Large-Scale Analysis of Viewing Behavior: Towards Measuring Satisfaction with Mobile Proactive Systems

Qi Guo*, Yang Song*

Google Inc.
1600 Amphitheater Parkway,
Mountain View, CA 94043
{qiguo, yangso}@google.com

ABSTRACT

Recently, proactive systems such as Google Now and Microsoft Cortana have become increasingly popular in reforming the way users access information on mobile devices. In these systems, relevant content is presented to users based on their context without a query in the form of information cards that do not require a click to satisfy the users. As a result, prior approaches based on clicks cannot provide reliable measurements of user satisfaction with such systems. It is also unclear how much of the previous findings regarding *good abandonment* with reactive Web searches can be applied to these proactive systems due to the intrinsic difference in user intent, the greater variety of content types and their presentations.

In this paper, we present the first large-scale analysis of viewing behavior based on the viewport (the visible fraction of a Web page) of the mobile devices, towards measuring user satisfaction with the information cards of the mobile proactive systems. In particular, we identified and analyzed a variety of factors that may influence the viewing behavior, including biases from ranking positions, the types and attributes of the information cards, and the touch interactions with the mobile devices. We show that by modeling the various factors we can better measure user satisfaction with the mobile proactive systems, enabling stronger statistical power in large-scale online A/B testing.

Keywords

large-scale log analysis, viewport modeling, satisfaction measures, mobile proactive systems

*The authors are ordered alphabetically. Work done while both authors were at Microsoft.

CIKM'16 October 24-28, 2016, Indianapolis, IN, USA

ACM ISBN 978-1-4503-4073-1/16/10.
DOI: http://dx.doi.org/10.1145/2983323.2983846

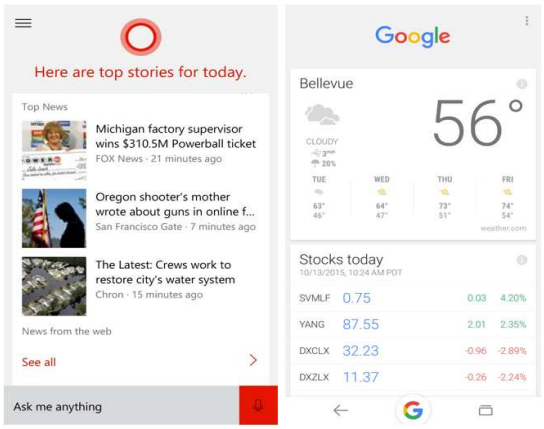


Figure 1: This figure shows the variety of information cards in two proactive systems: Cortana (left) and Google Now (right).

1. INTRODUCTION

Proactive systems such as Google Now and Microsoft Cortana have become increasingly popular on mobile devices in recent years. In these systems, relevant content is presented to the users based on the context and the personal interests without requiring users to submit a query. Due to the highly contextual and personal nature of the proactive system, it is challenging to measure the user satisfaction, especially on a large scale. One possible solution to this is to leverage interactions such as clicks as implicit relevance feedback from users, which were found to be successful in various applications and previous research [17, 7, 4, 14, 15, 20, 21, 34].

However, the presented content in the proactive systems is usually in the form of information cards, which often contains sufficient information without requiring users to click. As a result, previous approaches based on click signals would fall short as lack of click may actually represent *good abandonment* rather than dissatisfaction [24, 32] as was found in previous research for reactive Web searches [5]. Instead, viewing behavior or user attention, as captured by the viewport changes on mobile devices [10, 9, 23], could be valuable to derive satisfaction measures for the mobile proactive systems.

Yet, it is unclear how much of the previous findings regarding good abandonment and the viewport-based attention can be applied to the proactive systems, due to the

intrinsic difference in user intent [30] and the greater variety of content presented on individual impressions of the proactive systems. Figure 1 illustrates some examples of information cards for Google Now and Microsoft Cortana. As we can see, some information cards, like weather and stocks are simple answers that may not require users to click on to satisfy their information needs, while other cards such as news cards would require longer view time to parse the content and are more likely to receive clicks from users to read about the news articles of interest. Such variety imposes a great challenge to develop accurate and robust satisfaction measures from viewport-based viewing behavior that has not been addressed in previous research [9, 23].

Another limitation of most of the previous research [23, 10, 9] in this new area is that the data is collected from controlled user studies where the number of samples is small thus it is unclear how much the findings would be generalized on a large scale.

In this paper, we aim to fill in the gaps by conducting large-scale analysis of the viewport-based viewing behavior on mobile devices and focus on understanding and modeling the viewing behavior to better measure the user satisfaction with such systems. We make the following contributions:

- presents the first large-scale analysis of viewport-based viewing behavior on mobile devices;
- presents the first in-depth study on developing satisfaction measures for mobile proactive systems;
- identifies and characterizes a variety of biases that may impact the viewing behavior;
- demonstrates the effectiveness of viewport-based metrics for evaluating proactive systems on a large scale;
- demonstrates the improvements of viewport-based metrics through addressing the various identified biases.

2. RELATED WORK

Due to the difficulty of collecting large amount of user explicit labels for relevance measurement, in information retrieval community, researchers have elected to leverage implicit feedback from users to infer relevance. For Web search, click signals have been widely used as implicit feedback to measure user satisfaction at scale [16, 18, 33]. However, click-through is noisy as some clicks may be drawn because of the attractiveness of the search results rather than its intrinsic relevance [18]. As a result, users may bounce back from the landing page to the search engine result pages (SERPs) shortly after the click [18, 11] if the clicked result was not relevant, while spend longer time on the landing page if the search result was relevant [26].

To address the presentation bias of click-through, the post-click dwell time of the landing page is widely adopted [26, 20, 21, 34]. For example, in [26], the authors conducted a study focusing on the correlation between explicit feedback and news article reading time, and discovered that readers tend to spent significantly more time on interesting articles. In addition, they also discovered that article relevance only has weak correlation with their lengths, hence concluding that most readers only read part of the articles. However, some follow-up research works have suggested that the dwell time metric is not universally applicable to all types of retrieval tasks [20, 21], particularly for tasks with complex information needs [21]. Hence, the threshold needs to be carefully tuned according to the task types [34].

In addition, research has been conducted to measure a variety of user behavioral signals as implicit feedback. For example, in [4], the authors studied the relationship between mouse scrolling and page relevance and discovered their positive correlation. In [7], the authors collected over 30 implicit implicit measures from users. The study indicated that the combination of the right measures can lead to good predictions of user satisfaction. Among them, the exit type (how users left the page) and time to first click are good indicators to measure relevance. Furthermore, user behavior on the SERPs, when combined with page dwell-time and session level information, were found to significantly improve result ranking in the aggregate (e.g., [1]), and can be further improved by personalizing these measures (e.g., [25]). Fine-grained implicit feedback has been studied as well. For example, Buscher et al. [2] rely on eye-tracking data to determine which parts of a document have been read, skimmed, or skipped. The read and skimmed parts of the document were taken as relevant, while skipped document parts were ignored. The authors report considerable improvements for re-ranking of result lists, when including gaze-based feedback on the segment level compared to relevance feedback on the document level. The limitation of leveraging eye-tracking though is its lack of scalability due to its limited accessibility [2, 23].

Mouse activity is another important channel, emerged in recent years, to collect implicit feedback, which not only captures user attention at the similar fine granularity to eye-tracking but is also highly scalable due to its prevalence. One of the earliest research in this area is by Rodden et al., where the authors identified the coordination patterns between mouse and eye-movements [29]. Following this research, Guo and Agichtein [8] showed that gaze positions can be accurately predicted through modeling mouse movements, and Huang et al. [13] conducted more in-depth analysis and derived insights to improve the gaze-prediction models. Going beyond the regular Web search results, Navalpakkam et al. [27] conducted a controlled study to understand the mouse and eye movement patterns regarding knowledge panels on the right hand side of the SERPs, and also identified the coordination between the two. Mouse activity was also found to be useful for predicting search result relevance. In particular, cursor hovering and scrolling are found to better predict user clicks than other signals [14] and can be used as a good indicator to distinguish good and bad search abandonments [15], especially, for results that do not require a click to satisfy users, such as knowledge panels that often provide information snippets on the right hand side of the SERP [27]. Going beyond SERPs, Guo and Agichtein [8] also discovered stronger correlation of page relevance from mouse cursor movements compared to dwell time, which enables substantial improvements in relevance prediction and Web search result re-ranking. The relevance prediction model is further improved by Lagun et al. through mining the most frequent motifs of the mouse movements [22].

Despite the success, the aforementioned implicit feedback and behavioral signals were mainly designed for traditional desktop devices with keyboards and mice. However, with the increasing popularity of mobile touch-based devices such

as smart phones and tablet PCs, user behavior has greatly changed for Web search [31] and the existing behavioral signals need to be adapted to remain effective as implicit feedback and satisfaction measures for these new devices[10, 12, 9, 23]. Given the relatively small screens of the mobile devices, the viewport (i.e., the visible part of a Web page) is found to correlate well with user attention [23], and the way users change the viewports is found to be effective to identify the relevance of the viewed Web pages [9] and the elements on the SERPs [23] in two recent studies, respectively. In particular, Guo et al. [9] found that the "inactive time", i.e., the time spent on the stabilized viewports, and the speed and frequency users change the viewports are more accurate indicators of document relevance compared to the dwell time on the touch-enabled mobile devices. Complementarily, Lagun et al. [23] demonstrated a strong correlation between user attention in terms of eye-gaze movements and viewport-based viewing time for mobile Web searches, finding that increased scrolling and increased time below were strong signals of answer dissatisfaction. Yet, as far as we know, the studies on this topic have solely based on controlled user studies so far, and no studies have been conducted to analyze the viewing behavior based on viewports on mobile devices on a large scale, making it unclear how the existing findings might generalize.

Most recently, mobile proactive systems have become increasingly popular for accessing information on the mobile devices, which not only require *zero-query* but also present content to users in the form of cards that require *no clicks* to satisfy their information needs. To the best of our knowledge, Shokouhi and Guo [30] is the first and still only study on user interaction with these systems, which focuses on understanding the general usage patterns and discovering connections between the reactive Web search behavior and the user interactions with the proactive systems. Yet, the understanding of viewport-based viewing behavior and the modeling of the behavior to better measure user satisfaction remains an important opened question.

In this work, we aim to address these two important questions by conducting a large-scale analysis on viewport-based viewing behavior, with a focus on developing satisfaction measures of the newly emerged proactive systems. In particular, we compare and contrast the behavioral differences between traditional desktop search, mobile reactive search, and the zero-query proactive interaction paradigm, and previous findings from smaller scale controlled user studies. We show that the viewport-based viewing behavior can improve the measurement of user satisfaction with proactive systems compared to using click and dwell time based measures. We also identify a few important factors that may influence the viewing behavior of the users, and show that by modeling these factors, we may better measure user satisfaction.

3. DATA DESCRIPTION & TERMINOLOGY

The data set used in this paper was collected from a commercial personal digital assistant, namely Microsoft Cortana, which provides proactive recommendations to millions of users on their mobile devices every day, based on the context and their interests. While reactive impressions start with a query, a proactive impression is triggered when the user launches the digital assistant. Similar to the reactive search scenarios, our proactive logs are organized as *proactive impressions*, each of which consists of a ranking of proactive cards presented to the user together with corresponding interaction logs recorded such as clicks, viewports and scrolling. Each of the *proactive cards* are designed to satisfy a domain-specific set of information needs such as news, finance, traffic, sports, and travel (see examples in Figure 1). As the displayed content of the card is often sufficient to satisfy the user's information need without requiring a click [30], viewport logging is needed to infer how much time users spent viewing the content and, in turn, how likely they are satisfied with the content.

In our study, *viewport logging* is enabled through JavasScript embedded in the proactive impressions, and the viewport data is buffered and then sent back to the server through HTTP requests. We record the screen size, the positions and sizes of the cards rendered on the proactive impressions in pixels, as well as the viewport changing events with timestamps, allowing us to reconstruct the viewing behavior of the users and calculate the time users spent dwelling on each card.

Given the viewport logging, the *view time*, or *reading time*[1] of the card can then be derived by distributing durations of viewports to each card based on its *coverage* and *exposure* as defined in [23], where *coverage* is defined as how much of the card area was visible to a user and *exposure* is defined as how much of the viewport real estate did the card occupy. The total *view time* for a card is then computed as the sum of coverage and exposure-weighted time across all viewports. This calculation of view time was found to be best correlated with gaze data collected from eye-tracker [23] among a few different variants, exhibiting strong correlations (e.g., 0.7 for %time on an element between gaze-based and viewport-based times).

Now that we have view time inferred, we can further derive *user satisfaction metrics* from the view time. For example, we can threshold on the view time to derive the notion of a *SAT view*, which, in previous research (e.g., [30]) was determined as a view with duration above 30 seconds (following earlier work on determining *SAT click* [7]). However, no prior work was done to understand whether this widely used threshold of post-click dwell time is optimal for determining user satisfaction from proactive information card viewing without a click. This is the key research question we aim to address in this paper.

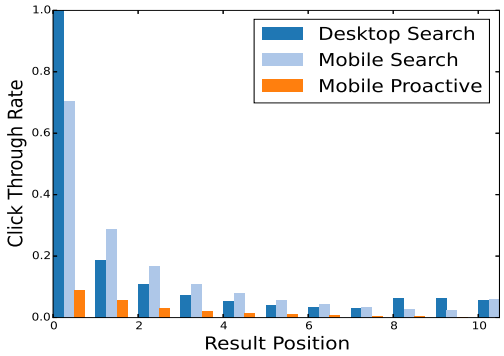
In our proactive logs, each user is represented using a consistent and anonymized identifier in the system. For our experiments we first randomly sampled 500,000 unique users from a period of 1 month between June 1 to July 1, 2015. In total, we observed over 3 million impressions from the proactive logs. We then extracted the mobile reactive logs from a commercial search engine (Microsoft Bing Mobile) for the same set of users to capture their mobile reactive queries and interactions during the same time period. In total, there are over 6.5 million reactive impressions. In addition, we collected 1-month search logs from the Web vertical of Microsoft Bing for a random sample of 500,000 users[2] in the U.S. search market. Those Web search users have a total of 3 million impressions. The statistics are summarized in Table 1.

[1]We use these two terms interchangeably in this paper.

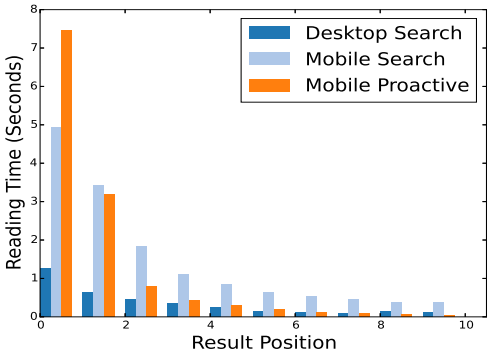
[2]The anonymized identifier for desktop search is not consistent with the mobile logs.

Table 1: Summary of data sets used in this paper.

Data Set	# Users	# Impressions
Mobile Proactive	500,000	3,182,863
Mobile Reactive	500,000	6,572,829
Desktop Search	500,000	3,993,096



(a) Result CTR (Scaled)



(b) Result Reading Time

Figure 2: Positional bias of results in three scenarios: Desktop search, Mobile search and Mobile Proactive.

4. ANALYZING VIEWING BEHAVIOR

In this section, we present in-depth analysis on the viewing behavior using large-scale interaction logs. We organize the discussion around a variety of factors that may influence the viewport-based viewing behavior and may lead to biases in designing user satisfaction metrics.

4.1 Positional Bias

The positional bias of Web search results have been extensively studied [16, 3, 18, 6]. Generally speaking, a result that is ranked higher tend to receive more clicks than a lower-ranked result, not only because its higher relevance but also because the higher chance of being viewed and higher *perceived* relevance due to its ranking position. In our study, we compare click through rate (CTR) and view time (calculated as described in Section 3) by position across the thee different experiences, namely, mobile proactive experience, mobile reactive search, and desktop reactive search.

In Figure 2 (a), we show the CTR[3] of top-10 results returned by the three experiences. In addition to the obvious positional decay, two interesting observations can be made. First, the CTR on the proactive cards are order-of-magnitude smaller compared to the reactive search scenarios, confirming that the proactive cards are indeed less likely to require a click to satisfy the users. Note that the content of the information cards tend to have higher quality on average compared to reactive search results by design, so the overall lower CTR on proactive cards cannot be explained by overall low quality [30]. The CTR for mobile reactive is also relatively lower compared to desktop search, which is likely due to the prevalence of instant answers that result in similar *good abandonments* [23]. Second, the decay of mobile scenarios are much more smoothed (decay factors 0.45 for proactive and 0.69 for mobile) compared to the sharp drop of CTR from the first position for desktop reactive search (decay factor: 1.48). This observation about reactive search scenarios is consistent with previous findings on lower reformulation rate on mobile, where users tend to examine more results due to lower network connection and more difficulty in typing as identified in previous research (e.g., [19, 9]).

Positional decay of view time is also observed as shown in Figure 2 (b). In contrast to CTR, the average view time for the two mobile scenarios are much higher, which can be also explained by the higher prevalence of answer like results presented in the mobile scenarios. Interestingly, while the decay of reactive searches appear to be pretty smoothed (decay factors 0.43 for desktop and 0.41 for mobile), there is a sharp drop of view time for mobile proactive (decay factor: 0.93). One explanation is related to the peek view of the proactive experience, where users may see the top half of the first card (under the reactive search box) before entering proactive. As a result, users are more likely to enter the proactive experience when the top card is relevant, and spend more time viewing it. This again confirms the utility of view time for proactive, as similar decay is not observed for the sparser signal in CTR for proactive.

The monotonic positional decay of view time distribution for mobile reactive search is also particularly interesting as it stands against findings in previous work [23] based on a controlled user study of 24 users and 20 tasks. In that study, the authors discovered a *bump* for the view time distribution around the second and third position, where the view time is calculated using the identical formula based on viewports as in this paper. The explanation of the difference may be due to the nuances in the small scale study – e.g., the knowledge answers that the study focused on tend to appear around the second position, which may attract more attention from users. In contrast, our analysis based on millions of impressions covers a wider variety of mobile search experiences, which is more *representative* and *generalizable*. In our future work, we plan to look into different slices of queries to further understand the variance among query classes in view time positional distribution.

4.2 Card Type Bias

Figure 3 shows the biases in view time and CTR for different card types. Note that, as certain types of cards may be ranked consistently higher than others, the higher view time and/or CTR they receive may be (in part) due to the posi-

[3]Due to the sensitivity of CTR, all reported CTR numbers in this paper are scaled w.r.t. the largest value in the context.

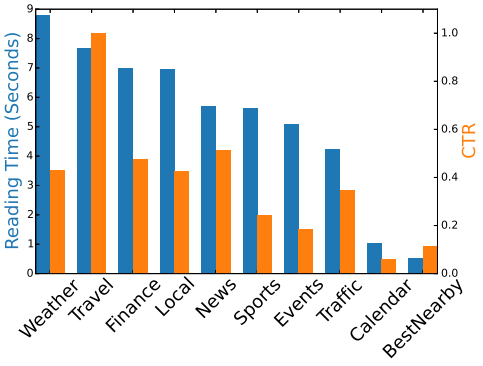


Figure 3: This figure shows the biases in view time and CTR for different card types. The bars are ordered decreasingly by view time (left y-axis).

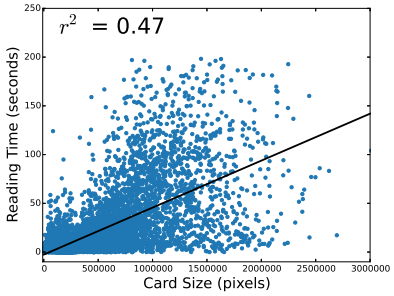


Figure 4: This figure shows the correlation between card size and view time. The card size is calculated by its area = width * height (pixels).

tion bias. To decouple from positional bias in this analysis, we consider card impressions that are at *top position*. As we can see, Travel and News cards have the highest CTR due to their interactive nature. In particular, the Travel card is a collection of mini-cards with snippets of relevant information (e.g., flight, traffic, nearby attractions) and often requires a click to get full information. Similarly, News card usually only shows the title and a snippet of the news and a click is required to obtain the full content. In contrast, cards such as Sports, which shows match scores, have high view time but low CTR as most of the needed information can be obtained via its displayed content.

4.3 Presentation Bias

The other aspect of card bias comes from the difference in their presentations. It is common to assume that richer presentation often leads to higher user engagement and better user satisfaction. Indeed, both eye-gaze and mouse-track studies shown that regardless of their relevance, results with images present always attract much more attention that those with plain text [23]. We conduct such study in our data and was able to observe the similar bias for the proactive systems. As shown in Tale 2, we can see that cards with images yield much higher view time and CTR, regardless of their positions and relevance.

In addition, in our scenario, we also study how the size of cards affect the view time. To achieve this, we measure the size of cards by its area of pixels. E.g. a card of size 300 by

Table 2: This table shows the difference between cards that have images and have no images. Statistical significance (SS) tests indicate both view time and CTR are different.

Card Type	View Time	CTR
With Images	2.07±2.34	0.019±0.03
No Images	1.53±1.55	0.011±0.01
SS Different?	Yes (p-val< 0.01)	Yes (p-val< 0.01)

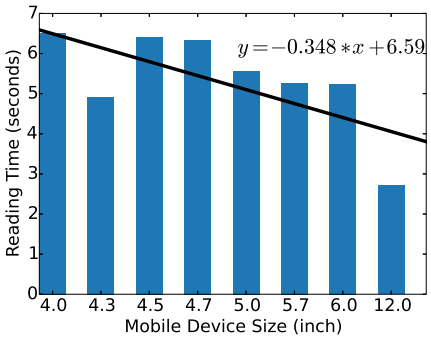


Figure 5: This figure shows the average view time per card by the size of the mobile devices.

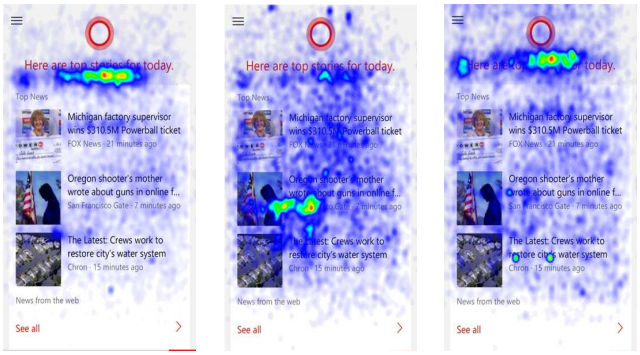
200 pixels has a 60,000 pixel-area. Figure 4 shows the relationship between card size and the average view time using a scatter plot. In general, we can observe that cards with large sizes tend to yield higher view time. The R-squared value of 0.47 also confirms a significant positive correlation between these two variables.

4.4 Device Bias

Another factor that may bias viewing behavior lies in the difference of the device size. While bigger devices may present multiple cards in one viewport, smaller devices may only present a fraction of a single card. To study this effect, we break down the impressions by the display size of devices, which is the diagonal measure of the screen. In Figure 5, we illustrates the effect of device sizes to the average viewing time for each proactive impression. In general, we observe a reduction of view time as the device size increases (except for those 4.3-inch phones). The best-fit one-degree polynomial curve has -0.34 slope and 5.56 residual scores, with an r-squared value of 0.48 and p-value of 0.04, confirming the negative correlation between view time and device size. This discovery indicates that bigger screens can help users fulfill their information need more efficiently, as users don't have to change viewports as frequently, which introduces unnecessary overhead. To enable more accurate measures of satisfaction across devices, adjusting for such bias can be beneficial.

4.5 Attention Shift on Swiping Directions

As shown in 4.1 and previous research [16, 3, 18, 6], users tend to focus their attention on the top of the screen, however, we hypothesize that the distribution of user attention may change as users swipe to change their viewports. To test this hypothesis at scale, we used clicks as proxy. The idea is that the click position would align with user attention even though clicks are sparser given the prevalence of proac-



(a) no swipe (b) swipe down (c) swipe up

Figure 6: This figure shows the heatmaps of user clicks after user swipes are performed. The data is collected from phones with screen size of 4.7 inches, which all have the same dimension (432*585 pixels).

Table 3: This table shows the user click statistics on 4.7-inch phones.

Swipe Direction	Count	Click Position (y-axis)
No Swipe ➡	295,311	265.48
Swipe Up ⬆	22,783	271.82
Swipe Down ⬇	198,105	313.69

tive cards not requiring a click to satisfy the users. In the future, we plan to confirm the findings with an eye-tracking study.

Figure 6 shows the heatmaps that represent the click distributions conditioned on the swipe directions. As expected, for initial viewports without a swipe and viewports with swipe-ups, we found that the clicks/attention are focused at the top of the screen (Figure 6 (a) and (c)), but for viewports with swipe-downs, we found that the clicks/attention are focused at the bottom of the screen (Figure 6 (b)). This makes sense as users may tend to focus on the fraction of the screen where new contents show up while swiping to change the viewports.

In Table 3, we report the statistics of user click positions for these three conditions. As we can see, most of the viewports with clicks are initial viewport (No Swipe), and users tend to swipe down (198,105) much more often than swipe back up (22,783). The click positions, relative to the viewport y-coordinate, are also consistent with what we observed in Figure 6. We further perform two-sampled independent t-test on each pair of user click data: ttest(NoSwipe, SwipeDown) = -135.474, p-val = 0; ttest(NoSwipe, SwipeUp) = -8.837, p-val = $1.05e^{-18}$; ttest(SwipeDown, SwipeUp) = 39.87, p-val = 0 and confirm that the differences between the three different conditions are statistically significant.

5. IMPROVING SAT METRICS

So far in the paper we have observed various factors that may bias the viewing behavior, hence impacting the effectiveness of using viewing behavior as a proxy of user satisfaction with the information cards. In this section, we show that our proposed model that addresses the various biases enabled better SAT metrics for measuring user satisfaction.

To evaluate the SAT metrics on a large scale, we ran an online A/B experiment for one week (from 3/11/2015 to 3/17/2015) with a sample of 1.3 million users, equally split

Algorithm 1: Measure Metric Sensitivity

Input: MT: The metric for evaluation,
$CTRL(MT)$: control users' MT values,
$TRET(MT)$: treatment user's MT values,
$N = \{10, 50, 100, 1000...\}$: the sample number,
M: number of times to repeat for each exp

for each n in N
  $winrate = 0$, $varList = Empty()$
  do: repeat M times
   $SampleCtrl \leftarrow SampleUser(n, CTRL(MT))$
   $SampleTret \leftarrow SampleUser(n, TRET(MT))$
   $SampleCtrlSum = sum(SampleCtrl)$
   $SampleTretSum = sum(SampleTret)$
   if $SampleTretSum > SampleCtrlSum$
    $winrate{+}{+}$
    $varList.add(1)$
   else
    $varList.add(0)$
   end if
  end do
  $winrate = winrate/M$, $std = STD(varList)$
end for

for control and treatment. Specifically, the control and treatment rankers were similar to the baseline default card ranker and the more sophisticated machine-learned Carre ranker, respectively, as described in [30], where the latter provides significantly better ranking quality compared to the former. We have also examined and triaged a variety of relevance and engagement metrics to confirm the better quality of the treatment ranker. In other words, we know for a fact that *treatment is better than control* and use this as the ground truth. We ensured that the magnitude of ranking quality improvement between this pair of experiments is representative of typical ranker releases, so that the findings are general.

To compare the effectiveness of different SAT metrics, we followed the procedure described in previous research [28]. Specifically, for a range of n, we randomly sample n users from each of treatment and control for m times, compute the means and variances for both, and see whether the difference in the metrics is in agreement with the known difference in ranking quality of control and treatment. A better metric is the one that has *higher agreement rate* with the ground truth with *lower variance*. Algorithm 1 sketches the details. In our experiment, we varied the number of sampled users n from 10 up to 500,000, and repeated $m = 10,000$ times for each n, to test metric stability and sensitivity.

For the baseline metric to compare with, we adopt the traditional click-based strategy as pseudo-relevance feedback labels, which is defined as:

- **SATClick**: For each proactive impression, if a user clicked on a card and viewed the content for more than 30 seconds, the card is considered relevant (gets a score of 1). Otherwise the card is non-relevant (gets a score of 0). The 30-second satisfaction cut-off of post-click dwell time has been found to be fairly optimal and has been widely used in prior work [7]. The comparison among different time threshold of SATClick is beyond the scope of this paper.

As a first attempt to combine view-based and click-based signals, we define a threshold-based metric based on card view time:

- **SATView(s)**: For each proactive card, if the view time of the card is greater than s seconds, the card is considered relevant. Otherwise the card is non-relevant. The view time calculation of an information card (as described in Section 3) follows previous research [23] .

In the literature, there has been no clear definition for the SAT cut-off value of SATView. Therefore, we aimed to fill the gap by experimenting with a variety of cut-off values s to measure the impact. In Table 4, we show the sensitivity values of SATClick, SATView(s) metrics with different s, as well the hybrid of these two types of metrics (i.e., SAT = SATClick **or** SATView(s)).

From the table, we can observe that when the sample rate is low (e.g., 10 random users), most of the metrics perform no better than random guess. With the increased number of sample users, some metrics gradually become more sensitive than others. The first and second columns show a head-to-head comparison between the performance of SATClick and SATView(30). We can easily observe that at all sample rate, SATView exhibited significantly higher win-rate and lower standard deviation. The result indicates that, *for the proactive scenario, SATView(30) is a more sensitive metric than SATClick.*

We then consider the hybrid version of the two metrics which is shown in the remaining columns of Table 4. In general, we observe that *combining SATClick and SATView indeed outperforms individual metrics significantly.* Comparatively, combining SATClick with SATView of longer view time ($\geq$ 10 seconds) tends to yield better performance than shorter view time (5 seconds). Nevertheless, statistical significance test does not indicate any significant difference among the sensitivity of SATView(10), SATView(15) and SATView(30). While SATView(15) and SATView(30) seem to converge faster to 1 with larger samples, SATView(10) shows higher sensitivity when the sample size is small.

5.1 Normalizing SATView by Card Size

So far we have observed a more sensitive metric that is a composition of two metrics using absolute time values to determine user satisfaction. As we recall in our previous discussion and shown in Figure 4, card view time is highly correlated with its area size, which leads to strong positive bias towards cards with large sizes. Therefore, we seek to correct such bias by proposing to normalize SATView by the size of the proactive cards, which is defined as (read as View Time per Pixel)

$$VTP = \frac{ViewTime}{CardWidth * CardHeight} \quad (1)$$

Table 5 depicts the results of VTP as a SAT metric, where *SATVTP (X PCTL)* is used to denote the use of X percentile of VTP as the SAT threshold. Overall, we observe that smaller VTP thresholds yield higher win-rate. Also, when the threshold is set to be too high, the metric may perform worse than random guess. Comparatively, VTP alone does not perform as well as SATView as a viewport-based metric as shown in the first two columns. However, the combination of SATVTP(25 PCTL) with SATClick (the last column) outperforms SATView(30) significantly and yields the best result in our analysis – on par with SATClick **or** SATView(30).

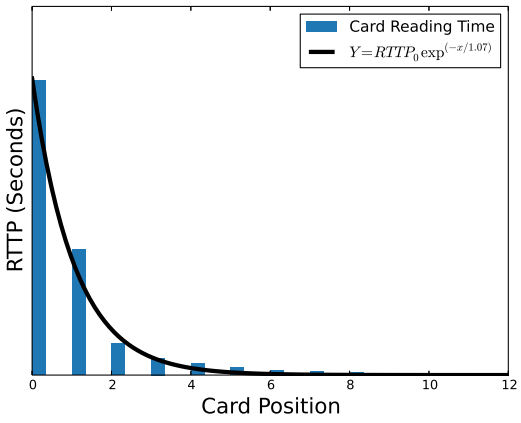


Figure 7: The fitting result of card view time using an exponential decay function.

The results from the above analysis suggest that in measuring user satisfaction in the proactive scenario, *both card view time and clicks* are essential in defining the correct SAT metric. Interestingly, normalizing card view time by size did not lead to significantly more sensitive metrics except for slight improvements on small number of samples.

Next, we conduct additional analysis to remove various other biases to further improve the metrics. We conducted analysis on both the normalized and unnormalized versions of the SAT metrics, and found that the normalized version yields similar yet more sensitive metrics with additional biases (e.g., position, card type) addressed, compared to the un-normalized version.

Therefore, due to the space constraint, we only report following results based on the normalized SAT metric next, which is defined as follows:

- **SATHybird** = SATClick **or** SATVTP (25 PCTL)

5.2 Addressing Positional Bias

The improvement of metric sensitivity so far encourages us to explore other opportunities to correct the view bias. In this section, we focus on addressing positional bias by dynamically adjusting SAT threshold for different positions. Recall that in the previous section setting VTP to its 25 Percentile yields the best performance, we thus take that as a baseline approach. From Figure 2 (b) we saw that the view time distribution looks like an exponential decay. Therefore, we propose to fit an exponential function to estimate the SAT view time at each position, which can be formulated as

$$SATVTP@i = N_0 e^{(-\frac{i}{\lambda})} \quad (2)$$

where i indicate the result position. Figure 7 shows the curve-fitting result. The optimal value of parameter λ is around 1.07, suggesting a linearly exponential decay. The starting value N_0 is set to be the best value of VTP_0, which according to our previous result, is 0.006.

Figure 8 illustrates an area plot comparing the original metric of SATHybrid and the enhanced SATHybrid with the position-decay equation integrated. In the figure, the red and blue lines correspond to the win-rate of baseline and position-decay metrics, respectively. The shaded areas around each line indicate the standard deviation of respec-

Table 4: This table shows the result of sensitivity analysis according to Algorithm 1, in the form of winrate±std. In general, metrics with higher win-rate and lower std are considered more sensitive.

# Users Sampled	Win-Rate					
	SATClick	SATView(30)	SATClick **or** SATView(5)	SATClick **or** SATView(10)	SATClick **or** SATView(15)	SATClick **or** SATView(30)
10	0.515 ± 0.500	0.526 ± 0.499	0.513 ± 0.500	**0.525 ± 0.499**	0.522 ± 0.500	0.523 ± 0.499
50	0.525 ± 0.499	0.548 ± 0.498	0.530 ± 0.499	0.545 ± 0.498	0.545 ± 0.498	**0.549 ± 0.498**
100	0.534 ± 0.499	0.561 ± 0.496	0.541 ± 0.498	0.562 ± 0.496	0.565 ± 0.497	**0.565 ± 0.496**
500	0.578 ± 0.494	0.629 ± 0.483	0.591 ± 0.492	**0.641 ± 0.480**	0.639 ± 0.480	0.638 ± 0.481
1,000	0.612 ± 0.487	0.676 ± 0.468	0.632 ± 0.482	**0.700 ± 0.458**	0.692 ± 0.461	0.695 ± 0.461
5,000	0.738 ± 0.440	0.847 ± 0.360	0.768 ± 0.422	**0.873 ± 0.333**	0.873 ± 0.334	0.865 ± 0.342
10,000	0.815 ± 0.388	0.930 ± 0.255	0.849 ± 0.358	0.952 ± 0.215	0.948 ± 0.222	**0.948 ± 0.221**
20,000	0.900 ± 0.300	0.984 ± 0.127	0.930 ± 0.255	**0.992 ± 0.087**	0.991 ± 0.094	0.989 ± 0.104
30,000	0.949 ± 0.220	0.997 ± 0.055	0.975 ± 0.156	0.998 ± 0.045	0.998 ± 0.032	**0.999 ± 0.032**
50,000	0.982 ± 0.133	1.000 ± 0.000	0.995 ± 0.071	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
100,000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

Table 5: This table shows the result of sensitivity analysis for the VTP metric in eq. (1), in the form of winrate±std. VTP Combined with SATClick becomes the most sensitive metric.

# Users Sampled	Win-Rate					
	SATView(30)	SATVTP (Mean)	SATVTP (25 PCTL)	SATVTP (50 PCTL)	SATVTP (75 PCTL)	SATClick or SATVTP (25 PCTL)
10	0.526 ± 0.499	0.500 ± 0.500	0.509 ± 0.500	0.502 ± 0.500	0.497 ± 0.500	**0.526 ± 0.499**
50	0.548 ± 0.498	0.499 ± 0.500	0.513 ± 0.500	0.504 ± 0.500	0.496 ± 0.500	**0.555 ± 0.497**
100	0.561 ± 0.496	0.497 ± 0.500	0.519 ± 0.500	0.503 ± 0.500	0.493 ± 0.500	**0.563 ± 0.495**
500	0.629 ± 0.483	0.489 ± 0.500	0.537 ± 0.499	0.506 ± 0.500	0.483 ± 0.500	**0.631 ± 0.477**
1,000	0.676 ± 0.468	0.489 ± 0.500	0.555 ± 0.497	0.516 ± 0.500	0.480 ± 0.500	**0.684 ± 0.465**
5,000	0.847 ± 0.360	0.457 ± 0.498	0.628 ± 0.483	0.529 ± 0.499	0.448 ± 0.497	**0.863 ± 0.326**
10,000	0.930 ± 0.255	0.451 ± 0.498	0.679 ± 0.467	0.537 ± 0.499	0.425 ± 0.494	**0.942 ± 0.218**
20,000	0.984 ± 0.127	0.426 ± 0.494	0.738 ± 0.440	0.544 ± 0.498	0.397 ± 0.489	**0.994 ± 0.105**
30,000	0.997 ± 0.055	0.398 ± 0.489	0.785 ± 0.411	0.558 ± 0.497	0.372 ± 0.483	**0.998 ± 0.035**
50,000	1.000 ± 0.000	0.410 ± 0.492	0.841 ± 0.366	0.587 ± 0.492	0.306 ± 0.461	1.000 ± 0.000
100,000	1.000 ± 0.000	0.369 ± 0.483	0.936 ± 0.245	0.607 ± 0.488	0.258 ± 0.438	1.000 ± 0.000

tive win-rate. It is evident to see that by adding position-decay to the SATVTP metric, the win-rate significantly increases. In addition, the area of standard deviation is also substantially smaller than the baseline. We can observe that the position-decay metric quickly converges to 100% agreement rate with the ground truth given only 20,000 random users, while the baseline does not converge until given 2.5-time more (50,000) users.

5.3 Addressing Card Type Bias

Furthermore, to address the bias of different card types, our studies have suggested that cards differ greatly in both view time and CTR, according to previous discussion in Figure 3. Therefore, we propose to use different threshold values for each type of answers.

We started with the näive version of per-card-type normalization via statistics such as mean and median, yet none resulted in significant improvements in metric sensitivity. What ended up working was the per-card-type SAT threshold derived from time-to-click. The intuition here is that users are more likely to be satisfied with viewing the card when they click (to explore more), thus the time-to-SAT-click distribution is closer to the SAT view distribution. To implement this idea, we first filter out impressions where no

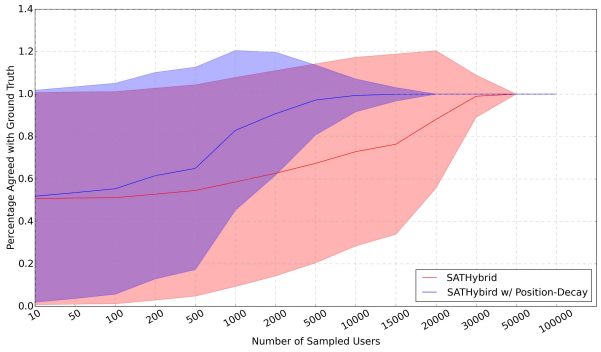


Figure 8: This figure shows the effect of addressing positional bias, which successfully increased the win-rate with less users (the blue line) as well as lowered the variance (the blue area).

cards receive any SAT clicks, thus leaving impressions with only *SAT cards* in our data. We then estimate the SAT for view time using the time-to-click from *the most-recent stable viewport* to the action of user clicks. After that, we aggregate these time periods for each type of card and compute

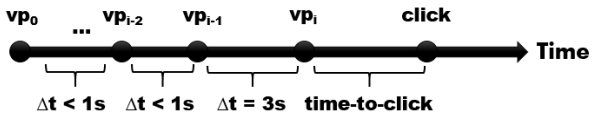


Figure 9: This figure shows the calculation of time-to-click value for setting the threshold of SAT views.

Table 6: This table shows the VTP SAT threshold derived for each type of answer, relative to the largest threshold of all types (i.e., News).

Card Type	Rel. VTP SAT Threshold (%)
News	100.00%
Local	57.62%
Weather	17.36%
Sports	15.35%
Traffic	14.05%
Travel	10.51%
BestNearby	10.39%
Calendar	6.85%
Finance	1.30%
Events	1.18%

the maximum likelihood of the mean and standard deviation to use as SAT threshold.

It is important to differentiate stable viewports from unstable ones when defining the SAT threshold. In unstable viewports, users are primarily searching/glancing instead of carefully viewing. We require a viewport to *at least last for one second* to be considered stable. Figure 9 illustrates an example.

Table 6 shows the SAT VTP threshold derived for each type of answer, relative to the largest threshold across all types (i.e., News). As we can see, the SAT VTP threshold for News card is significantly higher than other types of card, showing that users tend to read slower on news (i.e., spend more time viewing the same size of area) thus requiring higher threshold for users to be satisfied with news. At the other end of the spectrum are Finance and Events cards, which have much sparser content and require only quick glances for users to be satisfied.

In Figure 10, we compare the original version of SATHybrid with the card-type adjusted version. In particular, the card-type adjustment is even more effective compared to the position-bias adjustment, which converges even faster requring only 15,000 random users compared to the 20,000 random users required, as shown in Figure 8, for addressing the position bias.

5.4 Addressing Other Biases

Given the success of the addressing positional and card biases, a natural extension would be combining both of them to see whether a stronger metric can be found. In our experiment, we used a simple way to combine them by multiplying the card-specific SAT threshold (as shown in Table 6) at each position with the position-decay factor in eq. (2). Indeed, this combination further improves upon the SATHybrid with Card-Type Adjustment metric. However, the improvement is not statistically significant. Both metrics converge at 15,000 users and show very slight difference at 10,000 users (0.998 vs. 0.993 win-rate). This is understandable since the bias in card type may already encode part of the position bias as discussed earlier in Section 4.2.

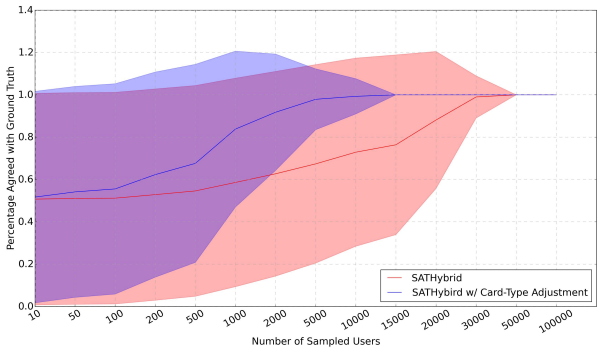


Figure 10: This figure shows the effect of addressing card-type bias, where we use the mean of view time to adjust each card type.

We have also made several efforts to address other types of biases identified in the earlier analysis in the paper, such as device size and shift of attention . Nevertheless, none of them were found to give significant further improvements over the original version of SATHybrid. We defer further investigation on this part in future work.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented, to the best of our knowledge, the first large-scale study of viewing behavior for mobile devices and the first in-depth analysis of viewport-based satisfaction metrics for mobile proactive systems. We compare and contrast the findings with previous research for reactive Web searches that are based on small-scale controlled user studies, identifying interesting similarities and differences. We have also identified and characterized a variety of factors that may influence the viewing behavior, including the positional bias, the various biases that are imposed by the types and attributes of the information cards, the devices, and how users interacted with the system.

Through running and analyzing data from a large-scale A/B live experiment, we demonstrated that the viewport-based metrics are more effective compared to metrics that are solely based on click-through and landing page dwell time in measuring user satisfaction with the proactive systems. We also showed that by addressing the various identified biases, in particular, through addressing the position and card type biases, we can further improve the satisfaction metrics significantly.

In the future, we plan to complement our analysis with controlled user studies to provide more qualitative insights into the proposed satisfaction measures. While having done some preliminary analysis comparing the two systems, we also plan to conduct large-scale analysis on viewing behavior for reactive Web searches, to have even deeper understanding of the similarities and differences in user behavior between the newly emerged mobile proactive systems and the long-existing reactive Web search systems.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06*, pages 19–26, 2006.

[2] G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 2991–2996, 2008.

[3] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *WWW '09*, pages 1–10, New York, NY, USA, 2009. ACM.

[4] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *IUI '01*, pages 33–40, 2001.

[5] A. Das Sarma, S. Gollapudi, and S. Ieong. Bypass rates: Reducing query abandonment using negative inferences. In *KDD '08*, pages 177–185, New York, NY, USA, 2008. ACM.

[6] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR '08*, pages 331–338, New York, NY, USA, 2008. ACM.

[7] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, Apr. 2005.

[8] Q. Guo and E. Agichtein. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW '12*, pages 569–578, 2012.

[9] Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *SIGIR '13*, pages 153–162, New York, NY, USA, 2013. ACM.

[10] Q. Guo, S. Yuan, and E. Agichtein. Detecting success in mobile search from interaction. In *SIGIR '11*, pages 1229–1230, New York, NY, USA, 2011. ACM.

[11] A. Hassan, Y. Song, and L.-w. He. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *CIKM '11*, pages 125–134, 2011.

[12] J. Huang and A. Diriye. Web user interaction mining from touch-enabled mobile devices. In *HCIR'12*, 2012.

[13] J. Huang, R. White, and G. Buscher. User see, user point: Gaze and cursor alignment in web search. In *CHI '12*, pages 1341–1350, 2012.

[14] J. Huang, R. W. White, G. Buscher, and K. Wang. Improving searcher models using mouse cursor activity. In *SIGIR '12*, pages 195–204, 2012.

[15] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: Using cursor movements to understand and improve search. In *CHI '11*, pages 1225–1234, New York, NY, USA, 2011. ACM.

[16] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

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

[18] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR 2005*, pages 154–161, 2005.

[19] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices. In *WWW '09*, pages 801–810, 2009.

[20] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback. In *SIGIR '01*, pages 408–409, New York, NY, USA, 2001. ACM.

[21] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *SIGIR '04*, pages 377–384, New York, NY, USA, 2004. ACM.

[22] D. Lagun, M. Ageev, Q. Guo, and E. Agichtein. Discovering common motifs in cursor movement data for improving web search. In *WSDM '14*, pages 183–192, New York, NY, USA, 2014. ACM.

[23] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR '14*, pages 113–122, New York, NY, USA, 2014. ACM.

[24] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR '09*, pages 43–50, New York, NY, USA, 2009. ACM.

[25] M. Melucci and R. W. White. Discovering hidden contextual factors for implicit feedback. In *Workshop on Contextual Information Retrieval (part of the 6th International and Interdisciplinary Conference on Modeling and Using Context)*, Jan. 2007.

[26] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *SIGIR '94*, pages 272–281, 1994.

[27] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *WWW '13*, pages 953–964, 2013.

[28] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *SIGIR '10*, pages 667–674, 2010.

[29] K. Rodden and X. Fu. Exploring how mouse movements relate to eye movements on web search results pages. In *Proceedings of ACM SIGIR 2007 Workshop on Web Information Seeking and Interaction*, pages 177–185, 2007.

[30] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *SIGIR '15*, 2015.

[31] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *WWW '13*, pages 1201–1212, 2013.

[32] Y. Song, X. Shi, R. White, and A. H. Awadallah. Context-aware web search abandonment prediction. pages 93–102, New York, NY, USA, 2014. ACM.

[33] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang. Enhancing personalized search by mining and modeling task behavior. In *WWW '13*, pages 1411–1420.

[34] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *CIKM '06*, pages 297–306, New York, NY, USA, 2006. ACM.