, with the results being statistically significant. In our system, changes in revenue and CTR could also be partly due to other factors, in particular, the tuning of the function for determining when to show sponsored results in the top slot. However, the results remained consistent through several retunings, thus we are confident that the results do reflect the accuracy of the models in predicting relevance.[8]

Since pure relevance did significantly better than max-examination at predicting CTR, it is not surprising that pure relevance also did better at predicting relevance. However, the results would have been very surprising if we had not presented the pure relevance model, and treated this solely as a comparison between the baseline and max-examination. From that perspective, max-examination does better at predicting CTR by leveraging co-click information, but because it incorrectly assigns credit to examination instead of relevance, actually does worse at predicting relevance.

7. CONCLUSIONS

Past work on user browsing models assumed that changes in CTR when conditioned on clicks in prior positions are due to changes in probability of examination. We showed that for sponsored search results, this fundamental assumption is contradicted by the data. We presented a plausible alternate conjecture: that relevance of the result for that query instance is strongly correlated with clicks on other results, and is responsible for a substantial portion of the changes in conditioned CTR. We proved this conjecture by finding a subset of query instances where the examination probability for certain positions is known to be close to 1, and showing that clicks on prior results still resulted in dramatic increases in CTR.

[7]Recall that $r_\phi(i) = \sum C_{\phi(i)} / \sum E_{\phi(i)}$. So we multiply by E_b to get $\Pr(C_i = 1)$ and then divide by E_m.

[8]At the time we ran the experiments, we expected max-examination to be more accurate than baseline. Thus we were highly motivated to get max-examination to work. This paper came about from our efforts to understand why we did not succeed!

We came up with new user browsing models that model changes in CTR (when conditioned on clicks in other positions) as caused by changes in instance relevance, or both relevance and examination. We also came up with an enhanced version of the UBM model, max-examination, that leverages information from both prior clicks as well as clicks below the current position, and predicts CTR better than the UBM model. Our new model, pure relevance, that attribute changes solely to instance relevance does significantly better at predicting CTR than the models that attribute CTR change solely to examination. In fact, the pure relevance model does only slightly worse than the more general model that attributes CTR change to both instance relevance and examination. This implies that changes in instance relevance account for a substantial portion of the change in CTR when conditioned on prior clicks.

Finally, we showed that evaluating user browsing models solely using offline analysis of CTR prediction can be problematic. A user browsing model may leverage information about clicks on other results (or other information about other results) to get superb accuracy when predicting CTR offline, but such an analysis cannot reveal whether the model is also correctly attributing credit between relevance and examination. If the model incorrectly attributes credit, it could end up with estimates of relevance and examination that are not very accurate in isolation, but whose product (CTR) is indeed accurate. We demonstrate that this is not purely theoretical, but indeed an important practical issue, by comparing the baseline/pure relevance and max-examination models in live experiments. Although the max-examination model does much better at CTR prediction in offline analysis, it does worse than the baseline/pure relevance models in predicting relevance, and therefore worse in live experiments. This reinforces our earlier conclusion that relevance is a key driver of changes in CTR when conditioned on other clicks, and also shows that directly evaluating relevance (through live experiments or human ratings) is an indispensable part of the evaluation of any user browsing model.

Future Work.

Our findings open up several directions for future work.

It would be interesting to see whether CTR changes for organic search results (when conditioned on prior clicks) are also substantially due to changes in instance relevance.

Quantitatively assigning credit between instance relevance and examination appears quite difficult. One approach might be to look at the values of the features corresponding to instance relevance and examination in the JRE model. However, these features are strongly correlated. Hence we found that with either iterative fitting or logistic regression, the values of the features are sensitive to the details of the algorithm and regularization, even though the final predictions are quite insensitive. Quantitative assignment of credit remains a challenging open problem.

A natural next step would be to generalize the pure relevance and JRE models to incorporate information about the relevance of prior results (Section 2.3), or the satisfaction of the user with the prior clicked results (Section 2.4). In particular, the session utility model [6] is intuitively appealing, but does not model examination or perceived relevance. A model that includes the key insights of both JRE and the session utility model would be very elegant.

8. REFERENCES

[1] Enquiro research, search engine results 2010. http://app.marketo.com/lp/enquiro/search-engine-results-2010.html?source=Search_Engine_Results_2010_whitepaper. (Free after registration.).

[2] E. Agichtein, E. Brill, and S. T. Dumais. Improving web search ranking by incorporating user behavior. In *ACM SIGIR Conference*, 2006.

[3] A. Aula and K. Rodden. Eye-tracking studies: more than meets the eye. http://googleblog.blogspot.com/2009/02/eye-tracking-studies-more-than-meets.html.

[4] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *World Wide Web Conference*, 2009.

[5] N. Craswell, B. Ramsey, M. Taylor, and O. Zoeter. An experimental comparison of click position-bias models. In *ACM Conf. on Web Search and Data Mining (WSDM)*, 2008.

[6] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *ACM Conf. on Web Search and Data Mining (WSDM)*, 2010.

[7] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *ACM SIGIR Conference*, 2008.

[8] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, and C. Faloutsos. Click chain model in web search. In *World Wide Web Conference*, 2009.

[9] F. Guo, C. Liu, and Y. Wang. Efficient multiple-click models in web search. In *ACM Conf. on Web Search and Data Mining (WSDM)*, 2009.

[10] T. Joachims. Optimizing search engines using clickthrough data. In *ACM Conf. on Knowledge Discovery and Data Mining (KDD)*, 2002.

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

[12] C. Liu, F. Guo, and C. Faloutsos. Bbm: Bayesian browsing model from petabyte-scale data. In *ACM Conf. on Knowledge Discovery and Data Mining (KDD)*, 2009.

[13] R. Pike, S. Doward, R. Griesemer, and S. Quinlan. Interpreting the data: Parallel analysis with sawzall. *Scientific Programming*, 13(4):277–298, 2005.

[14] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *ACM Conf. on Knowledge Discovery and Data Mining (KDD)*, 2005.

[15] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *World Wide Web Conference*, 2007.

[16] D. Sculley, R. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *ACM Conf. on Knowledge Discovery and Data Mining (KDD)*, 2009.

[17] D. Tang, A. Agarwal, D. O. Brien, and M. Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *ACM Conf. on Knowledge Discovery and Data Mining (KDD)*, 2010.

[18] Z. Zhu, W. Chen, T. Minka, C. Zhu, and Z. Chen. A novel click model and its applications to online advertising. In *ACM Conf. on Web Search and Data Mining (WSDM)*, 2010.

APPENDIX

A. EXAMPLE

We give a concrete example of how UBM (Equation 4) can be more accurate than the baseline model (Equation 2) in predicting CTR, but worse in predicting relevance.

Setup.

Assume that all changes in CTR when conditioned on other clicks are due to changes in instance relevance, not due to changes in examination. Let there be exactly 2 result positions, with examination probabilities

$$\begin{aligned}\Pr(E_1 = 1) &= \alpha_1 = 1\\ \Pr(E_2 = 1) &= \alpha_2 = 0.5\end{aligned}$$

Let the UBM parameters for the second position be:[9]

$$\begin{aligned}\beta_{2,p(1)=0} &= 0.5\\ \beta_{2,p(1)=1} &= 2\end{aligned}$$

We now consider a specific pair of results $s_1 = \phi(1)$ and $s_2 = \phi(2)$, with historical CTR of 0.2 and 0.1 respectively. Let there be 100 impressions. Since UBM is identical to baseline for the first position, we focus on s_2.

Baseline Model.

Relevance: Recall that in this scenario all CTR changes are due to changes in instance relevance. Thus the baseline model will correctly estimate $\Pr(E_2 = 1) = 0.5$, and correctly estimate relevance of s_2 as $\Pr(C_2 = 1|E_2 = 1) = \frac{0.1}{0.5} = 0.2$.

CTR: The baseline model does not make use of prior clicks, and will estimate $\Pr(C_2 = 1) = 0.1$. The absolute error (summed over 100 impressions) is:

$$10 \times |1 - \Pr(C_2 = 1)| + 90 \times |1 - \Pr(C_2 = 0)| = 18$$

UBM.

Relevance: Since $\Pr(C_1 = 1) = 0.2$ for s_1, UBM estimates $\Pr(E_2 = 1)$ as:

$$\begin{aligned}&\Pr(C_1 = 1)\alpha_2\beta_{2,p(1)=1} + \Pr(C_1 = 0)\alpha_2\beta_{2,p(1)=0}\\ &\quad = 0.2 \times 0.5 \times 2 + 0.8 \times 0.5 \times 0.5 = 0.4\end{aligned}$$

and therefore estimates relevance as $\Pr(C_2 = 1|E_2 = 1) = \frac{0.1}{0.4} = 0.25$. Notice that UBM ends up with an inaccurate estimate of relevance.

CTR: UBM model will estimate:

$$\begin{aligned}\Pr(C_2 = 1|C_1 = 1) &= \alpha_2\beta_{2,p(1)=1}r_{\phi(2)}\\ &= 0.5 \times 2.0 \times 0.25 = 0.25\end{aligned}$$

$$\begin{aligned}\Pr(C_2 = 1|C_1 = 0) &= \alpha_2\beta_{2,p(1)=1}r_{\phi(2)}\\ &= 0.5 \times 0.5 \times 0.25 = 0.0625\end{aligned}$$

Notice that the overall estimate for $\Pr(C_2 = 1) = \Pr(C_1 = 1) \times 0.25 + \Pr(C_1 = 0) \times 0.0625 = 0.1$ is correct. The estimates of $\Pr(C_2 = 1|C_1 = 1)$ and $\Pr(C_2 = 1|C_1 = 0)$ are also correct. The error in examination probability is exactly canceled out by the error in the relevance estimate.

The absolute error (over 100 impressions) is:

$$\begin{aligned}\text{Error} &= 5 \times |1 - \Pr(C_2 = 1|C_1 = 1)|\\ &\quad + 15 \times |1 - \Pr(C_2 = 0|C_1 = 1)\\ &\quad + 5 \times |1 - \Pr(C_2 = 1|C_1 = 0)|\\ &\quad + 75 \times |1 - \Pr(C_2 = 0|C_1 = 0)|\\ &= 5 \times 0.75 + 15 \times 0.25 + 5 \times 0.9375 + 75 \times 0.0625\\ &= 16.875\end{aligned}$$

The absolute error is less, though the relevance estimate is worse. It is easy to extend this example such that the positions of s_1 and s_2 exchange in live serving (by appropriately choosing bids for s_1 and s_2, resulting in lower revenue and CTR in live experiments.

B. REPEATABILITY

We discuss the repeatability of our analysis and experiments, assuming access to logs from a search engine. The key point is that while we may not have provided sufficient details for someone to exactly replicate what we did, our results do not depend on those omitted details, e.g, someone could use a different machine learning algorithm for predicting relevance, and we would expect them to get the same results.

In Section 3, recall that the ratio $L_i(1)/L_i(0)$ only depended on any errors in relevance estimation being roughly evenly distributed across the two sets. Thus we expect that one could use any reasonable machine learning algorithm for predicting relevance (e.g., a production system), and still get the same result wrt whether $L_i(1)/L_i(0)$ is different from 1 with statistical significance.

Similarly, in Section 5, we treat the output of the baseline model as given, and only fit the co-click dependent parameters. So the results (ordering of the models wrt accuracy) should again be independent of which machine learning algorithm is used for the baseline. Thus one can take the output of any production system for predicting relevance as given, and repeat our experiment.

In Section 6, it should be straightforward to compute the adjustments to relevance for max-examination if the production system uses a browsing model similar to the baseline model. If the production system uses a different browsing model, then one can use our methodology to first estimate the corrections to get relevance estimates for baseline, and then estimate a second correction to get the relevance estimates for max-examination. However, it may or may not be easy to run a live experiment with these corrections, based on the available infrastructure. If it is difficult to run live experiments, measuring relevance using human raters would be as effective, and would in fact nicely complement our experiments.

[9] An example setting that yields these parameters is when the average CTR in position 1 is 1/3, and $\frac{\Pr(C_2=1|C_1=1)}{\Pr(C_2=1|C_1=0)} = 4$ for any pair of results. Solving $\Pr(E_2 = 1) = \Pr(C_1 = 1)\alpha_2\beta_{2,p(1)=1} + \Pr(C_1 = 0)\alpha_2\beta_{2,p(1)=0}$ yields these values.