Revisiting Online Personal Search Metrics with the User in Mind

Azin Ashkan and Donald Metzler

Google Inc.
{azin,metzler}@google.com

ABSTRACT

Traditional online quality metrics are based on search and browsing signals, such as position and time of the click. Such metrics typically model all users' behavior in exactly the same manner. Modeling individuals' behavior in Web search may be challenging as the user's historical behavior may not always be available (e.g., if the user is not signed into a given service). However, in personal search, individual users issue queries over their personal corpus (e.g. emails, files, etc.) while they are logged into the service. This brings an opportunity to calibrate online quality metrics with respect to an individual's search habits. With this goal in mind, the current paper focuses on a user-centric evaluation framework for personal search by taking into account variability of search and browsing behavior across individuals. The main idea is to calibrate each interaction of a user with respect to their historical behavior and search habits. To formalize this, a characterization of online metrics is proposed according to the relevance signal of interest and how the signal contributes to the computation of the gain in a metric. The proposed framework introduces a variant of online metrics called *pMetrics* (short for *personalized metrics*) that are based on the average search habits of users for the relevance signal of interest. Through extensive online experiments on a large population of GMail search users, we show that *pMetrics* are effective in terms of their sensitivity, robustness, and stability compared to their standard variants as well as baselines with different normalization factors.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Retrieval effectiveness**; *Personalization.*

ACM Reference Format:
Azin Ashkan and Donald Metzler. 2019. Revisiting Online Personal Search Metrics with the User in Mind. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331266

1 INTRODUCTION

Online evaluation in information retrieval involves deploying a search engine to actual users with real-world information needs and assessing the performance of the system based on how these users interact with it [12]. Online evaluation is often used for controlled experiments and allows absolute or relative quality assessments with respect to a group of online quality metrics. Traditional online quality metrics are based on search and browsing signals, such as position and time of the click, modeling all users in exactly the same manner. As a result, quality metrics computed based on online evaluation may inaccurately reward or penalize a user's interaction with the ranked results. In other words, online evaluation tends to oversimplify real-world search tasks in such a way that the final metric value may overestimate or underestimate user satisfaction [21].

In addition, previous efforts [14, 15] show that satisfaction can be best explained with respect to the *outcome* (a.k.a *gain*) obtained through the search experience at the price of the *effort* spent by the searcher. While online metrics take into account the first factor via the implicit feedback signals (such as position and time of click), they do not account for difference in individuals' effort. In particular, these metrics do not take into account individuals' search habits and persistence. Different users tend to treat ranked results differently [1, 13, 23]; one may be a patient searcher in the sense that they take the time to browse all the way down on the result list as opposed to another user who usually cares about the very top ranked results and rarely clicks on the ones lower on the list. While traditional online effectiveness metrics assume all users have similar behavior, this paper proposes a framework to adopt these metrics to take into account variability of search and browsing behavior across individuals.

Modeling individuals' behavior in Web search may be challenging as the user's historical behavior may not be available (e.g., if the user is not signed into a given service). However, in personal corpora, a.k.a personal search [4, 29], individual users issue queries over their personal corpus (e.g. email, file, etc) while they are logged in the system. Given the availability of such a valuable source of information and the popularity of personal search, the existing evaluation metrics can be adopted to consider a user's habits and effort for finding information in their personal corpus. Even in the cold start case, where there is little or no historical information recorded for an individual, one can take into account the overall behavior of an average user in the system to emulate that individual's behavior resulting in a more accurate estimate of user satisfaction.

This paper proposes and validates a user-centric evaluation framework for personal search as the first yet significant step towards that goal. The main idea is to calibrate each user interaction with respect to the user's historical behavior and search habits. To formalize this, a characterization of online metrics is proposed according to the signal of interest and how the signal contributes to the computation of the gain in a metric. Then the framework introduces a variant of normalizing the online metrics based on the average search behavior of users for the signal of interest and its

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331266

contribution property. This is supported by the intuition that similar interaction signals obtained from different individuals should contribute differently to the overall metric value depending on the effort each individual spends comparing to their usual behavior in the system. As a result, variants of online metrics are formulated and validated that take into account the above factors, resulting in a suite of user-centric online metrics that we refer to as *pMetrics* (short for *personalized metrics*).

Our primary contributions can be summarized as follows:

- We propose a characterization of online metrics according to the signal of interest and how the signal contributes to the computation of the gain in a metric.
- We propose a user-centric evaluation framework for personal search, through which we introduce our notion of personalized online metrics (*pMetrics*) based on the average search habit of users for the signal of interest and its contribution property.
- Through extensive online experiments on a large population of GMail search users, we show that *pMetrics* are effective in terms of their sensitivity, robustness, and stability compared to their standard variants as well as baselines with different normalization factors.

The remainder of the paper is organized as follows: Section 2 provides an overview of the related work. A characterization of online metrics is proposed in Section 3. This serves as the basis of the proposed framework for user-oriented evaluation of personal search in Section 4. The results of the evaluation of metrics adopted with respect to the proposed framework are presented in Section 5. Section 6 summarizes the concluding remarks and future direction.

2 RELATED WORK

Offline and online quality metrics are widely used to measure the performance of search and retrieval systems. Offline metrics rely on relevance judgments to evaluate the effectiveness of a ranking algorithm based on the explicit feedback obtained from annotators on how satisfied they are with ranked results. Online metrics take a contrasting approach in a sense that they are based on the actual interactions between users and search systems in a natural usage environment. While offline evaluation is usually costly and cannot be done at large scale, it is often fast and less costly to collect online experiments data in modern systems, making it fairly easy to scale up online evaluation. On the other hand, online metrics are known to suffer from various form of biases, such as position and selection biases [16, 33].

Although both types of evaluation metrics have achieved success, there still remain questions about whether they can predict "actual" user satisfaction [5]. This may be more of an issue for online metrics as the online behavior of users can be affected by their search biases and habits that may need to be corrected for when inferring search success. Another reason that online metrics are the target of our study in this work is due to the popularity of online evaluation in personal search domains. Personal search is an important information retrieval task with applications such as email search [4, 10] and desktop search [8]. One major difference between personal and Web search is that in personal search scenarios each user has access only to their own private document corpus (e.g., emails, files, etc) while they are logged in during their entire interaction with the service. An important challenge in the context of personal search is the collection of explicit relevance judgments. Collection of TREC-like document relevance judgments by third party raters are difficult to obtain due to privacy restrictions [7]. In addition, since each user will have their own unique set of information needs and documents that evolve over time (e.g., new emails arrive every day), explicit relevance judgments may be prohibitively costly to maintain. Therefore, performing online evaluation through controlled experiments by utilizing click-through data as a noisy and biased source of relevance feedback has become essential for building highly effective personal search systems.

Search satisfaction is commonly measured using simple instruments in both offline and online metrics [5, 14, 15]: asking searchers or annotators whether they are satisfied on a binary or multi-point scale, or considering user interaction signals (e.g., click) as the notion of satisfaction. Sanderson et al. [22] study the correlation of user preferences and evaluation measures, arguing that there is much scope for refining effectiveness measures to better capture user satisfaction and preferences. While previous studies [2, 30] have explored how user behavior can contribute to gains in relevance through search personalization, the current paper is the first to study how historical user behavior can provide valuable signals for delivering a better evaluation of personal search systems.

Mao et al. [17] study the relationship between relevance, usefulness, and satisfaction and argue that traditional system-centric evaluation metrics may not be well aligned with user satisfaction. They suggest that a usefulness-based evaluation method should be defined to better reflect the quality of search systems perceived by users. In our proposed evaluation framework, we calibrate the notion of relevance from each interaction of a user with respect to the user's historical behavior and search habits that can be seen as a user-dependent notion of the usefulness of this interaction towards the computation of the overall gain from all interactions of this user with the system.

Other efforts [14, 15] show that satisfaction can be best explained with respect to the *outcome* (a.k.a *gain*) obtained through the search experience at the price of the *effort* spent by the searcher. They argue that the nature of a search task and the tenacity of the searcher are among the factors that can influence the types of search behavior observed, and therefore further studies are required to understand the role of these and other factors on behavior and search satisfaction. While traditional online metrics take into account the first factor (i.e., the gain) via the implicit feedback signals (such as position and time of click), they do not account for difference in individuals' efforts. In particular, these metrics do not take into account individuals' search habits and persistence.

One piece of related work on evaluation measures that does capture searchers' efforts is the one by Smucker and Clarke [25]. That work introduces an offline measure called time-biased gain (*TBG*), which takes into account the effort that a user spends to examine result snippets before reading the actual result document. They propose to use the time spent by the user to examine the snippets (that can depend on factors such as the document length) as the basis for discounting the relevance value of a document instead of the document rank. In more recent work [24], Smucker and Clarke extend *TBG* in the context of stochastic simulation

of user behaviors in order to capture variance in user behavior such that they create a single user model for each participant in their study. Furthermore, the U-measure proposed by Sakai and Dou [21], is another offline metric that considers the searcher's effort in measuring search satisfaction. Instead of discounting the value of a retrieved piece of information based on ranks, U-measure discounts it based on its position within the trailtext, where trailtext represents all the text the user has read during an information seeking process. Similar to *TBG*, U-measure also takes the document length into account.

Our work is also based on the idea of looking beyond item ids and rank to capture searchers' efforts. However, we consider discounting the gain of each interaction with respect to a users' average behavior and search habits specifically for online metrics. To this end, we propose a generic framework to calibrate online metrics with respect to the implicit interaction signals, such as position and time of click. To the best of our knowledge, this is the first work that focuses on capturing variability of user behavior in online metrics for personal search evaluation.

3 CHARACTERISTICS OF STANDARD ONLINE METRICS

Online quality metrics rely on user interaction signals, such as position and time of click, as opposed to offline metrics that are based on explicit relevance judgments. While interaction signals may introduce noise and bias in the evaluation, they are often cheap and fast to collect at large scale, which makes it relatively easy to scale up online metrics and monitor them frequently as opposed to their offline counterparts. In addition, online metrics rely on the actual experience of users who issued queries in the search system giving straightforward descriptions on how users would interact with the system.

As pointed out by Chen et al. [5], online metrics typically are i) click-based metrics; e.g., click-through rate and average click position, and ii) time-based metrics; e.g., query dwell time and average time to the first click. These metrics are typically computed as the average of gains across actions. Denoting the number of actions by n, an online metric M can be formulated as:

$$M = \frac{1}{n}\Sigma_{i=1}^{n} M_i$$

where M_i is the gain obtained from action i.

Examining the formulation of online metrics, we can characterize these metrics according to the signal of interest and how the signal contributes to the computation of gain in the metric:

- **Interaction Signal**: Online metrics vary based on the interaction signal that is obtained from search logs and used towards their computation. This signal may be *position* of the interaction which is usually a click (e.g., the click position in average click position metric), an *indicator* value (e.g., the binary click values used in the click-through rate metric), and *time* of the interaction (e.g. the time in query dwell time).
- **Contribution to Gain**: The way an interaction signal contributes to the computation of the gain may vary in online metrics. We find that for some metrics the signal of interest contributes linearly (e.g., click position for average click position) towards the gain, while there is an inverse contribution (e.g., position for mean reciprocal rank) for others.

A group of common online metrics along with their signal and contribution properties based on this characterization are depicted in the first four columns of Table 1. As can be seen in the second column of the Table, the right hand side of Σ is equally weighted across all actions (and therefore across all users) for all the listed metrics. In other words, the only factor that determines the gain is the value of the interaction signal, such that similar interactions of different users would be treated equally towards the final value of the metric.

While these interaction signals are valuable sources reflecting user behavior, they may introduce noise and bias in online evaluation. This paper is one step towards better adjustment of these signals such that they capture user's effort with respect to their current interaction as well as their usual search habits.

As an example, consider the position-based click metric *MRR* (mean reciprocal rank – the first metric listed in Table 1) in which the position of click contributes in the same fashion across all users (i.e., inversely) to calculate the gain obtained from each action. However, a user may be a patient searcher in the sense that they take the time to browse all the way down on the result list as opposed to another user who usually cares about the very top results (first or second) and rarely clicks on the ones lower on the list. Such a metric does not accurately reflect the fact that different members of the user population tend to behave differently than others. What if this metric is adopted to reflect the average click position of individuals as well?

As another example, consider the time-based *TTC* metric (time to click – the last metric in Table 1) that uses the time between the start of the search session and first click in that session as the signal of interest contributing linearly to compute the gain. The standard formulation of the metric measures the average time to first click across users by assuming they are all equally fast. However, an individual user may be a thorough examiner who takes a longer time on average to decide whether to click on a result. The question is whether this metric can be formulated according to the usual examination habit of the user, which can be captured from the click history of this individual.

4 A USER CENTRIC EVALUATION FRAMEWORK

In this section, we extend the online quality metrics with respect to the two factors described for these metrics earlier: interaction signal and contribution to the gain. The average search behavior of users for the signal of interest and how the signal contributes to the metric's gain form the basis of the proposed normalization for these metrics. This is supported by the intuition that similar interaction signals obtained from different individuals should contribute differently to the overall metric value depending on the effort each individual spends compared to their usual behavior in the system.

4.1 Proposed Formulation of Online Metrics

Given action i with the signal s_i (e.g., click position, dwell time, etc), we denote the average signal value of the corresponding user by $\bar{s}_i$.

Table 1: Different types of online metrics and their personalized variant based on the proposed framework.

Metric family	Standard metric	Signal s_i	Contribution	w_i	pMetric
mean reciprocal rank	$MRR = \frac{1}{n}\Sigma_{i=1}^{n}\frac{1}{r_i}$	click position r_i	inverse	$\log(\frac{\bar{r}_i}{r_i}+1)$	$pMRR = \frac{\Sigma_{i=1}^{n}\frac{\log(\frac{\bar{r}_i}{r_i}+1)}{r_i}}{\Sigma_{i=1}^{n}\log(\frac{\bar{r}_i}{r_i}+1)}$
average click position	$ACP = \frac{1}{n}\Sigma_{i=1}^{n} r_i$	click position r_i	linear	$\log(\frac{r_i}{\bar{r}_i}+1)$	$pACP = \frac{\Sigma_{i=1}^{n}\log(\frac{r_i}{\bar{r}_i}+1)\times r_i}{\Sigma_{i=1}^{n}\log(\frac{r_i}{\bar{r}_i}+1)}$
click-through rate	$CTR = \frac{1}{n}\Sigma_{i=1}^{n} c_i$	binary click indicator c_i	linear	$\log(\frac{c_i}{\overline{CTR}_i}+1)$	$pCTR = \frac{\Sigma_{i=1}^{n}\log(\frac{c_i}{\overline{CTR}_i}+1)\times c_i}{\Sigma_{i=1}^{n}\log(\frac{c_i}{\overline{CTR}_i}+1)}$
abandonment rate	$AR = \frac{1}{n}\Sigma_{i=1}^{n} a_i$	binary abandonment indicator a_i	linear	$\log(\frac{a_i}{\overline{AR}_i}+1)$	$pAR = \frac{\Sigma_{i=1}^{n}\log(\frac{a_i}{\overline{AR}_i}+1)\times a_i}{\Sigma_{i=1}^{n}\log(\frac{a_i}{\overline{AR}_i}+1)}$
time to click	$TTC = \frac{1}{n}\Sigma_{i=1}^{n} t_i$	time t_i	linear	$\log(\frac{t_i}{\bar{t}_i}+1)$	$pTTC = \frac{\Sigma_{i=1}^{n}\log(\frac{t_i}{\bar{t}_i}+1)\times t_i}{\Sigma_{i=1}^{n}\log(\frac{t_i}{\bar{t}_i}+1)}$

We consider a normalization factor of w_i for each action of the user that depends on $\bar{s}_i$, the current signal s_i, and a metric dependent parameter ρ:

$$w_i = f(\bar{s}_i, s_i, \rho)$$

This normalization factor can be seen as a measure of how well the ranking system fulfills the expectation and general behavior of users. For instance, if the system delivers beyond the user's average behavior, the metric should reflect that as a positive impact towards the overall score of the system. As a result, the proposed online metric (pM), normalized with respect to the user's historical behavior, can be formulated as:

$$pM = \frac{\Sigma_{i=1}^{n} w_i M_i}{\Sigma_{i=1}^{n} w_i} \tag{1}$$

which is the weighted average of the standard gain with respect to the factor w_i across the actions.

We refer to the proposed user-centric variant of online metrics as *pMetrics* (short for *personalized metrics*), and denote it by pM for the given standard metric M. For instance, the personalized variant of *MRR* is denoted by *pMRR*, the personalized variant of *CTR* is denoted by *pCTR*, and so on.

4.2 Normalization Factor

Here we consider a log-based function f to account for the impact of a user's past behavior through the normalization factor w_i. Investigating variants of f in a more systematic way is a future direction for this work, but we compare the log-based function with a linear-based candidate function in the experimental evaluation.

Assuming a log-based weighting function, the way that the average behavior of the user is formulated with respect to their current behavior can be decided based on the type of the contribution the signal has to the gain of the metric. If the signal of interest contributes linearly towards the gain, the higher the observed signal value is for an action the higher the value of the gain is for the metric in that action. In this case, we want to reward the system slightly higher if the current signal (s_i) is also higher than the user's usual behavior ($\bar{s}_i$). In other words, in case of a "linear" contribution to the gain, we want to consider a normalization weight proportional to $\log(\frac{s_i}{\bar{s}_i})$. Otherwise, if the contribution to the gain is "inverse", w_i should be proportional to $\log(\frac{\bar{s}_i}{s_i})$ in order to take into account that the system should be rewarded more if the current signal value is lower than user's historical behavior recorded for that signal.

To implement this idea, we define a binary parameter, denoted by ρ that is metric dependent. This parameter is 1 if the signal's contribution towards the metric's gain is linear, and 0 otherwise. As a result, the proposed normalization factor is formulated as follows:

$$w_i = \log[(\frac{s_i}{\bar{s}_i})^{2\rho-1} + 1] \tag{2}$$

The intuition behind this weighting function can be explained further with an example. Consider a position based metric, such as *MRR*, where the signal of interest is the position of click and it contributes inversely to the metric's gain. This means that $\rho = 0$ in Equation 2. We also consider $\bar{s}_i = \bar{r}_i$ and $s_i = r_i$, respectively, representing the average position of click and the current position of click for the user corresponding to action i. As a result, the normalization factor for the personalized variant of *MRR* is set as:

$$w_i = \log[(\frac{r_i}{\bar{r}_i})^{(2\times 0-1)} + 1] = \log(\frac{\bar{r}_i}{r_i} + 1)$$

and the underlying intuition can be explained as follows:

(1) If the current click position and the average click position are the same, we do not want to change the gain obtained from this action, because the system appears to be able to fulfill as expected according to the general behavior of the user; hence,

$$r_i = \bar{r}_i \Rightarrow w_i = 1.$$

(2) If the user usually clicks on lower positions (towards the bottom of the list) than their click position in the current action (i.e., $\bar{r}_i > r_i$), we want to reward the performance of the system in this action, because it appears to provide the user with an experience with less effort needed than usual;

$$\bar{r}_i > ri \Rightarrow \log(\frac{\bar{r}_i}{r_i} + 1) > 1 \Rightarrow w_i > 1$$

(3) If the user usually clicks on higher positions than their click position in the current action (i.e., $\bar{r}_i < r_i$), we want to reduce the gain of this action's click on the overall performance of

Table 2: An example of how variants of *MRR* discussed in this paper can be computed with respect to different users' search habits and the way their actions contribute towards the gain for these metrics.

	Action 1	Action 2	Action 3	Action 4	Action 5
Current click position (r_i)	2	3	1	3	-
Average click position for the user of this action ($\bar{r_i}$)	2	1	2	3	2
Log-based normalization: $w_i = log(\frac{\bar{r_i}}{r_i} + 1)$	1	0.41	1.58	1	1
Linear-based normalization: $w_i = \frac{\bar{r_i}}{r_i}$	1	0.33	2	1	1
$MRR_i = \frac{1}{r_i}$ (standard gain)	0.5	0.33	1	0.33	0
$pMRR_i = w_i MRR_i$ (log-based gain)	0.5	0.13	1.58	0.33	0
$pMRR_i = w_i MRR_i$ (linear-based gain)	0.5	0.11	2	0.33	0
$MRR = \frac{1}{5}(0.5 + 0.33 + 1 + 0.33 + 0) = 0.43$ $pMRR$ (log) $= [1/(1 + 0.41 + 1.58 + 1 + 1)] \times (0.5 + 0.13 + 1.58 + 0.33 + 0) = 0.50$ $pMRR$ (linear) $= [1/(1 + 0.33 + 2 + 1 + 1)] \times (0.5 + 0.11 + 2 + 0.33 + 0) = 0.55$					

the system to account for the higher than expected effort they spent;

$$\bar{r}_i < ri \Rightarrow \log(\frac{\bar{r}_i}{r_i} + 1) < 1 \Rightarrow w_i < 1$$

The above intuition can be better illustrated through an example in Table 2. Assume there are five actions recorded in the search log based on which we want to calculate *MRR*. The second row in the Table shows the click position for these actions. Note that the fifth action is a no-click, and its gain is set as zero in the *MRR* formulation. The third row of the Table, on the other hand, shows the average click position of the user corresponding to each action. This average click position can be calculated from the historical interactions of users with the system. If there is no historical action recorded for a user (i.e., the cold start case), the average click position of all users is used. It is noted that the numbers in this row are for example purposes, and in reality they do not need to be integers.

The following two rows in Table 2 (i.e., the fourth and fifth rows) respectively represent the log- and linear- based normalization factors computed in terms of the current click signal and the average click signal for each action. The next three rows show the gain value that each action contributes towards the metric for each of the standard *MRR*, the log-based *pMRR*, and the linear-based *pMRR*. For instance, the first action contributes the same gain of 0.5 towards all three metrics because the current click behavior and the historical click behavior of the corresponding user are the same (i.e., $r_i = \bar{r_i} = 2$). Whereas, for the second action, both the log-based and linear-based variants of *MRR* contribute less compared to the standard variant of the metric. The only difference is that the log-based variant has a more conservative normalization than the linear-based, and as a result we see gain contributions of 0.13 versus 0.11 for these settings respectively, as opposed to the standard metric which has a gain of 0.33 from the second action. Finally, the value of each metric is computed according to Equation 1 as shown in the last row of Table 2.

In the above example, we dived deep into *MRR*, where the metric's signal of interest is the click position that contributes inversely to the metric's gain. Similar arguments hold for cases with click position as the signal and linear contribution (e.g., *ACP* metric - average click position) and time-based metrics (e.g. *TTC* metric - time to click).

However, for metrics with binary signal (e.g. CTR with click or no click as the only two possible values of the metric's signal), we need to measure the average behavior of the user ($\bar{s}_i$ in Equation 2) with respect to the historical value of the metric for each individual. The rest of the normalization factor calculation is similar to what explained above. For instance, for the *CTR* metric, we set $\rho = 1$ since the binary click signal contributes linearly to the *CTR*'s gain. We then consider $\bar{s}_i = \overline{CTR_i}$ and $s_i = c_i$, respectively, representing the average click-through rate and the current click signal for the user corresponding to action i. As a result, the normalization factor for the personalized variant of *CTR* is set as:

$$w_i = \log[(\frac{c_i}{\overline{CTR_i}})^{(2\times 1-1)} + 1] = \log(\frac{c_i}{\overline{CTR_i}} + 1)$$

As stated before, for cold start cases where there is little or no historical information recorded for an individual, one can take into account the overall behavior of an average user of the system to emulate the individual's behavior. Hence, we define $\bar{s}_0$ to denote the average value of the signal of the interest across all users, and let $\bar{s}_i = \bar{s}_0$ in Equation 2 for cases where the user of action i has a value of zero recorded for their average historical behavior at the time of evaluation.

Following the proposed formulation in Equation 1 and the normalization variants described above, the personalized variant of commonly used online metrics are presented in the last column of Table 1. Different forms of metrics are listed in the Table that are based on the factors described in our characterization of online metrics in Section 3.

It is worth mentioning that although *ACP* and *CTR* share respectively similar characteristics with *TTC* and *AR* - in terms of how their signal contributes to the gain and therefore how their *pMetric* variants are formulated - we include them all in the Table for the completeness of our presented list of online metrics that are commonly used in the personal search evaluation.

5 EVALUATION OF THE PROPOSED METRICS

A suite of experiments are conducted in this section to empirically evaluate the effectiveness of the proposed metrics and compare them with their standard counterparts and baselines with different normalization factors.

5.1 Experimental Setting

The data set used for our experimental study consists of GMail search log data. GMail is one of the most popular and widely used email providers. It has over 1.5 billion users[1], many of whom rely on search to find their personal emails.

For search, GMail uses an overlay to show relevance ranked search results as a user types. This evaluation specifically focuses on the relevance ranked email results displayed in the overlay. The search overlay disappears when the user clicks on one of the results or when they press the "enter" key (thus triggering a chronologically sorted search). The search overlay shows up to six relevance ranked email results for each query. Given the specifics of the user interface, each search query is associated with at most one click.

Given the sensitive nature of personal email content, all of the data analyzed in these experiments have been rigorously aggregated and anonymized in line with industry-wide best practices and in accordance with all relevant terms of service and business contracts.

For all the experiments reported in this section, we use a window of 30 days from different points of time in 2018 and consider a notion of d, such that the last d days constitute the *observation period* based on which $pMetrics$ are computed. The preceding $30 - d$ days are the *estimation period* used to compute each user's average behavior. In these experiments, unless stated otherwise, we set $d = 10$. This value has been chosen so that we can dedicate the majority of the time from the beginning of the study window to compute the average behavior of users while there is still enough time left as our observation period to obtain sufficient number of actions for the computation of the metrics. The value of d is varied in the second group of experiments (Section 5.3) where we study the discriminative power of the proposed metrics.

We focus our experiments on several tens of millions of users randomly sampled from the population of users of the service that were active during the experimentation time period. The interaction signals recorded for our experiments are the following:

- Position of click (r_i): varies from 1 to 6 based on whether the user clicked on any of the results displayed in the action i.
- Indicator of click (c_i): 0 or 1 indicating whether the user clicked on any of the displayed results in the action i.
- Time of click (t_i): the time from when the query is issued until the click (if any) occurs in the action i.

For each interaction signal in a metric, we also compute the average behavior across *all users* in the estimation period. This global estimate is substituted for an individual user's average behavior if there is no interaction history in the observation period. This global average is denoted by $\bar{s}_0$ in general form in Section 4, and by $\bar{r}_0$, $\bar{c}_0$, and $\bar{t}_0$ for the interaction signals collected in our study.

Various online experiments are conducted with respect to the above settings. Each $pMetric$, its standard form, and the variant with a linear weight function are calculated for each of these settings for comparison purposes. Each group of experiments is explained in more detail next. In summary, we aim at addressing the following research questions through these experiments:

- How **sensitive** are $pMetrics$ in detecting changes in randomized tests?
- How **discriminative** are $pMetrics$ across runs?
- How **stable** are $pMetrics$?
- How **correlated** are $pMetrics$ with their standard counterparts?

5.2 Sensitivity in Randomized Tests

A good online metric should be able to detect changes through A/B experiments effectively. Hence, a reasonable approach to validate the effectiveness of the proposed metrics is to purposefully degrade the quality of the underlying search engine's results through randomization to observe whether the proposed metrics are at least as good as their standard counterparts in detecting the change.

For this purpose, we randomly divide our user population into seven experiment buckets. All users within a given bucket experience the search results returned by exactly one of the following seven settings:

- E_C: The existing (production) ranking algorithm is used by the system.
- E_R: Fully randomizes the ordering of the top six items returned by the production ranking algorithm.
- $E_{i,i+1}$ for $i \in \{1, ..., 5\}$: Randomly swaps the production system's results in positions i and $i + 1$, known as *FairPairs* [18].

We end up with six A/B experiments such that they vary based on their treatment group (B), which is one of E_R, $E_{1,2}$, ..., $E_{5,6}$. All experiments share the same control group (A) that is E_C. Ideally, in all these A/B experiments, a perfect metric should be able to detect a performance decrease in control versus treatment.

Different variants of popular online metrics are computed for this set of A/B experiments. For each metric, the point estimate of the relative change in treatment versus control is computed at a significance level of 0.05 using a Jackknife statistical test [9] with 20 buckets. These numbers are reported in percentage form in Table 3. If a metric detects the change to be statistically significant, the corresponding number appears in bold in the Table.

Of the standard online evaluation metrics, CTR appears to be the most consistent in terms of detecting system pairs to be statistically significantly different. The same trend is also observed in the proposed $pMetric$ variant of CTR in Table 3.

Note that some of the system changes experimented with, such as those FairPairs that randomly swap items at lower positions in the ranked list, may not be easily detectable via A/B tests. This is primarily due to the relatively sparse number of clicks observed at these positions as a result of the underlying click distribution and due to factors such as position bias. For these reasons, it is not too surprising to see that none of the metrics, standard or personalized, are able to reliably detect such subtle changes in behavior.

It is also observed that for all cases, the proposed log-based personalized variant of the metrics can detect statistically significant changes in system pairs just as well as the standard variant of each metric. The baselines with the linear-based weight function, on

[1] https://twitter.com/gmail/status/1055806807174725633

Table 3: Sensitivity in randomized tests: The control setup is the same in all tests whereas the treatments are based on different settings that change the results list at random. The numbers do not represent the absolute value of metrics, but they reflect the percentage change of each metric in treatment from control.

Metric family	Comparison pair A/B	point estimate (%) Standard	Proposed log-based *pMetric*	Linear-based variant
MRR	E_C/E_R	**-26.5901**	**-31.7119**	**-32.9302**
	$E_C/E_{1,2}$	**-8.6315**	**-11.5692**	**-12.5413**
	$E_C/E_{2,3}$	-0.8017	-0.9011	-0.897
	$E_C/E_{3,4}$	-1.0374	-0.9011	-1.271
	$E_C/E_{4,5}$	-0.7669	-0.9011	-0.6772
	$E_C/E_{5,6}$	-1.4646	-1.6603	-1.6678
CTR	E_C/E_R	**-11.069**	**-13.2651**	**-9.4612**
	$E_C/E_{1,2}$	**-3.008**	**-4.5422**	**-2.1373**
	$E_C/E_{2,3}$	-0.799	-1.5307	-0.4311
	$E_C/E_{3,4}$	-1.0499	-3.1233	-0.5184
	$E_C/E_{4,5}$	-1.0251	-1.6247	-1.0822
	$E_C/E_{5,6}$	**-1.7228**	**-4.3949**	-0.5155
TTC	E_C/E_R	**+1.8339**	**+1.9263**	+6.2718
	$E_C/E_{1,2}$	**+0.5889**	**+0.9216**	+1.8828
	$E_C/E_{2,3}$	+0.1559	+0.2265	+2.0705
	$E_C/E_{3,4}$	+0.0352	+0.161	-0.7703
	$E_C/E_{4,5}$	+0.2646	+7.2407	+1.3783
	$E_C/E_{5,6}$	+0.3554	+0.5322	-0.3083

the other hand, appear weaker in detecting some of the changes. In particular, the personalized variant of *TTC* with linear-based weighting is found to detect no statistically significant change in pairs, whereas the proposed log-based (*pMetric*) variant of the metric, as well as the standard one, detect significant changes in some cases. This may be due to the fact that a linear-based weighting may not be suitable for a signal like time that tends to have large absolute values and high variance. A log-based weighting would provide a more tempered way of taking the average behavior of individuals into account, suggesting that the proposed framework suits a wide group of interaction signals in online metrics.

5.3 Discriminative Power

We also compare the proposed metrics with the baselines in terms of their discriminative power [19]. The discriminative power is commonly used for evaluating the robustness of quality metrics. Given a collection of runs and a quality metric, the discriminative power of the metric is calculated as the percentage of the run pairs that are detected as statically significant based on this metric and using a pairwise significance test. A low value of the discriminative power for a metric indicates that the metric may not be useful for drawing conclusion from experiments, whereas a metric with relatively higher discriminative power can be regarded as being more sensitive in detecting changes.

The experiments conducted in this section aim at calculating the discriminative power of *pMetrics* as well as the baselines to evaluate how consistent they are across runs; in other words, how often the metrics can detect differences between runs with high confidence. The online experiments that we set up for this purpose are the following:

- Experiments that display differing numbers of relevance ranked email results. Values tried include 3, 4, 5, 6, 7, and 8.
- Experiment in which the ranking of the top results returned by the production ranking algorithm is randomized, similar to the fully randomized experiments described in the previous subsection.

For each metric and every pair of experiments in the collection, we conduct a pairwise significance test to measure the statistical significance of the difference in the mean of the experiments pair. The number of statistically significantly different pairs for each metric determines the metric's discriminative power.

Since we have 7 experiments (6 displaying different numbers of results and one that randomizes the displayed results order), we have $\binom{7}{2} = 21$ experiment pairs to test. We use a Jackknife statistical test [9] with 20 buckets and $\alpha = 0.05$. The pairwise comparisons are repeated across five days, varying d from 8 to 12 days to collect enough data to compute the average discriminative power of each metric. The analysis is conducted for *MRR*, *CTR*, and *ACP* and for three settings: standard, *pMetric*, and linear weighting, resulting in 3x3=9 metrics overall. As a result, we conducted 945 statistical tests for this round of the study: 21 (pairs) × 9 (metrics) × 5 (days) = 945 statistical tests conducted.

The results of this study are depicted in Figure 1. Comparing standard metrics only (blue bars in the plot), it appears that *MRR*-based metrics are more discriminative than *CTR*- and *ACP*-based metrics in general. Comparing different variants of *pMetrics* (red

Figure 1: Discriminative power of metrics under the pairwise significance test with a significance level of 0.05.

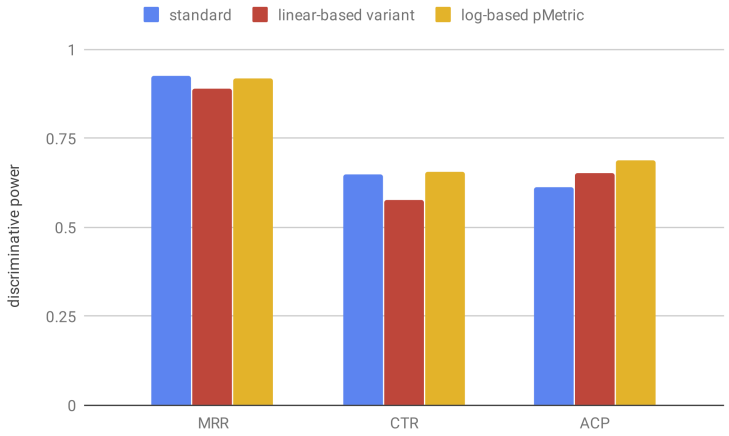

Table 4: Percentage of A/A tests with no statically significant difference detected by *pMetrics*, their linear-based variant, and the standard metrics.

Metric family	Setting	% of A/A tests with no significant change
MRR	standard	94.77
	linear-based variant	90.20
	proposed log-based *pMetric*	95.43
CTR	standard	94.20
	linear-based variant	93.31
	proposed log-based *pMetric*	96.73
ACP	standard	94.45
	linear-based variant	90.86
	proposed log-based *pMetric*	95.84

bars versus orange bars), the proposed log-based *pMetrics* appear to outperform their linear-based variants in terms of the discriminative power across all metric types, suggesting once again that a log-based normalization is indeed better suited for these metrics. Finally, comparing standard metrics against the proposed log-based *pMetric* (blue bars against orange ones), we observe that the log-based *pMetrics* are more discriminative than or at least as good as their standard counterparts in detecting a change throughout the experiments, suggesting these metrics are reliable online metrics for controlled experiments.

5.4 Stability in A/A Tests

An online quality metric is ideally expected to detect no significant difference between two groups of the same population experiencing the same system. However, the natural variability in the population usually results in some false positives, or type I errors, indicating that a behavior change exists when in reality there should not be a difference between the two control groups. This motivates us to validate the proposed metrics in terms of their stability in A/A tests and study how they can handle user sampling bias in such tests compared to standard online metrics.

It is expected that the proposed metrics are, at the least, as stable as standard metrics in A/A tests. We evaluate this via setting up a set of A/A experiments by randomly placing our user population into 50 control groups. We then conduct a Jackknife statistical test (with 20 buckets) between every control pair, resulting in $\binom{50}{2} = 1225$ tests per metric. For each metric, we calculate the percentage of A/A tests in which the metric detects no statistically significant change at a significance level of 0.05.

Table 4 shows the results of this study. As expected, the proposed *pMetrics* outperform their counterparts in terms of their stability in A/A tests. The results observed here are similar to those observed in other experiments. Specifically, the proposed log-based variants clearly outperform the linear-based variants. Furthermore, the log-based variants consistently outperform the standard metrics for each metric family. These results suggest that the proposed framework can be seen as a step closer to reduce the effect of user bias in evaluating online controlled experiments.

5.5 Correlation Study

Correlation studies are commonly performed to evaluate evaluation metrics empirically. This is based on the idea that metrics that

Table 5: Kendall's τ correlation of *pMetrics* as well as the linear-based variant of *pMetrics* with the standard metrics.

Metric family	*pMetric* variant	Correlation with standard metric
MRR	linear-based variant	0.7222
	proposed log-based *pMetric*	0.7778
CTR	linear-based variant	0.3889
	proposed log-based *pMetric*	0.4444
TTC	linear-based variant	0.0555
	proposed log-based *pMetric*	0.2222

correlate poorly with standard metrics are probably poor metrics [3]. What is commonly performed in this type of empirical evaluation is that quality metrics are evaluated based on how similarly they rank systems compared to the established metrics in order to verify that the new metrics measure the same thing on average. Kendall's τ [26] is a well-established rank correlation measure for comparing system rankings generated by different metrics [6, 20, 28, 32].

In this set of experiments, we randomly divide our user population into nine buckets and expose all users of each bucket to experience the search results returned by a different ranking model. The ranking models include a number of experimental models that were being evaluated at the time.

Kendall's τ is used to measure the stability of the rankings of these experimental runs under the *pMetrics* as well as their linear-based variants with respect to the rankings obtained from the standard metrics. Note that Kendall's τ ranges from +1 to -1, with +1 indicating perfect agreement and -1 indicating the opposite.

The results of this analysis are presented in Table 5. As it is shown in the Table, the proposed log-based variant of *pMetrics* correlates better with standard metrics as opposed to the linear-based variant. This further confirms the previous empirical results presented in this section suggesting that *pMetrics* offer a better formulation with respect to the user's average behavior comparing to their linear-based counterparts.

6 CONCLUSION

Traditional online quality metrics are based on the search and browsing signals, such as position and time of the click, modeling all users in exactly the same manner. As a result, quality metrics computed based on the online evaluation may inaccurately reward or penalize a user's interaction with the ranked results. While these metrics take into account a gain-based computation of the final score with respect to the implicit feedback signal from each interaction of each user, they do not account for differences in individuals' effort.

Given the availability of logged-in information in personal search, the existing evaluation metrics can be adopted to consider a user's habits and effort for finding information in their personal corpus. Hence, in this paper, we propose a user-centric framework to calibrate *online personal* search metrics to take into account variability of search and browsing behavior across individuals.

We first propose a characterization of online metrics according to the signal of interest and how the signal contributes to the computation of the gain in a metric. Then we motivate and introduce a log-based weighting function and use it to normalize online metrics in our framework based on the average search behavior of users for the signal of interest and its contribution property. This is supported by the intuition that similar interaction signals obtained from different individuals should contribute differently to the overall metric value depending on the effort each individual spends comparing to their usual behavior in the system. As a result, we formulate and validate variants of online metrics that take into account the above factors, and refer to them as *pMetrics* (personalized metrics).

Through extensive online experiments on a large population of GMail search users, we evaluate the effectiveness of *pMetrics* and compare them with their standard counterparts and baselines with linear-based normalization factors. Through randomized experiments that purposefully degrade the quality of the search results, we show that the proposed metrics are as sensitive as their standard variants in terms of detecting statistically significant changes while they outperform their linear-based counterparts. We also compare *pMetrics* with the baselines in terms of their discriminative power across a collection of runs. The results of this study confirms that the proposed *pMetrics* outperform their linear-based variants in terms of the discriminative power, suggesting once again that a log-based normalization is indeed better suited for these metrics. In addition, we observe that *pMetrics* are generally more discriminative than their standard counterparts in detecting a change throughout the experiments, suggesting that they are reliable online metrics for controlled experiments. In the third group of experiments, we evaluate the stability of *pMetrics* in A/A tests. The observed results are similar to those from the previous experiments, indicating that *pMetrics* consistently outperform their linear-based variants as well as the standard metrics. Finally, through a correlation study, we compare *pMetrics* with their linear-based variants in terms of how correlated they are with the standard metrics. The results observed here show that *pMetrics* correlate better with standard metrics as opposed to the linear-based variants, further supporting the previous empirical results presented in the paper.

Overall, the proposed framework can be seen as the first yet significant step towards reducing the effect of user bias in evaluating online controlled experiments. While the proposed log-based normalization factor appears to be suited for these metrics, investigating different variants of the normalization in a more systematic way is a future direction. In addition, different types of search tasks [11, 27, 31] as well as the variability of user behavior across these tasks are among factors that can be considered to better calibrate these metrics with respect to user behavior in personal search.

REFERENCES

[1] Azin Ashkan and Charles LA Clarke. 2012. Modeling browsing behavior for click analysis in sponsored search. In *Proceedings of the* 21^{st} *ACM international conference on information and knowledge management*. ACM, 2015–2019.

[2] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the* 35^{th} *international ACM SIGIR conference on research and development in information retrieval*. ACM, 185–194.

[3] Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the* 27^{th} *annual international ACM SIGIR conference on research and development in information retrieval*. ACM, 25–32.

[4] David Carmel, Guy Halawi, Liane Lewin-Eytan, Yoelle Maarek, and Ariel Raviv. 2015. Rank by time or by relevance?: Revisiting email search. In *Proceedings of the* 24^{th} *ACM international on conference on information and knowledge management*. ACM, 283–292.

[5] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the* 40^{th} *international ACM SIGIR conference on research and development in information retrieval*. 15–24.

[6] Charles LA Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 75–84.

[7] Dotan Di Castro, Liane Lewin-Eytan, Yoelle Maarek, Ran Wolff, and Eyal Zohar. 2016. Enforcing k-anonymity in web mail auditing. In *Proceedings of the ninth ACM international conference on Web search and data mining*. ACM, 327–336.

[8] Susan Dumais, Edward Cutrell, Jonathan J Cadiz, Gavin Jancke, Raman Sarin, and Daniel C Robbins. 2016. Stuff I've seen: a system for personal information retrieval and re-use. In *ACM SIGIR Forum*, Vol. 49. ACM, 28–35.

[9] Bradley Efron. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* 68, 3 (1981), 589–599.

[10] David Elsweiler, Morgan Harvey, and Martin Hacker. 2011. Understanding re-finding behavior in naturalistic email interaction logs. In *Proceedings of the* 34^{th} *international ACM SIGIR conference on research and development in information retrieval*. ACM, 35–44.

[11] David Elsweiler and Ian Ruthven. 2007. Towards task-based personal information management evaluations. In *Proceedings of the* 30^{th} *annual international ACM SIGIR conference on research and development in information retrieval*. ACM, 23–30.

[12] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online Evaluation for Information Retrieval. *Foundations and trends in information retrieval* 10, 1 (2016), 1–117.

[13] Botao Hu, Yuchen Zhang, Weizhu Chen, Gang Wang, and Qiang Yang. 2011. Characterizing search intent diversity into click models. In *Proceedings of the* 20^{th} *international conference on World Wide Web*. ACM, 17–26.

[14] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the eighth ACM international conference on Web search and data mining*. ACM, 57–66.

[15] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. In *Proceedings of the* 37^{th} *international ACM SIGIR conference on research and development in information retrieval*. ACM, 607–616.

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the* 28^{th} *annual international ACM SIGIR conference on research and development in information retrieval*, Vol. 51. ACM, 4–11.

[17] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does Relevance Mean Usefulness and User Satisfaction in Web Search?. In *Proceedings of the* 39^{th} *International ACM SIGIR conference on research and development in information retrieval*. ACM, 463–472.

[18] Filip Radlinski and Thorsten Joachims. 2006. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *Proceedings of the national conference on artificial intelligence*, Vol. 21. 1406–1412.

[19] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the Bootstrap. In *Proceedings of the* 29^{th} *annual international ACM SIGIR conference on research and development in information retrieval*. ACM, 525–532.

[20] Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Zhicheng Dou, and Chin-Yew Lin. 2010. Simple Evaluation Metrics for Diversified Search Results.. In *EVIA@ NTCIR*. 42–50.

[21] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In *Proceedings of the* 36^{th} *international ACM SIGIR conference on research and development in information retrieval*. ACM, 473–482.

[22] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the* 33^{rd} *international ACM SIGIR conference on research and development in information retrieval*. ACM, 555–562.

[23] Si Shen, Botao Hu, Weizhu Chen, and Qiang Yang. 2012. Personalized click model through collaborative filtering. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 323–332.

[24] Mark D Smucker and Charles LA Clarke. 2012. Modeling user variance in time-biased gain. In *Proceedings of the symposium on human-computer interaction and information retrieval*. ACM.

[25] Mark D Smucker and Charles LA Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the* 35^{th} *international ACM SIGIR conference on research and development in information retrieval*. ACM, 95–104.

[26] Alan Stuart. 1983. Kendall's tau. *Encyclopedia of statistical sciences* (1983).

[27] Sarah K Tyler and Jaime Teevan. 2010. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 191–200.

[28] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing and management* 36, 5 (2000), 697–716.

[29] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. 2016. Learning to rank with selection bias in personal search. In *Proceedings of the* 39^{th} *international ACM SIGIR conference on research and development in information retrieval*. ACM, 115–124.

[30] Ryen W White and Diane Kelly. 2006. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the* 15^{th} *ACM international conference on information and knowledge management*. ACM, 297–306.

[31] Steve Whittaker, Tara Matthews, Julian Cerruti, Hernan Badenes, and John Tang. 2011. Am I wasting my time organizing email?: a study of email refinding. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 3449–3458.

[32] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proceedings of the* 31^{st} *annual international ACM SIGIR conference on research and development in information retrieval*. ACM, 587–594.

[33] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the* 19^{th} *international conference on World Wide Web*. ACM, 1011–1018.