Mining Touch Interaction Data on Mobile Devices to Predict Web Search Result Relevance

Qi Guo*
Microsoft
One Microsoft Way
Redmond, WA 98052 USA
qiguo@microsoft.com

Haojian Jin
Emory University
400 Dowman Drive
Atlanta, GA 30322 USA
haojian.jin@emory.edu

Dmitry Lagun
Emory University
400 Dowman Drive
Atlanta, GA 30322 USA
dlagun@emory.edu

Shuai Yuan
Emory University
400 Dowman Drive
Atlanta, GA 30322 USA
syuan3@emory.edu

Eugene Agichtein
Emory University
400 Dowman Drive
Atlanta, GA 30322 USA
eugene@mathcs.emory.edu

ABSTRACT

Fine-grained search interactions in the desktop setting, such as mouse cursor movements and scrolling, have been shown valuable for understanding user intent, attention, and their preferences for Web search results. As web search on smart phones and tablets becomes increasingly popular, previously validated desktop interaction models have to be adapted for the available touch interactions such as pinching and swiping, and for the different device form factors. In this paper, we present, to our knowledge, the first in-depth study of modeling interactions on touch-enabled device for improving Web search ranking. In particular, we evaluate a variety of touch interactions on a smart phone as implicit relevance feedback, and compare them with the corresponding fine-grained interactions on a desktop computer with mouse and keyboard as the primary input devices. Our experiments are based on a dataset collected from two user studies with 56 users in total, using a specially instrumented version of a popular mobile browser to capture the interaction data. We report a detailed analysis of the similarities and differences of fine-grained search interactions between the desktop and the smart phone modalities, and identify novel patterns of touch interactions indicative of result relevance. Finally, we demonstrate significant improvements to search ranking quality by mining touch interaction data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – relevance feedback, search and selection process.

*Work was done as a Ph.D. student at Emory University

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Keywords

Mobile search behavior; touch interaction models; implicit relevance feedback

1. INTRODUCTION

Accurately estimating document relevance is at the core of Web search or information retrieval in general. Implicit feedback from users are good indicators of document relevance, among which, dwell time, or the time user spends on visiting the clicked Web search result document, has been found to be indicative of document relevance [9, 36] and has been successfully applied in various other Web search applications [35, 18, 10]. However, dwell time may be misleading – a user may spend 30 seconds reading a relevant paragraph in a document or struggling finding the information and skimming through during the 30 seconds.

In previous studies [33, 14, 23, 16, 22], fine-grained interactions such as mouse cursor movements and scrolling were found to be indicative of searcher interests and preferences. In particular, post-click behavior were found to capture the different viewing patterns of relevant and non-relevant document, resulting in a more accurate prediction of intrinsic document relevance [16], compared to using dwell time information alone. For example, extensive and slow mouse movements were found to correlate with reading a relevant document while fast scrolling are found to correlate with skimming a non-relevant document.

Recently, touch-enabled mobile devices, such as smart phones and tables, have become an increasingly popular modality for users to search and browse the Web. Despite the success of modeling fine-grained interactions on a personal computer with a mouse and a keyboard as the primary input devices, fine-grained Web search interactions on these new devices, where users zoom and swipe instead of using a mouse, are less understood. While these touch interactions may also provide additional signals of document relevance beyond dwell time, we are not aware of any reports of modeling these fine-grained touch interactions as implicit relevance feedback. Thus, we explore the following two research questions:

- Do users view search result documents differently on a *touch-enabled mobile device* compared to using a *desktop computer* with a mouse and keyboard as the primary input device?
- Do users behave differently on a *touch-enabled mobile de-*

vice when viewing a *relevant* search result document compared to a *non-relevant* document?

To answer these research questions, we analyzed searcher behavior on a smart phone and a desktop computer in two user studies, which involved 56 participants in total and hundreds of unique Web search queries and page examinations. In particular, we identify behavioral patterns that correspond to viewing relevant and non-relevant documents. Using these insights, we design a variety of touch interaction features to capture these patterns, and analyze the correlations of these feature values with the document relevance. We also compare the differences and similarities of the interaction patterns across these two modalities. Finally, we develop relevance prediction models based on the touch interaction features, and demonstrate significant improvements over using dwell time information alone.

In summary, our contributions include:

- A characterization of patterns of examination and touch interactions that correspond to viewing a relevant or non-relevant Web search result document (Section 4).
- MTI, a novel model of relevance prediction that captures fine-grained Web search interactions on touch-enabled mobile devices (Section 5).
- A comparison of the differences and similarities of fine-grained Web search interactions on a touch-enabled mobile device and a desktop PC (Section 6).
- Empirical evidence that MTI is more effective than using dwell time information alone, for ranking Web search result documents using the estimated relevance (Section 7).

Next, we briefly survey the background and related work to put our contribution in context.

2. RELATED WORK

Using page dwell time for inferring relevance has a long history in the information retrieval community, with mixed conclusions about its utility. Some of the early research done in the area of implicit feedback in information retrieval was that of Morita and Shinoda [30]. They conducted a study where participants were asked to provide explicit feedback about interestingness of news articles that they have read. The study focused on the correlation between reading time and explicit feedback, while considering document length and additional textual features. They noted that there is a strong tendency to spend more time on interesting articles rather than on uninteresting ones. Similar findings have also been reported in [9] and [11]. Furthermore, Morita and Shinoda found only a weak correlation between the lengths of articles and associated reading times, indicating that most articles are only read in parts, not in their entirety.

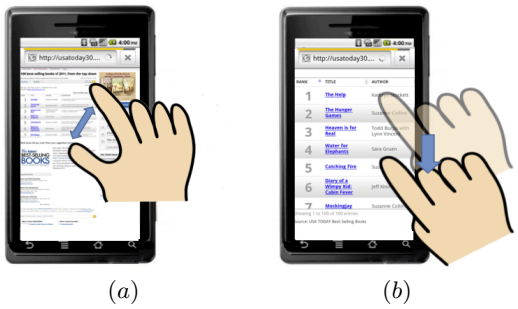
Interestingly, dwell time does not always correlate with relevance. Kelly and Belkin [27] tried to reproduce the results of Morita and Shinoda in a different, more complex information retrieval scenario, yet found no correlations between display time and explicit relevance ratings for a document. In a subsequent, naturalistic study, Kelly and Belkin [28] found again no general relationship between display time and the users' explicit ratings of the documents' usefulness. Instead, they observed high variation of display time with respect to different users and different tasks. Recently, White and Kelly [36] reported that adjusting display time thresholds for implicit feedback according to task type leads to improved retrieval performance, while adjusting the thresholds according to individual users degraded performance. This stands in contrast to findings of a prior study by Rafter and Smyth [32] who showed for one specific task type that display time is correlated with user interest, especially after individually adjusting the measure. In summary, while dwell time does appear to contain relevance information, previous studies have found many different interpretations of it, with no clear consensus of the relationship to document relevance.

Additional implicit measures have been examined on the object level (e.g., document paragraph or page item) as well. On one hand, it has been found that good indicators of interest include the amount of scrolling on a page [9], click-through [11, 25], and exit type for a Web page [11]. On the other hand, mouse cursor movements and mouse clicks while viewing a document do correlate with user interest [9]. Furthermore, user behavior on the Search Engine Result Pages (SERPs), when combined with page dwell-time and session level information, can significantly improve result ranking in the aggregate (e.g., [1]), and can be further improved by personalizing these measures (e.g., [29]).

Other previous efforts focused on modeling more explicit user interactions. Golovchinsky et al. [12] focused on user-created annotations on documents such as highlightings, underlinings, circles, and notes in margin. They used this kind of feedback to infer relevance of document passages. In a document search scenario utilizing query expansion, they reported a significant improvement of the annotation-based feedback technique over explicit relevance feedback on the document level. Ahn et al. [2] followed a similar idea but used the concept of a personal notebook where users could paste text passages worth remembering. The two previous approaches require some form of explicit and therefore rare user interactions (e.g., annotating, copying and pasting) to work properly. In contrast, Buscher et al. [5] only rely on implicit data, and determine which parts of a document have been read, skimmed, or skipped by interpreting eye movements. 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.

Our work builds on the previous research on connecting searcher examination patterns to user interest and document relevance. In particular, mouse cursor tracking has emerged as a useful tool for understanding searcher behavior. Previous research has shown the coordination between the searcher gaze position and mouse movement over the search results [33, 15, 21], as well as association between mouse cursor movements and searcher attention [13, 14, 31] and result preferences [23, 16, 22]. Most closely related to our work, Guo and Agichtein [16] found correlations between fine-grained interactions such as mouse cursor movements and document relevance, and showed that better relevance prediction can be achieved by incorporating these additional fine-grained behavioral signals compared to using dwell time information alone.

Recently, with the increasing popularity of touch-enabled devices, there have been a number of recent studies aiming to characterize and understand mobile search intent (e.g., [8]) and behavior (e.g., [26, 7]). Yet, the new types of interactions available on these touch-enabled devices are not yet sufficiently understood for optimizing the mobile searcher experience. In a recent position paper, Huang and Diriye argue for modeling of touch interactions for Web users [20]. However, to the best of our knowledge, there has been little empirical research on mining touch interactions for improving Web search. One exception is the recent work by Guo et al. [17], where the authors studied the problem of predicting search suc-



(a) (b)

Figure 1: (a) A searcher uses two fingers to pinch out (zoom in) a Web page on a smart phone touch screen; (b) A searcher uses one finger to swipe the page down to see the upper part of the page.

1	What was the best selling book (title and author) of 2011 in the US? How many copies of the book were sold in 2011 in the US?
2	Find three vegetarian restaurants near the Lenox Square.
3	What is the average temperature for the Dallas,SD for winter? Summer?"
4	How many pixels must be dead on a iPad 3 before Apple will replace it? Assume the tablet is still under warranty.
5	In what year did the USA experience its worst drought? What was the average precipitation in the country that year?
6	Is the band State Radio coming to Atlanta,GA within the next year? If not, when and where will they be playing closest?
7	Find the working hours of the Target store nearest to Emory University.

Table 1: Tasks descriptions

cess and satisfaction on a smart phone, and found that by modeling the amount of touch interactions more accurate predictions can be achieved.

In summary, our work extends previous research by offering the first in-depth study of modeling touch interactions for web search. We focus on the core problem of predicting document relevance for improving search result ranking. We present the first comprehensive analysis of touch interactions in search, with the aim of distinguishing the examination patterns of viewing a relevant or a non-relevant search result document.

3. USER STUDY

In this section, we introduce the two user studies conducted in the mobile and desktop settings respectively, and the interaction data collected in these studies.

3.1 Web Search on a Smart Phone

In the first user study, we capture the touch interactions during Web search on a smart phone, such as zooming and swiping (two examples are shown in Figure 1). To capture such interactions, we developed a modified version of the Chrome browser application for an Android phone, due to the popularity of this mobile operating system. All user actions, including clicks, touches, swiping, zooming, bookmarking page navigation and other interactions were logged with timestamps, and sent to a remote server via HTTP for further analysis. Similar instrumentation can be implemented for iPhones or Windows phones, as the types of touch interactions for viewing Web pages are largely shared across these platforms.

For each touch event, we additionally recorded the following information:

- Action type (namely, down, move and up)
- Number of touch points (i.e., number of fingers with contact with the screen)
- X and Y coordinates of the touch points
- Touch pressure, with a normalized value between 0.0 (no touch) and 1.0 (full force)
- Touch "size" with a normalized value between 0.0 (the smallest detectable size) to 1.0 (largest detectable size).

A gesture (a sequence of touch events starting with "down" and ends with "up"), may result in specific browser actions. For example, a pinch-out (or pinch-in) gesture with two fingers would result in zooming – an enlargement (or shrinkage) of a Web page (Figure 1 (a)) while a continued sequence of touch movements would result in swiping – a change of the viewing content ((Figure 1 (b)). Both zooming and swiping events are recorded in addition to the primitive touch events, with the former storing additional information about the new and old display scales, while the latter type providing additional information about the current vertical and horizontal page offsets.

Twenty-six subjects were recruited for this user study (fifteen females and eleven males). The mean age of participants was 21.7. All the subjects were undergraduate and graduate students from Emory University, with a variety of majors (e.g., psychology, economics, computer science, history and business), and all had some experience with Web search on a mobile device with a touch screen.

We designed the search tasks for this study to be representative of common Web search tasks on mobile devices. Table 1 reports the search tasks, together with the descriptions provided to the participants. As we can see, the tasks are a mixture of different topics, and highlight some geographical intents that are identified as significant part of mobile information needs in previous research [8, 26]. The search tasks were also designed to be difficult to solve with a search engine (i.e., the answer was not easily found on a single page). This is particularly valuable, since these more difficult and long-tailed search tasks are the main challenge for the state-of-the-art search engines. To distinguish oneself from the others, a search engine provider need to ensure that they do a good job on such search tasks.

Prior to starting the study, participants were given a tutorial on using the phone, including clicking, zooming, swiping, and changing the screen orientation. Next, a "warm-up" task was given to each participant to familiarize her with the procedure. Following the warm-up task, seven search tasks were given to each participant. To begin each task, the participants were presented a task description and an initial query to ensure starting with the same first set of result. After launching the initial query using the Google search engine, the participants were free to use the search engine however they chose in order to complete the task. Although participants generally agreed that the initial queries were reasonable for each task, they reformulated queries as needed.

Each time the participants navigated away from a non-search result document, or "landing page", they were asked (through a pop-up window) the degree to which the document satisfied the task. The rating was on a five point scale, with "1" indicating the page was "Bad", meaning it "did not satisfy the information need at all", and "5" indicating that the page was "Perfect", meaning it "com-

pletely satisfied the information need". The participants had the option to skip the rating, if they had not viewed the page.

In total, we obtained 393 unique Web page views, associated with explicit relevance judgments by the 26 participants. Our modified version of the browser application and the data used in this paper is available for the research community at http://ir.mathcs.emory.edu/data/SIGIR2013/.

3.2 Web Search on a Desktop PC

For the desktop setting, the data set we used was collected in a user study conducted by researchers at the University of Massachusetts [10]. The search behavior data of the participants was tracked, containing the URLs the searchers visited, the fine-grained interactions with the browsed pages, such as clicks, mouse cursor movements, and scrolling, the time-stamp of each page view and interaction was also recorded. The user population and search tasks in this second study are similar to that of the first user study we conducted to analyze mobile Web search behavior. The original dataset of the second user study is publicly available online [1] (additional information can be found along with the original dataset).

Similar to the first study, upon leaving a non-search page, the searchers were asked the degree to which their information need of the task was satisfied by the page on a five point scale. In total, 666 unique Web page views are provided, associated with explicit relevance judgments by the 30 participants.

4. TOUCH INTERACTION PATTERNS ON LANDING PAGES

In this section, we analyze the patterns of landing page examination and touch interactions. In general, we observed that the examination is typically a mixture of "reading" and "skimming" behavior, as found in previous research for the desktop setting [16]. In particular, "skimming" typically happens when the user is still searching the needed information, while "reading" typically happens when the user found the (seemingly) relevant information and is examining it more thoroughly.

The corresponding observable behavior, however, on a touch-enabled mobile device is different from a desktop personal computer due to the difference in the input and output device form factors. On a desktop personal computer, users scroll to change the content of the page (if the page is long enough to scroll), and may use the mouse cursor to focus attention (e.g., vertical coordination or text highlighting) [15, 16]. In contrast, on a touch-enabled mobile device, users use their fingers to swipe to change the displayed content, and stop touching the screen to read; they may pinch-out (zoom) to enlarge the web page if the content is not readable, or if they found something particularly interesting to examine more closely.

Based on analyzing the replays of touch interactions on hundreds of page views, we found that fast swiping behavior appears to indicate "skimming" behavior, while slow swiping/inactivity, and zooming suggest "reading" behavior or focused attention, where the former is an indicator of viewing non-relevant content, and the latter is an indicator of viewing relevant content. Figure 2 provides some illustrative examples of these interactions. Figure 2 (a) and (b) present two landing page examinations by a study participant, each of which is represented by the upper vertical coordinates of the smart phone view port (gray lines) as a result of swiping, and the corresponding view port scales (red lines) and zooming occurrences (red dots) over time. As we can see, both of the two page views have roughly 30 seconds dwell time, which would be considered relevant based on previous research [9, 11]. However, from the interaction patterns, we can tell that the two cases are actually different. As shown in Figure 2 (a), the user briefly paused, swiped down, briefly paused (possibly finding something seemingly relevant), and then swiped up again, without stopping to spend a significant amount of time reading any particular region of the page. Upon leaving the page, the user rated the page as "Bad". In contrast, in Figure 2 (b), the same user swiped down a little before a zoom-in, followed by a long period of inactivity (possibly indicating reading the content carefully). Upon leaving the page, the user rated the page as "Perfect". If we had only considered the dwell time information, we would predict the document of Figure 2 (a) as relevant due to the long dwell time. In contrast, in Figure 2 (b), the touch interactions data indicate that the user spends less time to find the needed information (i.e., less extensive swiping for changing the view port), while spending more time on "reading" or consuming the information (i.e., by observing the longer inactivity period following the zoom-in gesture), thus suggesting the document to be relevant.

The left part of Figure 2 (c) shows the corresponding heat map of the page examination shown in Figure 2 (b), where the darker the shade of green, the longer is the time that the user remained inactive or spent on reading the content that is presented in the corresponding view port. The right part of Figure 2 (c) shows the enlargement of the corresponding region where the user spent most of the time on after the zoom-in gesture. As we can see, the region indeed contains the needed information for the user (i.e., the worst drought year of the US history and the corresponding yearly average precipitation of that year, needed for completing Task 5 of Table 1). This further demonstrates that inactivity and zoom-in may be good indicators of user's reading behavior and, in turn, document relevance.

In addition to the amount of swiping, zooming and inactivity, a closer look at Figure 2 (a) and (b) reveals the potential of behavioral sequences in distinguishing viewing a relevant or a non-relevant document. For example, the periods of inactivity in Figure 2 (a) are followed by extensive swiping gestures, suggesting the dissatisfaction of the user after "reading"; in contrast, the inactivity in Figure 2 (b) is followed by the end of the page view, suggesting that the user was satisfied with the information she read about.

Next, we present our proposed approach of capturing these behavioral patterns and empirical results to demonstrate the utility of such modeling, which would provide further support to our observations about the "reading" and "skimming" behavior. However, to more comprehensively verify these assumptions, an eye-tracking study of searching on a mobile device (similar to the one reported in [3]) would be needed, suggesting a promising future extension of this work.

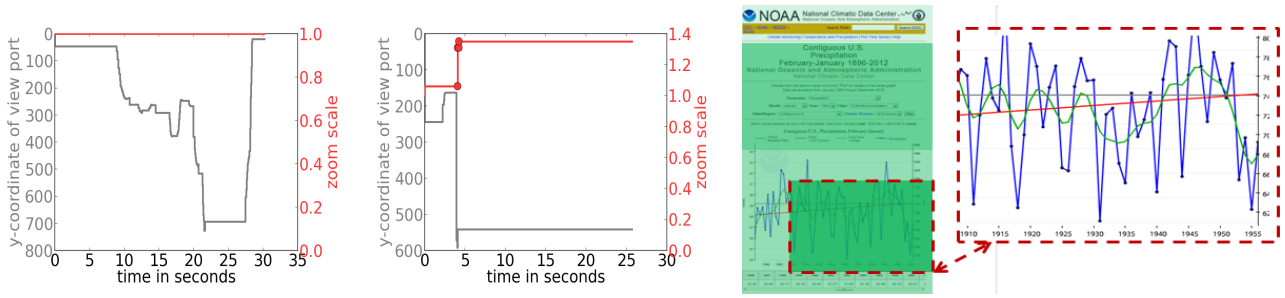
5. MODELING TOUCH INTERACTIONS

In this section, we describe our proposed *Mobile Touch Interaction (MTI)* features to capture the page examination patterns on a touch-enabled mobile device, with the goal of estimating document relevance. The full list of MTI features with brief descriptions is reported in Table 2, and expanded below.

5.1 Dwell Time

Dwell time, or document viewing time, has been extensively used as an indicator of document relevance – the longer the dwell time, the more likely the document is relevant. As typically done, dwell time is defined as the interval, in seconds, between the time the page is loaded and the time the searcher leaves the page. We use

[1] http://ciir.cs.umass.edu/~hfeild/downloads.html



(a) $dwell : 30s\ (rel : 1)$ (b) $dwell : 25s\ (rel : 5)$ (c)

Figure 2: Two example landing page examinations, represented by the vertical swipe coordinates (gray lines), the view port scales (red lines) and zooming occurrences (red dots) over time. The x-axis represents the time from page load in seconds, and the y-axes represent the vertical page offset (i.e., upper vertical coordinate of the corresponding view port) resulted from the swipe event and the zooming scale: (a) a user viewing an non-relevant document; (b) the same user viewing a relevant document; (c) the heat map for the page examination shown in (b) and the corresponding enlargement of the page region that received the most attention from the user (i.e., the long period of inactivity or "reading" after a zoom-in). The darker the green in (c) the longer the time the user remained inactive or spent on reading the content that is presented in the corresponding view port.

Group	*Feature Description*
Dwell	*dwell:* time of the page view in seconds
Zooming	*gestcnt:* number of touch gestures *gestfreq:* gestcnt/dwell *pressure:* average touch pressure, normalized to the range of 0 and 1 *touchsize:* average touch size, normalized to the range of 0 and 1
	zoomcnt: number of zooming gestures *zoomfreq:* zoomcnt / dwell *zoomdist:* accumulated zooming scale changes *zoomspeed:* zoomdist / dwell *zoommax:* maximum zooming scale
Swiping	*swipecnt:* number of vertical swipes *swipefreq:* swipecnt / dwell *swipedist:* total vertical swipe distance *swipespeed:* swipedist / dwell *swipemax:* maximum vertical swipe coordinate
Inactivity	*inactive_total:* accumulated inactive time *inactive_pct:* inactive_total / dwell *inactive_avg:* average inactive time *inactive_max:* maximum inactive time
Transitions	s_i-s_j*_cnt:* number of transitions from state s_i to observation s_j during the page view *transitions_cnt:* total number of observation transitions during the page view s_i-s_j*_prob:* s_i-s_j_cnt / transitions_cnt
User	*user_id:* the ID of the user who viewed the page

Table 2: Feature descriptions

dwell time both as a baseline to compare against and as a feature in our full model.

5.2 Gestures

As mentioned in Section 3, a gesture is a sequence of touches with x and y coordinates, starting with a "down" touch and ends with an "up" touch. A gesture may trigger effects such as zooming (pinching) and view port changing (swiping). The modeling of zooming and swiping are moved to their own feature groups. This group of features aim to capture the overall amount of efforts of the user (on a "syntactical" level) without knowing the "semantic" meaning of the gesture. Features include the number and frequency of the gestures, which were found to be negatively correlated with task-level searcher satisfaction in previous preliminary study of touch interactions [17]. In addition, we also capture the average pressure and touch sizes of the gestures, assuming they correlate with searcher's frustration level and could be indicative of document relevance. That is, viewing non-relevant document may cause users to become frustrated and thus change the touch pressure and size. Similar features such as mouse pressure were identified as frustration indicator in previous research [10].

5.3 Zooming

The zooming behavior typically occurs when users would like to take a closer look at the Web page, which happens relatively frequently on a mobile device with a limited screen size (especially, when users are viewing a Web page that is not optimized for mobile devices [17]), and infrequently happens on a desktop computer where screen sizes are typically much bigger. Zooming scale, as we noted in Section 4, can be indicative of degree of user interest. As pointed out by Huang and Diriye in their position paper [20], zooming may correspond to the mouse cursor movements in the desktop setting, such as text highlighting [34, 16] and hovering [23], which were found to be indicators of searcher interest. In this feature group, we capture the count, frequency of the zooming behavior, as well as the maximum zooming scale, the accumulated changes in zooming scale and the speed of the scale changing, following the representation of mouse cursor movements of the PCB model proposed by Guo and Agichtein [16].

5.4 Swiping

The swiping behavior on a touch screen results in the change of the view port similar to the scrolling behavior on a desktop computer. Due to the small screen sizes of mobile devices, users tend to swipe more often on such devices compared to scrolling on a desktop computer. Nevertheless, the change of view port on both of the modalities (as shown in Section 4 and previous research [16], respectively) tend to indicate that the user is still in the searching mode, thus extensive swiping when viewing a Web page is likely to suggest non-relevance. Following the previous research in the desktop setting [16], we focus on modeling the vertical swipe for this group, capturing the overall amount, the frequency, and speed

of the swiping behavior, as well as the overall swipe distance and its maximum in pixels.

5.5 Inactivity

As noted in Section 4, a period of inactivity on a touch-enabled mobile device seems to be a good indicator of "reading" behavior. This is in part due to the touch nature of the device – as a result, inactivity is typically needed to keep the screen still for careful reading (after adjustment of the view port through zooming and swiping). In contrast, inactivity was not found to be a good indicator of relevance in the desktop setting with a mouse and a keyboard as the primary input devices [16]. In the desktop setting, some users remain inactive when reading, while some other users use mouse cursor as the reading-aid [33, 15, 21, 34] and remain inactive when nothing interest them. The effectiveness of the inactivity may also attribute to the smaller screen sizes of the mobile devices – a long period of inactivity on a large PC screen may correspond to reading a specific area or skimming through the whole view port while a long period of inactivity on a small mobile device screen is less likely to correspond to skimming behavior (as it would take much less time). Empirical evidence would be provided in the next section to illustrate the difference of inactivity in these two modalities. For this feature group, we define inactivity as the time period of more than one second between a pair of consecutive touch behavior, and compute the total inactive time, and average and maximum time of inactive periods.

5.6 Sequence

Section 4 sheds some light on the potential of modeling behavioral sequence. For example, inactivity followed by the end of the page view is an indicator of viewing a relevant page while inactivity followed by swiping before ending may suggest that the document were seemingly relevant but actually not. The features in this group aim to capture these sequential patterns. We define the fine-grained interactions during a page view as a sequence of observed states $S = s_0 \rightarrow s_1 \rightarrow ... \rightarrow s_n$, where $s_i \in \{ZI, ZO, SD, SU, SS, IS, IM, IL\}$. The possible states are summarized in Table 3.

Given a sequence of states, we then count the occurrences of transitions $s_i \rightarrow s_j$ to characterize the behavioral sequence when viewing the page. We also normalize the transition counts by the total transition count of each page view to generate another set of transition features.

5.7 User Variability

Previous research has discovered significant variations in behavioral patterns among different users [15, 6]. For example, some users are found to be more "active" in examination than others [6]. Modeling such user variations was shown useful in improving estimating document relevance [16] and other Web search applications [15]. In this paper, we consider a simple modeling of user variation through incorporating participant IDs as additional features, which were shown effective in previous research [15].

6. CHARACTERIZING MTI FEATURES

We first characterize the "atomic" (non-sequential) MTI features, and evaluate their utilities as implicit relevance feedback through calculating their correlations with self-reported relevance judgements. We also compare MTI features with the corresponding fine-grained interaction features for the desktop setting (e.g., the PCB features presented in [16]), when available. Then, we characterize and evaluate the utility of the Markov transitions between features, for implicit feedback.

State	*Description*
START	start of the page view
ZI	zoom in
ZO	zoom out
SD	swipe down
SU	swipe up
SS	swipe still (i.e., horizontal swipe)
IS	short period of inactivity (1 to 5 secs)
IM	medium period of inactivity (5 to 20 secs)
IL	long period of inactivity (longer than 20 secs)
END	end of the page view

Table 3: Possible states used in sequence modeling.

Feature	*Mobile*	*PC*
dwell time	44.3 (36.2)	31.2 (29.9)
swipe/scroll distance	2808.8 (4007.6)	570.2 (1770.7)
swipe/scroll speed	66.9 (96.1)	17.8 (45.1)
swipe/scroll maximum	1352.6 (1915.4)	275.4 (835.4)
inactive percentage	0.59 (0.21)	0.57 (0.22)
total inactive time	28.0 (27.2)	19.1 (22.5)
average inactive time	4.3 (6.2)	4.7 (4.3)
maximum inactive time	14.8 (18.5)	9.2 (13.0)
zoom count	0.89 (2.71)	n/a
touch size	0.15 (0.06)	n/a
cursor ymax	n/a	595.5 (236.1)
cursor xspeed	n/a	108.5 (120.4)

Table 4: Means and standard deviations of representative fine-grained interaction features on the two different modalities

6.1 Comparing the Two Modalities

In Section 4 and Section 5, we compare the fine-grained interactions on the two modalities qualitatively. This section presents some statistics to demonstrate the similarities and differences in a more quantitative way.

First, we compare the overall statistics between the two modalities, including means and standard deviations of the features, which are summarized in Table 4. As we can see, some features are primarily used for the desktop PC setting, such as the wide variety of features that aim to capture mouse cursor movements [16]. Some representative examples include the maximum y-coordinate and the horizontal mouse moving speed. Similarly, some other features are primarily for touch-enabled mobile devices, such as zooming counts and touch sizes.

Despite the obvious differences in the types of common interactions, the two modalities also share some interaction features, such as dwell time, view port changing (i.e., swiping vs. scrolling) and periods of inactivity. While these shared interaction features can be considered as similarities of the two modalities, a closer look reveals noticeable differences. For example, dwell time and inactive time overall tend to be longer for the mobile setting, which may be explained by the slower searching and reading on this modality due to the relatively small screen size. Also likely due to partly the smaller screen size and partly the easiness in changing the view port through swiping, we can see that mobile users swipe much more and faster compared to scrolling on a desktop PC.

Second, we evaluate the MTI feature utility by computing the correlation with explicit relevance judgements provided by the users and compare with the feature utility for the counterpart fine-grained

interaction features for the desktop setting. The findings for representative features of the two modalities are summarized in Table 5.

As we can see, the dwell time information is consistently effective, exhibiting significant, but weak positive correlations around 0.17 for both modalities. Gesture count and frequency, in accordance with previous research [17], were found to be indicators of searcher dissatisfaction, both exhibiting negative correlations with the explicit relevance judgements. This makes sense as more efforts the user needs to make to find the information the more likely the user is dissatisfied and the document is more likely to be not relevant. In contrast, the touch size and touch pressure were found to be only weak indicators of document relevance, exhibiting insignificant correlations, which is in line with what has been reported previously [10] about pressure of using mouse on desktop computer. Somewhat surprising was the positive correlation (instead of negative correlation as assumed), which may be explained by the association of interaction speed with touch size/pressure – for example, when users swipe faster, the touch sizes of their fingers tend to be smaller compared to leaving their fingers on the screen to slowly swipe.

Similar to the scrolling behavior on a PC, the swiping behavior on touch-enabled mobile device also exhibits negative correlations with the document relevance – the faster the users swipe, the more likely the user was still searching and not satisfied by the document content. One subtle difference appear to be the decreased importance of knowing the actual number of pixels the user swipe. For example, the swipe speed on a mobile device exhibits only -0.165 correlation compared to the -0.212 correlation of its counterpart scroll speed on a PC, while the swipe and scroll frequency on the two modalities exhibit both around -0.210 negative correlations with document relevance. One explanation is the effects brought in by zooming behavior on the mobile device – with a zoomed-in view of the Web page, the same amount of swiping effort may only translate to smaller change in terms of pixels for the view port.

Zooming behavior, as expected, are mostly positively correlated with document relevance, but not in a significant way. The strongest correlation from this group is the insignificant positive correlation of 0.039 from the zoom distance features, which captures the accumulated amount of zooming scale changes. One explanation of the insignificance of zooming behavior is its relatively rare occurrences – with more and more Web sites provide mobile customized views of their page contents, zooming behavior may become less and less needed. Another factor that may contribute to the insignificance is the noise brought in by zooming on a mobile unfriendly non-relevant document, in which case, similar to the perception bias in click signals that are found in previous research [16], zooming is just a necessity for being able to view the seemingly relevant document and is not sufficient to reveal the actual relevance by itself.

The features of inactivity, which aim to capture the "reading" behavior on a touch-enabled mobile device, exhibit significant and positive correlations with document relevance. This accords with our intuition that inactivity on such device is indicative of "reading" and, in turn, indicative of relevance. Interestingly, while the overall statistics of inactivity features for the two modalities appear similar (in Table 4), the correlations for the same set of features on the desktop PC setting are mostly insignificant (except for the total inactive time, which may borrow strength from the overall dwell time as it has strongest correlation with the overall dwell time among all the features in the group). The maximum inactive time is among the strongest features in the group for mobile, exhibiting a significant positive correlation of 0.284. Its counterpart for the desktop PC, however, only has an insignificant correlation of 0.059. This observation supports our assumption that inactivity on a desktop PC (with bigger screen size, and mouse and keyboard as primary input devices) does not necessarily correspond to the reading behavior.

Feature	*Mobile*	*PC*
dwell time	0.171*	0.167*
gesture count	-0.042	n/a
gesture frequency	-0.201*	n/a
touch size	0.032	n/a
touch pressure	0.020	n/a
swipe/scroll count	-0.043	-0.008
swipe/scroll frequency	-0.210*	-0.206*
swipe/scroll distance	-0.065	-0.092*
swipe/scroll maximum	-0.052	-0.025
swipe/scroll speed	-0.165*	-0.212*
zoom count	0.026	n/a
zoom frequency	-0.008	n/a
zoom distance	0.039	n/a
zoom maximum	0.021	n/a
inactive percentage	0.244*	-0.044
total inactive time	0.232*	0.124*
average inactive time	0.276*	0.016
maximum inactive time	0.284*	0.059
cursor count	n/a	0.164*
cursor frequency	n/a	-0.082*
cursor ymax	n/a	0.243*
cursor xspeed	n/a	-0.143*

Table 5: Pearson's correlation between individual implicit feedback feature and the actual explicit judgments of document relevance (* indicates $< .05$ statistical significance).

Even though inactivity features are not good relevance indicators for the desktop PC setting, we found that its exclusive features of mouse cursor movements instead provide substantial utility as implicit relevance feedback. In accordance with the previous research [16], we found that the amount of mouse cursor movements (e.g., cursor ymax) indicates user's interest level of the document, exhibiting significant positive correlation with document relevance, while faster mouse movements seems to suggest "skimming" behavior, exhibiting significant negative correlation with document relevance.

In summary, we have identified interesting similarities and differences of the fine-grained interactions between the two modalities, both in terms of their overall characteristics and utility as implicit relevance feedback. Next, we will characterize and evaluate the utility of modeling behavioral sequence as Markov transitions.

6.2 Evaluating Markov Transitions

Figure 3 summarizes the Markov state transitions for viewing a document on a touch-enabled mobile device, estimated from our dataset. In the figure, the nodes represent the possible states, and the directed edges represent the transitions between the two connected states with the width of the edge signifying the likelihood of the corresponding transitions. The figure reveals some interesting sequential patterns. For example, the START state is most likely to be followed by IM (Inactivity Medium), suggesting users tend to remain inactive for a intermediate period of time in the beginning to get themselves familiar with the page content in this initial view port before deciding the next step (e.g., swipe, zoom or end). The heaviest outgoing edge for IM, interestingly, is SD (i.e., swipe down), which is much thicker compared to the outgoing edge to END, suggesting that in most cases, users would not find the rele-

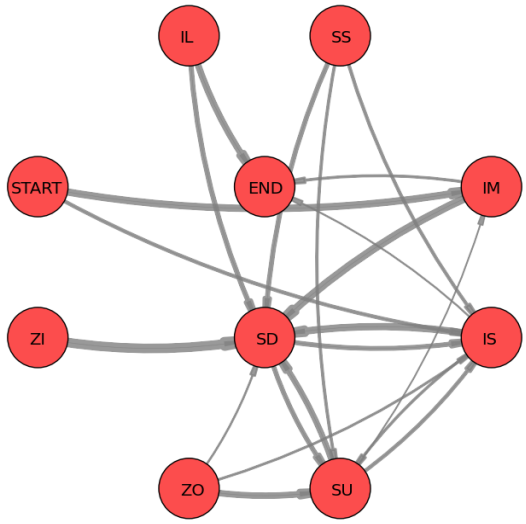


Figure 3: Markov state transitions for interactions of viewing a document on a touch-enabled mobile device. The transition probabilities are indicated by the line width; transitions with probabilities of less than 0.1 are not shown.

vant information in the very first view port (neither would exit the page view immediately) and tend to swipe down to view further. As a matter of fact, SD is the largest receiving state, with many dominant edges pointing from other states. This may also be explained by the relatively small screen size and the easiness of swiping on a touch-enabled mobile device – it is not hard to imagine, users often need to swipe down to continue reading no matter whether they have found something promising (e.g., transitioning from state IL or ZI), or have not found something interesting (e.g., transitioning from IS), or would like to re-examine (e.g., transitioning from SU). Another two interesting prominent transitions are ZO to SU and IL to END, where the former signifies the typical pattern of going back (e.g., to revisit previously seen content) while the latter signifies a typical "happy ending" where the user ends the page view after a long period of "reading".

Table 6 highlights example state transitions that are particularly indicative of document relevance, with each row reporting the likelihoods of a particular transition under the relevant and non-relevant Markov models and the corresponding likelihood ratio. Self-reported judgments greater or equal to 3 ("good") were considered "relevant" while "Fair" and "Bad" were considered "non-relevant" in this analysis. IL-END is one of the most interesting transitions: the chance of seeing such "happy ending" is 1.6 times more likely to happen for viewing a relevant document as compared to reading a non-relevant document. Its counterpart IS-END, in contrast, signifies an opposite trend, exhibiting a 0.558 likelihood ratio. Interestingly, a swipe down after IL (i.e., IL-SD) turns out to be a strong negative indicator of relevance. This makes sense as a continued searching after a long read is likely to suggest that the user was not (completely) satisfied. Another interesting example is START-IS, which is 1.4 times more likely to happen when viewing a relevant document. This may correspond to cases where users found something interesting or promising and quickly react thus shorten the period of time for the initial orienteering. Regarding swiping, SS-SU (i.e., horizontal swipe followed by a upward swipe), which signifies the re-examination, also appears to be a strong indicator of non-relevance. Zooming-related transition features, while having relative big likelihood ratios (e.g., SU-ZO), are not necessarily good predictors as they occur rarely.

Transition	*Relevant Lik.*	*Non-relevant Lik.*	*Lik. Ratio*
IL-SD	0.255	0.488	0.522
SS-SU	0.145	0.265	0.546
IS-END	0.095	0.170	0.558
ZO-IM	0.034	0.059	0.586
SU-IS	0.311	0.278	1.118
SD-ZO	0.003	0.002	1.414
IM-END	0.267	0.187	1.429
SU-ZO	0.049	0.030	1.609
START-IS	0.372	0.266	1.400
SS-IS	0.362	0.230	1.575
IL-END	0.638	0.395	1.615

Table 6: Relevant and Non-relevant Likelihoods (*Lik.*), Likelihood Ratio (*Lik. Ratio*) for Example Markov State Transitions

In summary, we have characterized the common behavioral sequences using Markov transition probabilities, and demonstrated the potential of modeling these transitions for estimating document relevance.

7. PREDICTING DOCUMENT RELEVANCE

Next, we describe our experiments on predicting document relevance, and re-ranking documents based on the predicted relevance scores.

7.1 Prediction Model

We treat the relevance prediction as a regression problem, and use a learning algorithm based on an ensemble of regression trees, which is representative of what major search engines tend to use for training ranking algorithms. In particular, similar to the previous research [16], we used Bagging (Bootstrap Aggregation) [4], which outputs a numerical outcome as prediction by aggregating multiple versions of a single predictor through plurality voting. The multiple versions of the single predictor are formed by making bootstrap replicates of the learning set, and using these as new learning sets. The single predictor, or weak learner, used was the C4.5 regression tree, which splits the tree at each node by choosing the attribute that is most effective in terms of normalized information gain. The advantages of this non-linear ensemble regressor include its expressiveness, which can capture complex interactions among a variety of features, resistance to over-fitting, and the relative computational efficiency compared to other non-linear learning algorithms such as neural networks.

7.2 Evaluation Metric

We evaluated the performance of predicting document relevance by computing the standard information retrieval measure of NDCG (Normalized Discounted Cumulative Gain) for the corresponding re-ranking of the search results. Specifically, given a ranked list of documents for a search task, $NDCG_k$ [24] measures the quality of a ranked list at position k, as follows:

$$NDCG_k = \frac{DCG_k}{IDCG_k}, DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

where $IDCG_k$ is the DCG_k value of the ideal ranking with respect to the actual document relevance, and the rel_i is the relevance judgment, which is at a five point scale. DCG_k aims to

penalize the ranked list with highly relevant documents appearing at lower positions, with the graded relevance value reduced proportionally to the position of the result in the list. $NDCG_k$ of 1.0 indicates a perfect ranking that is identical to $IDCG_k$, with smaller values indicating worse rankings. We first compute $NDCG_k$ for each individual search task, and then average the scores into one $NDCG_k$ to summarize the quality of the ranked list provided by each method. We evaluate this score for various k values and highlight $k = 1, 3, 10$, which are the values reported to be used by commercial search engines.

7.3 Methods Compared

We consider the following methods for estimating document relevance, including three variants of the MTI model, and two baseline models using dwell time information and the original search result rank, respectively. All the models are trained with Bagging with regression trees (Section 7.1).

ResultRank: This model uses the original search result rank as the feature to predict document relevance, which encodes the belief in relevance that the search engine holds and is typically derived from thousands of ranking signals. Presumably, the smaller the rank value (i.e., the higher the document was ranked), the more relevant the document is likely to be. However, if the search engine fails in accurately estimating the relevance, the rank would become uninformative. For the viewed documents in the search trail that were not ranked in the SERP, the rank of the landing page (i.e., the origin of the search trail the document was on) is used.

DwellTime: This model uses dwell time information as the feature to predict document relevance, which has been widely used in previous research as implicit relevance feedback [36, 16], and successfully applied to addressing a variety of real-world Web search applications [35, 18, 10, 19].

MTI_P: A variant of the MTI model that adapts the feature representations proposed by the PCB model [16] – for example, MTI_P incorporates features for swiping behavior that are similar to the PCB scrolling features, features for zooming and raw gestures similar to PCB features for mouse cursor movements that aim to capture user degree of interest, efforts and engagements. In other words, the features from the Dwell Time, Gesture, Swiping and Zooming groups that are described in Section 5 are used to train this model.

MTI_PI: A variant of the MTI model that adapts the feature representations proposed by the PCB model and adds the newly proposed inactivity features. In other words, the features from Gesture, Swiping, Zooming, and Inactivity groups that are described in Section 5 are used.

MTI_PIM: A variant of the MTI model that adapts the feature representations proposed by the PCB model and adds the newly proposed inactivity and Markov transition features that are described in Section 5.

MTI_PIM_U: A variant of the MTI model that incorporates User Variability features (Section 5) in addition to all MTI_PIM features, aiming to model variations in user behavior.

7.4 Results

We now report our results on re-ranking the viewed documents using the estimated relevance. For training and testing, we use 10-fold cross-validation, with 10 randomized experimental runs. We report $NDCG_k$, averaged over all the search tasks across different users.

The overall experimental results are summarized in Figure 4 across multiple K values. As we can see, all of the four variants of the MTI model substantially and significantly outperform the two baseline models across all K values. The relatively poor performance of the ResultRank baseline is expected due to the difficulty of the assigned search tasks, which was also observed in previous research [16]. Among the MTI variants, the simplest version in MTI_P already results in significant gains, on top of which, adding inactivity features MTI_P results in another substantial lift. Adding Markov transition features (MTI_PIM) and user features (MTI_PIM_U) achieve further incremental gains with smaller yet statistically significant margins. Interestingly, all the models based on implicit feedback seem to perform particularly well at the top positions (e.g., $K < 3$). One explanation is that the search interaction patterns are more consistent for viewing very relevant documents (e.g., with "perfect" rating on the five point scale), thus resulting in more accurate estimation. The improvements over the stronger DwellTime baseline are further summarized in Table 7 for $K = 1, 3, 10$. As we can see, the improvements are around 20% and are substantially higher for top positions, which is desirable as these are the ranking positions that matters most to Web searchers and are of particular interest to search engines.

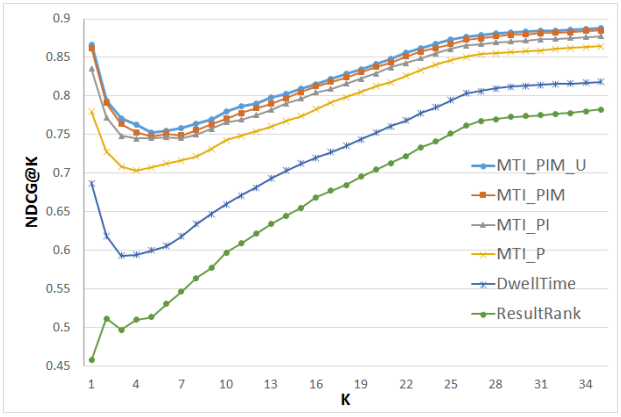


Figure 4: NDCG at K (K=1, 2, ..., 35) for the different variants of the MTI model compared to the dwell time and result rank baseline models. The differences between the different methods are statistically significant at p < .05 level.

Method	*K=1*	*K=3*	*K=10*
MTI_PIM_U	0.866 (+26%)	0.771 (+30%)	0.780 (+18%)
MTI_PIM	0.862 (+25%)	0.764 (+29%)	0.771 (+17%)
MTI_PI	0.835 (+22%)	0.749 (+26%)	0.766 (+16%)
MTI_P	0.780 (+14%)	0.709 (+19%)	0.743 (+13%)
DwellTime	0.687	0.593	0.659

Table 7: NDCG at K (K=1, 3, 10) for the different variants of the MTI model compared to the dwell time baseline model. The differences between the different methods are statistically significant at p < .05 level.

8. CONCLUSIONS

In this paper we introduced a new model for representing Web search interactions on touch-enabled mobile devices, which captures not only dwell time information, but also fine-grained user interactions, such as zooming, swiping and inactivity. To our knowledge, our proposed Mobile Touch Interaction (MTI) model is the first successful attempt to exploit such "low-level" behavioral signals to identify the basic patterns of "reading" and "scanning" behavior, as well as more complex combinations of these on a touch-

enabled mobile device, represented by expressive features to capture these examination patterns automatically. We also presented in-depth comparative analyses of the fine-grained interaction modeling for the mobile and the desktop Web search settings, identifying interesting similarities and differences.

Our experimental results show that the proposed behavioral signals indeed correlate with searchers' explicit judgments of document relevance, and provide additional valuable information beyond dwell time alone. For example, we found that periods of inactivity, as being indicative of "reading" behavior on a touch-enabled mobile device, are among the most predictive signals of document relevance on this new Web search modality, but not for the desktop PC setting. We also found that swiping behavior, being similar to the scrolling behavior in their functionality of changing view port, are indicative of information searching or "skimming" and can be used as negative implicit feedback. Furthermore, we found that incorporating the behavioral sequence and user information can additionally improve document relevance estimation. In combination, these signals enable MTI to exhibit substantial improvements on relevance prediction accuracy, consequently improving document ranking by up to 30% compared to using dwell time data alone.

In summary, we have laid the groundwork for exploiting fine-grained search behavior for improving relevance estimation and search result ranking for touch-enabled devices, providing a significant advance towards improving the mobile search experience.

9. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant IIS-1018321. The authors also thank Henry Feild for sharing the data of the second user study for the desktop PC setting and the reviewers for valuable comments to help improve the paper.

10. REFERENCES

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

[2] J.-w. Ahn, P. Brusilovsky, D. He, J. Grady, and Q. Li. Personalized web exploration with task models. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 1–10, 2008.

[3] R. Biedert, A. Dengel, G. Buscher, and A. Vartan. Reading and estimating gaze on smart phones. ETRA '12, pages 385–388, New York, NY, USA, 2012. ACM.

[4] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.

[5] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. SIGIR '08, pages 387–394, 2008.

[6] G. Buscher, R. W. White, S. Dumais, and J. Huang. Large-scale analysis of individual and task differences in search result page examination strategies. WSDM '12, pages 373–382, 2012.

[7] K. Church and N. Oliver. Understanding mobile web and mobile search use in today's dynamic mobile landscape. In *Proc. of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 2011.

[8] K. Church and B. Smyth. Understanding the intent behind mobile information needs. In *Proceedings of the 14th international conference on Intelligent user interfaces*, IUI '09, pages 247–256, 2009.

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

[10] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. SIGIR '10, pages 34–41.

[11] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), 2005.

[12] G. Golovchinsky, M. N. Price, and B. N. Schilit. From reading to retrieval: freeform ink annotations as queries. SIGIR '99, pages 19–25.

[13] Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. SIGIR '08, pages 707–708, 2008.

[14] Q. Guo and E. Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. SIGIR '10, pages 130–137, 2010.

[15] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. CHI EA '10, pages 3601–3606, 2010.

[16] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. WWW '12, 2012.

[17] Q. Guo, S. Yuan, and E. Agichtein. Detecting success in mobile search from interaction. SIGIR '11, 2011.

[18] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: user behavior as a predictor of a successful search. WSDM '10, pages 221–230.

[19] A. Hassan, Y. Song, and L.-w. He. A task level user satisfaction metric and its application on improving relevance estimation. CIKM '11, 2011.

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

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

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

[23] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. CHI '11, pages 1225–1234, 2011.

[24] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. SIGIR '00, pages 41–48, 2000.

[25] 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), 2007.

[26] 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. WWW '09, pages 801–810, 2009.

[27] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. SIGIR '01, pages 408–409, 2001.

[28] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. SIGIR '04, pages 377–384, 2004.

[29] M. Melucci and R. W. White. Discovering hidden contextual factors for implicit feedback. In *CIR'07*, pages –1–1, 2007.

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

[31] V. Navalpakkam and E. Churchill. Mouse tracking: measuring and predicting users' experience of web-based content. CHI '12, pages 2963–2972, 2012.

[32] R. Rafter and B. Smyth. Passive profiling from server logs in an online recruitment environment. In *Proc. of ITWP*, 2001.

[33] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. CHI EA '08, pages 2997–3002, 2008.

[34] R. W. White and G. Buscher. Text selections as implicit relevance feedback. SIGIR '12, pages 1151–1152, 2012.

[35] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. CIKM '09, 2009.

[36] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. CIKM '06, pages 297–306, 2006.