Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data

Yisong Yue*
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
yyue@cs.cornell.edu

Rajan Patel
Google Inc.
1600 Ampitheatre Parkway
Mountain View, CA 94043
rajan@google.com

Hein Roehrig
Google Inc.
1600 Ampitheatre Parkway
Mountain View, CA 94043
hein@google.com

ABSTRACT

Leveraging clickthrough data has become a popular approach for evaluating and optimizing information retrieval systems. Although data is plentiful, one must take care when interpreting clicks, since user behavior can be affected by various sources of presentation bias. While the issue of position bias in clickthrough data has been the topic of much study, other presentation bias effects have received comparatively little attention. For instance, since users must decide whether to click on a result based on its summary (e.g., the title, URL and abstract), one might expect clicks to favor "more attractive" results. In this paper, we examine result summary attractiveness as a potential source of presentation bias. This study distinguishes itself from prior work by aiming to detect systematic biases in click behavior due to attractive summaries inflating perceived relevance. Our experiments conducted on the Google web search engine show substantial evidence of presentation bias in clicks towards results with more attractive titles.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance Feedback

General Terms

Experimentation, Human Factors, Measurement

Keywords

Presentation Bias, Implicit Feedback, Click Logs

1. INTRODUCTION

Evaluating the quality of search result rankings has traditionally relied on explicit human judgments or editorial labels. While effective, the prohibitive cost of acquiring human judgments makes it difficult to apply these evaluation approaches at scale for large search services such as commercial search engines. It is also impractical to acquire manually labeled query/document relevance judgments for every possible retrieval domain, such as medical, law, physics, etc. As such, collecting implicit user feedback – typically click logs – has grown tremendously in popularity in recent years.

But what can we interpret from clicks? It is well known that users are biased towards clicking on higher ranked results (cf. [14]) – this is the so called position bias effect. But users must also judge relevance based on summaries rather than the actual results themselves. Summaries typically include titles, URLs, and query dependent abstracts (or snippets), and often have matching query terms highlighted using boldface font. As a simple thought experiment, consider two equally relevant results with one having more bolded query terms in the title. Intuitively, we might expect click behavior to favor the more attractive title. Thus, even in the absence of position bias, a result's perceived relevance might noticeably differ from its actual relevance.

A particularly illuminating study by Clarke et al. [6] found that click inversions (when a lower ranked document receives more clicks than a higher ranked one) cannot be entirely explained by the lower ranked document being more relevant. They found that click inversions tend to co-occur with additional factors such as lower ranked documents having comparatively more matching query terms in the titles.

In this paper, we quantify the effect of bolded keyword matches in the title and abstracts on the attractiveness of the result. Our analysis controls for both position bias and rated relevance (as judged by human raters), and is based on data collected using a portion of search traffic from the Google web search engine. To control for position bias, we collected data using the FairPairs algorithm [19], which allows us to interpret clicks as preference judgments between two documents. To control for quality, we gathered human preference judgments for a subset of our clickthrough data. Our findings show that clicks are measurably biased by attractive titles inflating perceived relevance. We will also discuss possible ways to adjust for title attractiveness bias.

For the rest of this paper, we proceed by first overviewing related work. Section 3 describes our data collection methodology, which includes collecting both clickthrough data and explicit human judgments. Section 4 describes our analysis on measuring attractiveness bias. We then discuss ways to adjust for bias in Section 5, and conclude with a discussion of limitations and avenues for future work.

2. RELATED WORK

The earliest studies on evaluating retrieval systems commonly used the Cranfield methodology (such as many tasks in TREC [23]), which relies on explicit relevance judgments

*This study was conducted while the first author was an intern at Google.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

collected from human experts. Given a query and a ranking of documents produced by a retrieval system, metrics such as Average Precision, NDCG and Mean Reciprocal Rank (cf. [9]) can be used to evaluate ranking quality.

Unfortunately, acquiring explicit relevance judgments is quite costly and time consuming, making it difficult to apply at scale for large search services such as commercial search engines. It is also infeasible to collect explicit relevance judgments across a variety of search domains such as patent or medical search. Additionally, some metrics based on human judgments have been shown to not necessarily correlate with more user-centric performance measures [22]. Consequently, collecting usage logs such as clickthrough data has become increasingly popular in recent years.

Accurate interpretation of usage logs is an area of intense study (see [15] for an overview). The problem of position bias – that users tend to click more on higher ranked results – is well documented in the literature [14]. Since users typically scan results in rank order, clicking on higher ranked results does not necessarily indicate relevance. One way to leverage usage data as an evaluation metric is by adjusting for bias post-collection [24]. One can also use implicit feedback to design new signals (or features) to be used by ranking functions [1, 5].

One can also interactively modify rankings to preemptively control for position bias. For example, if two competing results were randomly shown in the the original and reversed orders equally often, then clicks might correspond to relative preference between the two results (e.g., is result A more relevant than result B?). Prior work have considered interactively extracting both pairwise document preferences [19] as well as ranking level preferences [21]. We will use the FairPairs algorithm proposed by Radlinski & Joachims [19] to control for position bias and collect document-level preference feedback. Craswell et al. [8] also gathered feedback from swapping adjacent pairs.

Our approach is related to prior work on probabilistic user behavior models [2, 8, 10, 5]. Unlike previous approaches, the click models we consider aim to tease apart the effect of attractiveness bias from that of other factors such as relevance and position bias. Most prior work studying the impact of result summaries focused instead on how search difficulty can vary depending on summary quality [6, 16, 17].

The problem of learning ranking functions is a topic of great interest in the machine learning community. Of particular relevance to this study are methods which optimize over labeled pairwise preferences (e.g., result A better than result B). Existing approaches build upon a variety of conventional techniques including boosting [11], SVMs [13, 4], and neural networks [9]. Implicit preference feedback (such as the data gathered for this study) can be naturally integrated into the objective functions of such approaches.

3. DATA COLLECTION

3.1 Collecting Fair Pairs

Since users typically scan results in rank order, one can reasonably interpret clicks on a lower ranked result as implicit preference feedback over an unclicked higher ranked results [14]. One way to control for position bias is by randomly showing two adjacent results in either the original or swapped order. Since both results appear at the both positions equally often (in expectation), then intuitively, we can simply count clicks to determine relative preference. We use the term **Fair Pair** to denote the pairing of two rank-adjacent results (e.g., as determined by the incumbent ranking function). A Fair Pair presented at rank i indicates that its two results were presented at ranks i and $i+1$.

Algorithm 1 FairPairs Randomization

Let $R \leftarrow (d_1, \ldots, d_n)$ be the results for some query.
Randomly choose $k \in \{0, 1\}$ with uniform probability.
if $k = 0$ **then**
  for $i \in \{1, 3, 5, \ldots\}$ **do**
    Swap d_i and d_{i+1} in R with 50% probability.
  end for
else
  for $i \in \{2, 4, 6, \ldots\}$ **do**
    Swap d_i and d_{i+1} in R with 50% probability.
  end for
end if
Present R to the users, recording clicks on results.

The FairPairs algorithm for collecting clickthrough data, first proposed by Radlinski & Joachims [19], is formally described in Algorithm 1. For each query, the original ranking is first partitioned into Fair Pairs via one of two schemes chosen at random. The first scheme (1-2 partitioning) groups results at ranks 1 & 2, 3 & 4, 5 & 6, and so forth into Fair Pairs, whereas the second scheme (2-3 partitioning) groups results at ranks 2 & 3, 4 & 5, 6 & 7, and so forth into Fair Pairs (leaving rank 1 unaffected). Afterwards, each Fair Pair is presented to the user in either its original or swapped order with 50% probability. The decision on whether to swap each Fair Pair is made independently of other Fair Pairs.

For example, given original ranking (A,B,C,D,E,F,G), a random swapping using the 1-2 partitioning might result in (B,A,C,D,F,E,G), whereas a random swapping using the 2-3 partitioning might result in (A,B,C,E,D,G,F). Its relatively low impact on presented rankings makes the FairPairs algorithm an attractive way to control for position bias on click data gathered from commercial search traffic .

We call a click on the bottom presented result in a Fair Pair a **bottom click**, and a click on the top presented result a **top click**. Radlinski & Joachims showed that clicks correspond to unbiased relative preference feedback under assumptions which we paraphrase in simplified form:

ASSUMPTION 1. *User click behavior depends on the actual relevance of documents. In other words, perceived relevance does not deviate from actual relevance.*

ASSUMPTION 2. *Suppose in a Fair Pair (d_i, d_j) that document d_i is more relevant than d_j. Then the probability of bottom (or top) clicking on d_i is greater than the probability of bottom (or top) clicking on d_j.*

Under these assumptions, it can be shown that, when using FairPairs randomization on (d_i, d_j), d_i is more relevant than d_j *if and only if* clicks on d_i are more likely than clicks on d_j (see Theorem 1 in [19]). We will verify the validity of these assumptions. In particular, we find that Assumption 1 does not hold in our data set due to attractiveness bias.

3.1.1 Implementation Details

In our implementation, not all rankings could be modified. Many queries, often navigational ones, have one very relevant result with much higher quality than the other results. In such cases, we did not modify the original ranking.

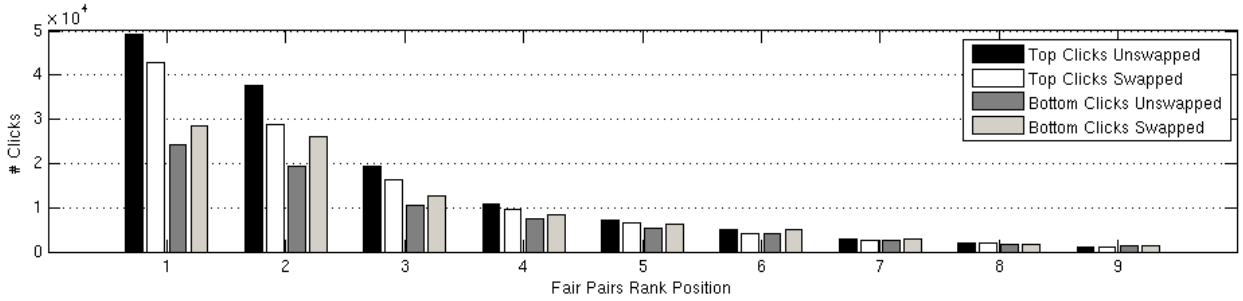

Figure 1: Absolute click counts for Fair Pairs from rank position 1 to rank position 9. Fair Pairs at position i randomly swap results at rank i and $i+1$.

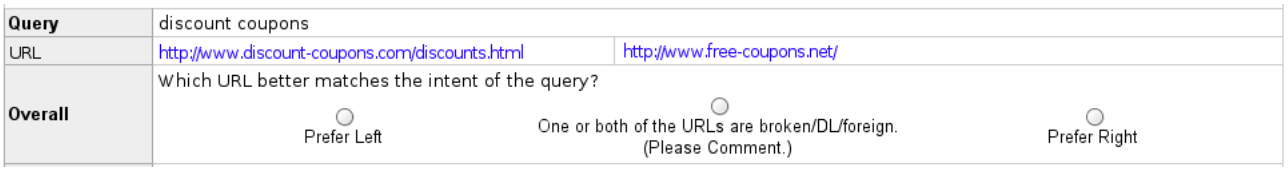

Figure 2: An example rating task. Human judges are not given the result summaries and must determine preference based purely on the site content and their interpretation of query intent.

Table 1: Click agreement with human raters

	Top Click	Bottom Click
Unswapped	217/122 (64%)	106/91 (54%)
Swapped	152/136 (52%)	170/124 (58%)
Total	353/274 (56%)	276/215 (56%)

Other overriding modifications include queries that required injecting image, video or news results. Thus, our collected data set might not be completely representative of the true distribution of user behavior.

3.1.2 Fair Pairs Data set

We collected Fair Pairs data on a portion of Google web search traffic from 8/1/2009 to 8/20/2009. In total 439,246 clicks were collected, comprising of 255,112 top clicks and 184,134 bottom clicks. Figure 1 shows the breakdown for the top 9 Fair Pairs. Unswapped and swapped refer to Fair Pairs where the original ordering was unchanged and swapped, respectively. Clicks were anonymized such that they could not be linked with a particular user or cookie.

Note that, at each rank position, swapped bottom clicks are more likely than unswapped bottom clicks. This implies that click behavior tends to favor the original higher ranked result regardless of swapping. Given a high-quality search engine (which typically ranks more relevant results higher), then this provides empirical support for Assumption 2.

We also note that results at rank 10 received more clicks than results at rank 9 (see Fair Pairs at rank 9 in Figure 1). We conjecture that this is caused by users skipping to the bottom results and scanning in reverse rank order (before deciding to load the second page of results).

3.2 Collecting Human Judgments

We collected explicit preference judgments from human raters on a small subset of Fair Pairs. These preference judgments can be considered ground truth, and can be used to verify the accuracy of Fair Pairs data, as well as any presentation bias effects the click data may contain.

Figure 2 shows an example rating task. For each selected Fair Pair, a human rater is presented with the query and two competing URLs, and must choose which result to prefer after examining both web pages. The two URLs are shown to raters in randomized order. Each selected Fair Pair is judged by five raters and each rater must explicitly prefer one result since we already have a clear click preference.

We selected 1150 Fair Pairs in total to be rated, which includes 650 top clicks and 500 bottom clicks. Of these 32 were discarded due to issues such as broken URLs and non-English results. The inter-judge agreement was 70% as measured by randomly sampling a query and two ratings for that query. This is a relatively difficult rating task since the two competing results are typically similar in quality.

Table 1 summarizes the agreement between clicks and human raters. We find that top and bottom clicks correlate about equally well with rater preferences. The rater agreement for both top clicks (56% $\pm$0.04) and bottom clicks (56% $\pm$0.04) are positive with 95% confidence. One reason for the seemingly low agreement is the fact that raters must infer intent using only the query. For instance, a query might be associated with a primary (75% of users) and a secondary intent (25% of users), whereas raters always judge based on the primary. Overall, it seems both top and bottom clicks can be useful indicators of relative quality.

4. ATTRACTIVENESS BIAS

The fundamental research question we ask is, "how does attractiveness or perceived relevance impact click behavior?" In this study, we focused on the attractiveness of titles and abstracts as measured by the number of matching query terms. Since our click data was collected on result pages

Table 2: Fitting Rated Clicks Model using multivariate logistic regression and 500 bootstrap samples. * indicates statistical significance at the 0.05 level.

Parameter	Mean	95% CI	Interpretation
w_0	0.653 *	±0.183	Original higher result clicked on 68% when preferred by human raters and presented on top.
w_T	0.150 *	±0.120	Click odds increase by x1.16 from base rate for every additional bolded title term.
w_A	0.039	±0.120	Click odds increase by x1.04 from base rate for every additional bolded abstract term.
w_S	-0.435 *	±0.209	Being presented on bottom decreases click odds by x0.65 from base rate.
w_H	-0.360 *	±0.215	Raters preferring original lower result decreases click odds by x0.70 from base rate.

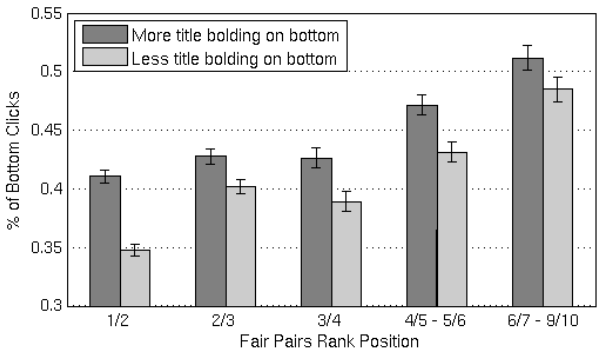

Figure 3: Comparing bottom click frequency when the bottom result has more and less title bolding. Fair Pairs at position $i/i+1$ randomly swap results at ranks i and $i+1$.

with query bolding (all matching query terms in the summary are displayed in boldface), we will interchangeably refer to matching query terms in title as **title bolding**, and in the abstract as **abstract bolding**. Note that different systems generate abstracts differently, so specific measurements (from this study) that are based on abstracts may not be broadly applicable.

As a motivating case study, we first measured click percentages under different conditions such as how often the clicked result had more title bolding than the unclicked result. Figure 3 shows the bottom click percentages conditioned on whether the bottom result has more or less title bolding. We observe a clear trend of clicks favoring more title bolding across all rank positions. Note that bottom clicks are typically less likely than top clicks (e.g. due to position bias). We also observe a similar trend for top clicks. All measurements are made using 500 bootstrap samples of our entire Fair Pairs data set.

Of course, one should expect query/title similarity to be correlated with relevance. It is therefore unsurprising that clicks would favor results with more title bolding. To control for quality, we estimated a click model using human rated data. We will show results confirming that title attractiveness positively biases click behavior. Using an additional assumption, we can further extend our click model to incorporate all (unlabeled) Fair Pairs data.

4.1 Click Analysis using Human Ratings

In order to measure the effects of attractiveness, we estimated a logistic regression model [12] to predict click probabilities using human rated data. Using this model, we can explicitly quantify the impact of title and abstract bolding on click behavior beyond its correlation with rated quality. Recall from Section 3.2 and Figure 2 that human raters must examine both web pages, and do not view result summaries. The Rated Clicks Model is formally defined as

Definition 1. ***(Rated Clicks Model)***

$$P = \frac{1}{1+\exp(-z)},$$

where

$$z = w_0 + w_T X_T + w_A X_A + w_S X_S + w_H X_H,$$

and

$P =$ *probability of clicking on original higher result*

$X_T =$ *title bolding difference of original higher - lower*

$X_A =$ *abstract bolding difference of original higher - lower*

$$X_S = \begin{cases} 0 & \textit{in top of presented Fair Pair.} \\ 1 & \textit{in bottom of presented Fair Pair.} \end{cases}$$

$$X_H = \begin{cases} 0 & \textit{original higher preferred by majority of raters.} \\ 1 & \textit{original lower preferred by majority of raters.} \end{cases}$$

By bolding difference we mean the number of bolded words in the originally higher results minus the the number of bolded words in the originally lower result.

This model predicts the probability of clicking on the original higher ranked result based on 4 factors: the title bolding difference, abstract bolding difference, whether the Fair Pair was swapped, and human-rated relative quality. By controlling for quality and position, statistically significant positive estimates of w_T and w_A would imply that click behavior is biased towards more attractive titles and abstracts, respectively, beyond their correlation with relevance.

Table 2 shows the results of fitting the Rated Clicks Model using human rated Fair Pairs data. All estimates are made using 500 bootstrap samples on the human rated data. To help interpret this model, we first note that the mean estimate of the click rate on original top results in the base condition (presented on top and preferred by raters) is

$$\frac{1}{1+\exp(-w_0)} \approx 0.68.$$

When raters preferred the original lower result ($X_D = 1$), the click odds, $P(1-P)$, decrease from base odds by a factor of $\exp(w_H) \approx 0.65$. When the Fair Pair is swapped ($X_S =$

Table 3: Fitting Rated Agreement Model using multivariate logistic regression and 500 bootstrap samples. * indicates statistical significance at the 0.05 level.

Parameter	Mean	95% CI	Interpretation
w_0	0.258 *	±0.062	Original higher result is preferred by raters 57% overall.
w_T	0.018	±0.060	Preference odds increase by x1.02 from base rate for every additional bolded title term.
w_A	0.058	±0.060	Preference odds increase by x1.06 from base rate for every additional bolded abstract term.

Table 4: Human rater agreement with original ranking as categorized by title bolding difference.

Title Bolding Diff	Raters Agree	Disagree
Orig Higher - Lower ≤ -2	45	25
Orig Higher - Lower $= -1$	71	82
Orig Higher - Lower $= 0$	375	279
Orig Higher - Lower $= 1$	98	72
Orig Higher - Lower ≥ 2	41	30
Total	630	488

1), then the click odds decrease from the base odds by a factor of $\exp(w_S) \approx 0.70$.

We can interpret the mean estimate of w_T to indicate that, for every additional bolded title term, the clicks odds of the original top result improve by a factor of $\exp(w_T) \approx 1.16$ over the base clicks odds after adjusting position and quality. Although the variance of the estimate is somewhat large due to the limited amount of data, we can nonetheless conclude with 95% confidence that title bolding has a positive effect on clicks. This suggests that user click behavior is affected by attractive titles inflating perceived relevance, which implies that Assumption 1 does not hold in practice.

We observe a slightly positive effect from abstract bolding, although the effect is not significant with 95% confidence. Fitting the Rated Clicks Model to predict click probabilities on the original lower results yields similar results.

4.2 Title Bolding vs Relative Quality

As touched on earlier, title bolding is correlated with quality, so overall, one should expect more clicks on results with more title bolding. Although it is relatively straightforward to detect attractiveness bias using human rated data to control for relative quality, for reasons of cost we can only collect such ratings on a small subset of Fair Pairs data.

On the other hand, virtually all successful search engines predict relevance using not only query/title similarity, but also a wide range of other features. For example, a document with relatively little title bolding might achieve a high ranking due to factors such as having relevant body content or high PageRank [18]. This leads us to consider the following assumption:

Assumption 3. *The relative quality between two adjacent documents, conditioned on their original rank positions (as computed by a high-quality ranking function), is independent of their query/title similarities.*

Although Assumption 3 is idealistic, if it approximately holds in practice, then it can be exploited to detect title attractiveness bias in unlabeled Fair Pairs data. We will justify the approximate validity of Assumption 3 using the human rated data. Intuitively, if rater agreement with the original ranking is strongly correlated with the title bolding difference between the two results, then Assumption 3 would be significantly violated. We can also restate the assumption with respect to abstract bolding:

Assumption 4. *The relative quality between two adjacent documents, conditioned on their original rank positions (as computed by a high-quality ranking function), is independent of their query/abstract similarities.*

Table 4 provides an insightful view of the human rated data. Overall, we observe no strong trend between title bolding difference and rater agreement with the original ranking. We confirm this observation by estimating a logistic regression model to predict rater agreement with the original ranking based on title and abstract bolding differences.

Definition 2. ***(Rated Agreement Model)***

$$P = \frac{1}{1 + \exp(-z)},$$

where

$$z = w_0 + w_T X_T + w_A X_A,$$

and

$P =$ *probability of raters favoring original higher result*

$X_T =$ *title bolding difference of original higher - lower*

$X_A =$ *abstract bolding difference of original higher - lower*

Table 3 shows the fitted parameters of the Rated Agreement Model using human rated data. Overall, the mean estimate of w_T (0.018 ± 0.06) suggests little to no correlation. The estimate on w_T remains virtually unchanged (and in fact becomes even more neutral) if we remove abstract bolding from the Rated Agreement Model. This suggests that Assumption 3 is approximately satisfied in practice. The effect from abstract bolding is somewhat stronger, so the validity of Assumption 4 is more questionable.

4.3 Click Analysis using All Fair Pairs

We now turn our attention to detecting attractiveness bias using the entire Fair Pairs data set. Figure 4 shows the bottom click percentages on swapped and unswapped Fair Pairs conditioned on whether the bottom result has more or less title bolding. In both cases, clicks clearly favor results with more title bolding. Note that swapped Fair Pairs present the original higher result (which is on average more relevant) at the lower rank. Using Assumption 3, we can assume that relative quality between adjacent results is independent of

title bolding. Since clicks clearly favor results with more title bolding, this implies that attractive titles (as measured by title bolding) noticeably biases clicks in the entire Fair Pairs data set. We also observed a similar trend for top clicks, as well as a stronger trend when restricting to Fair Pairs with title bolding differences of 2 or greater.

More formally, we estimated the effects of title and abstract attractiveness using a probabilistic model of click behavior. The All Clicks Model described below is analogous to the Rated Clicks Model considered in Section 4.1. In the absence of rated quality judgments, we instead rely on Assumptions 3 & 4 and assume that the expected quality difference between the original higher and lower results is independent of title and abstract bolding. Although our analysis in Section 4.2 casts doubt on the validity of Assumption 4, we shall see in the following that clicks are not strongly influenced by abstract bolding. We therefore do not require Assumption 4 in order to interpret most of the presentation bias effects detected by our model.

Since data is plentiful, we can consider many different factors simultaneously. In particular, the All Clicks Model will jointly consider the effects of rank-adjacent position bias at different rank positions as well as the effects of title and abstract bolding on both top and bottom clicks.

DEFINITION 3. ***(All Clicks Model)***

$$P = \frac{1}{1+\exp(-z)},$$

where

$$z = w_0 + w_T X_T + w_{\bar{T}} X_{\bar{T}} + w_A X_A + w_{\bar{A}} X_{\bar{A}} + \sum_{I \in \mathcal{I}} w_I X_I,$$

and

$P =$ *probability of clicking on original higher result*

$X_T =$ *unswapped title bolding diff of orig higher - lower*

$X_{\bar{T}} =$ *swapped title bolding diff of orig higher - lower*

$X_A =$ *unswapped abstract bolding diff of orig higher - lower*

$X_{\bar{A}} =$ *swapped abstract bolding diff of orig higher - lower*

$$X_I = \begin{cases} +1 & \textit{from unswapped in position group } I \\ -1 & \textit{from swapped in position group } I \\ 0 & \textit{not in position group } I \end{cases}$$

The position groups we considered are Rank 1, Rank 2, Rank 4-5, Rank 6-9, and Rank 10+, that is,

$$\mathcal{I} = \{\ 1,\ 2,\ 3,\ 4\text{-}5,\ 6\text{-}9,\ 10+\ \}.$$

For example, if a Fair Pair at rank 3 (comparing ranks 3 and 4) presented results in swapped order, then $X_{\{3\}} = -1$ and all other $X_I = 0$.

The All Clicks Model makes a distinction between presentation bias effects on top and bottom clicks. Since we model clicks on the original higher result, then predicted click probabilities in a swapped Fair Pair correspond to bottom click probabilities. For example, X_T is non-zero only for unswapped Fair Pairs. This allows us to examine whether title bolding affects top and bottom clicks differently.

Table 5 shows the details of our fitted model. We observe a statistically significant effect from title bolding for both

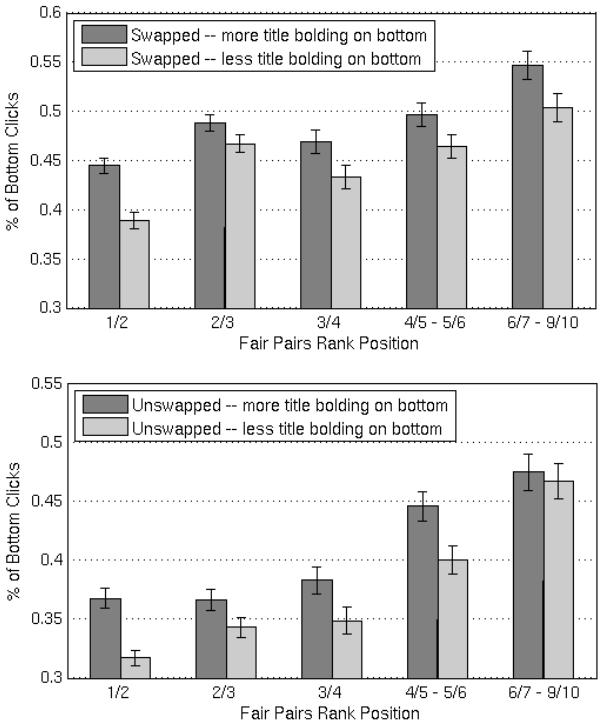

Figure 4: Comparing swapped and unswapped bottom click frequency when the bottom result has more and less title bolding. Fair Pairs at position $i/i+1$ randomly swap results at ranks i and $i+1$.

swapped and unswapped Fair Pairs. For example, for every additional bolded title term, the click odds, $P/(1-P)$, of swapped (bottom) clicks on the original higher results improves by a factor of $\exp(w_{\bar{T}}) \approx 1.06$. The mean estimate of the title bolding effect is smaller than the analogous estimate for the Rated Clicks Model (see Table 2), although the two estimates have overlapping confidence intervals. The smaller estimate is in part due to the All Clicks Model not controlling for relative quality and instead relying on Assumption 3. We observe only a weak effect from abstract bolding.

Note that title and abstract bolding effects are slightly stronger for bottom clicks than top clicks, although the differences are not significant. Intuitively, we should expect bottom clicks to be more susceptible to attractiveness bias since more attractive results can help counter the negative position bias effect from being presented lower in ranking.

Table 5 also allows us to estimate the relative impact of position versus title attractiveness across different rank positions. We can see that position effects are much stronger than title and abstract bolding effects.

5. FAIR PAIRS AS EVALUATION METRIC

Leveraging pairwise preference feedback – such as Fair Pairs data – can be an attractive way to evaluate retrieval systems. Such feedback can also be easily integrated into the objective functions of many learning to rank algorithms (e.g., [11, 4, 9]). However, attractiveness bias is a cause for concern. Recall from Figure 4 that if we used Fair Pairs feedback unmodified (e.g., evaluate retrieval systems based on

Table 5: Fitting All Clicks Model using multivariate logistic regression and 500 bootstrap samples. * indicates statistical significance at the 0.05 level.

PARAM.	MEAN	95% CI	INTERPRETATION
w_0	0.184 *	±0.007	Original higher result is clicked on 55% overall.
w_T	0.060 *	±0.008	Unswapped click odds increase by x1.06 from base rate for every additional bolded title term.
$w_{\bar{T}}$	0.061 *	±0.009	Swapped click odds increase by x1.06 from base rate for every additional bolded title term.
w_A	0.007	±0.009	Unswapped click odds increase by x1.007 from base rate for every additional bolded abstract term.
$w_{\bar{A}}$	0.014*	±0.008	Swapped click odds decrease by x1.01 for every additional bolded abstract term.
$w_{\{1\}}$	0.561 *	±0.011	Click odds increase by x1.75 from base rate if on top in presented Fair Pairs at rank 1.
$w_{\{2\}}$	0.390 *	±0.012	Click odds increase by x1.48 from base rate if on top in presented Fair Pairs at rank 2.
$w_{\{3\}}$	0.372 *	±0.016	Click odds increase by x1.45 from base rate if on top in presented Fair Pairs at rank 3.
$w_{\{4-5\}}$	0.198 *	±0.014	Click odds increase by x1.22 from base rate if on top in presented Fair Pairs at ranks 4-5.
$w_{\{6-9\}}$	0.009	±0.014	Click odds increase by x1.01 from base rate if on top in presented Fair Pairs at ranks 6-9.
$w_{\{10+\}}$	0.054 *	±0.009	Click odds increase by x1.06 from base rate if on top in presented Fair Pairs at ranks 10+.

the raw sum of Fair Pairs ordered correctly), then we might incorrectly conclude that our incumbent ranking function is undervaluing query/title similarity. In this section, we will discuss strategies to adjust for attractiveness bias.

Given explicit knowledge of users' preferences, one natural evaluation metric would be to compute the expected number of document pairs that are ordered correctly. We can express such a metric more formally as

$$Q(f) = \mathbf{E}_{(q,d_1,d_2,p)}\left[\mathbf{1}_p(f(q,d_1), f(q,d_2))\right], \quad (1)$$

where (q, d_1, d_2, p) is a tuple indicating the query, two competing documents, and explicit preference from a user (sampled from a population of users), and $\mathbf{1}_p(\cdot,\cdot)$ is an indicator function which equals 1 if the ordering produced by f agrees with user preference. However, since we only observe clicks, our proxy evaluation metric, in its simplest form, is

$$\tilde{Q}(f) = \mathbf{E}_{(q,d_1,d_2,c)}\left[\mathbf{1}_c(f(q,d_1), f(q,d_2))\right],$$

where $\mathbf{1}_c(\cdot,\cdot)$ is an indicator function which equals 1 if the ordering produced by f agrees with the click. This is clearly not an unbiased estimate of $Q(f)$ due to presentation bias effects such as from title attractiveness.

5.1 Adjusting using Click Odds

We can use the click models examined in Section 4 to correct for presentation bias. Conditioned on other factors (e.g., click position, swapped or unswapped, etc), the click probability can be modeled as

$$P = \frac{1}{1+\exp(-(w_T X_T + w_A X_A + \ldots))},$$

where w_T and w_A are the title and abstract bolding bias effects, respectively. One straightforward way to correct for title attractiveness bias is to weight each Fair Pairs data entry by $\exp(-w_T X_T)$ where X_T is the title bolding difference between the clicked and unclicked results. This yields the following approximation of (1),

$$\hat{Q}(f) = \mathbf{E}_{(q,d_1,d_2,c)}\left[e^{-w_T X_T}\mathbf{1}_c(f(q,d_1), f(q,d_2))\right].$$

5.2 Learning to Predict Human Agreement

We can also learn to predict human agreement on clicks. Having an estimate of human agreement (conditioned on the click) allows us to approximate (1) as

$$\hat{Q}(f) = \mathbf{E}_{(q,d_1,d_2,c)}\mathbf{E}_{p|q,d_1,d_2,c}\left[\mathbf{1}_p(f(q,d_1), f(q,d_2))\right].$$

Table 6: Predicting human agreement. Comparing accuracy of learned model from baseline.

MODEL	ACCURACY	TRUE POS	FALSE POS	TRUE NEG	FALSE NEG
Baseline	0.563	629	489	0	0
Trained	0.587	521	354	135	108

In a preliminary experiment, we trained logistic regression models to predict human agreement with clicks using the human rated Fair Pairs data. Our feature representation consists primarily of indicator step functions defined using raw signals such as title bolding differences and click position (similar to Section 4.2 in [25]). This representation allows linear models to fit a non-linear decision surface w.r.t. the raw signals. To mitigate the risk of overfitting, we also used bagging [3] to learn an ensemble of models. We measured accuracy using hold-out data via 10-fold cross validation. The results in Table 6 show a modest performance gain over the baseline which counts all clicks as agreement. As a compromise for higher accuracy on positive predictions, the learned model also incurs false negative errors.

6. DISCUSSION

In this paper, we examined the effects of result attractiveness on click behavior. We used the FairPairs algorithm to collect clicks in order to control for position. Our findings on human rated data show substantial evidence of presentation bias from title attractiveness as measured by bolded query terms in the title. It would be interesting and useful to identify more sophisticated ways to measure attractiveness; e.g., we have not considered the attractiveness of the displayed result URL. Its length, bolding, and recognizable domain may have a significant impact.

We further extended our analysis to incorporate all unlabeled Fair Pairs. Interpreting the All Clicks Model requires making an independence assumption between relative qual-

ity of adjacent results and title bolding differences. While this assumption appears justified for adjacently ranked results (as computed by a high-quality ranking function), we do not expect it to hold for non-adjacent results.

We also discussed possible ways to adjust for presentation bias in order to yield better evaluation metrics. We restricted our attention to methods which adjust off-line on pre-collected data. Interactively modifying results on-line to control for bias or obtain more accurate estimates is an interesting direction for future work. Our analysis also assumed that feedback collected from FairPairs is representative of all pairs of results. However, we only collected feedback for adjacent pairs and our feedback is limited to documents ranked highly by the incumbent ranking function. Another direction for future work is to incorporate active learning (cf., [20]) in order to gather a more representative sample of user preferences. Using Fair Pairs data for evaluation also does not account for result-set diversity [26, 7].

It would also be interesting to develop general models of browsing behavior which directly account for attractiveness bias. Existing browsing models [8, 10, 5] incorporate properties such as users typically browsing in rank order, intermediate clicks in a session being less relevant, and users often viewing one result below the last clicked result. We might see further benefit from incorporating attractive bias.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incroporating user behavior. In *ACM Conference on Information Retrieval (SIGIR)*, 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *ACM Conference on Information Retrieval (SIGIR)*, 2006.

[3] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with svms. *Information Retrieval Journal*, 2009.

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

[6] C. Clarke, E. Agichtein, S. Dumais, and R. White. The influence of caption features on clickthrough patterns in web search. In *ACM Conference on Information Retrieval (SIGIR)*, 2007.

[7] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. Mackinnon. Novelty and diversity in information retrieval evaluation. In *ACM Conference on Information Retrieval (SIGIR)*, 2008.

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

[9] P. Donmez, K. Svore, and C. Burges. On the local optimality of lambdarank. In *ACM Conference on Information Retrieval (SIGIR)*, 2009.

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

[11] Y. Freund, R. Iyer, R. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research (JMLR)*, 4:933–969, 2003.

[12] D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, 2000.

[13] T. Joachims. Optimizing search engines using clickthrough data. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 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 Transactions on Information Systems (TOIS)*, 25(2), April 2007.

[15] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *ACM SIGIR Forum*, 37(2):18–28, 2003.

[16] R. Khan, D. Mease, and R. Patel. The impact of result abstracts on task completion time. In *WWW Workshop on Web Search Result Summarization and Presentation*, 2009.

[17] M. Murata, H. Toda, Y. Matsuura, and R. Kataoka. Query-page intention matching using clicked titles and snippets to boost search rankings. In *ACM Joint Conference on Digital Libraries (JCDL)*, 2009.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[19] F. Radlinski and T. Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *Conference of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2006.

[20] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *ACM International Conference On Knowledge Discovery and Data Mining (KDD)*, 2007.

[21] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Conference on Information and Knowledge Management (CIKM)*, 2008.

[22] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasts. In *ACM Conference on Information Retrieval (SIGIR)*, 2006.

[23] E. M. Vorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[24] K. Wang, T. Walker, and Z. Zheng. Pskip: Estimating relevance ranking quality from web search clickthrough data. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[25] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *ACM Conference on Information Retrieval (SIGIR)*, 2007.

[26] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metric for subtopic retrieval. In *ACM Conference on Information Retrieval (SIGIR)*, 2003.