Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and other Post-click Searcher Behavior

Qi Guo
Mathematics & Computer Science Department
Emory University
qguo3@emory.edu

Eugene Agichtein
Mathematics & Computer Science Department
Emory University
eugene@mathcs.emory.edu

ABSTRACT

Result clickthrough statistics and dwell time on clicked results have been shown valuable for inferring search result relevance, but the interpretation of these signals can vary substantially for different tasks and users. This paper shows that that post-click searcher behavior, such as cursor movement and scrolling, provides additional clues for better estimating document relevance. To this end, we identify patterns of examination and interaction behavior that correspond to viewing a relevant or non-relevant document, and design a new Post-Click Behavior (PCB) model to capture these patterns. To our knowledge, PCB is the first to successfully incorporate *post-click* searcher interactions such as cursor movements and scrolling on a landing page for estimating document relevance. We evaluate PCB on a dataset collected from a controlled user study that contains interactions gathered from hundreds of unique queries, result clicks, and page examinations. The experimental results show that PCB is significantly more effective than using page dwell time information alone, both for estimating the explicit judgments of each user, and for re-ranking the results using the estimated relevance.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Design, Experimentation, Human Factors

Keywords

post-click search behavior, relevance estimation

1. INTRODUCTION

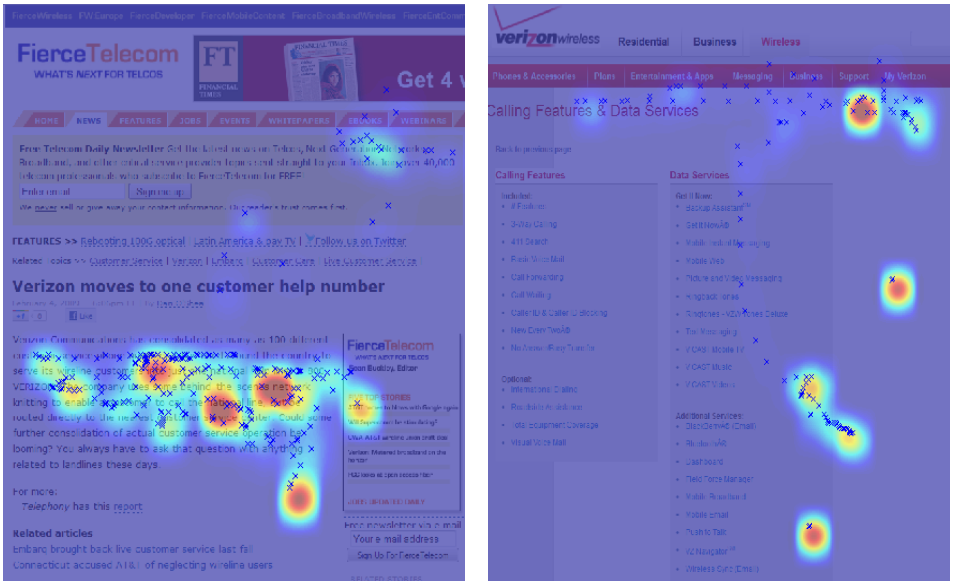
Estimating document relevance is at the core of information retrieval ranking and evaluation. Unfortunately, this task is notoriously difficult: even the notion of relevance itself varies for different tasks and users (e.g., [5, 34]). In this paper, we argue that *post-click search behavior* provides essential evidence for estimating the "intrinsic" page relevance for a search task.

While previous research has made great use of result clickthrough data (e.g., [3, 25, 13, 12]), the usefulness of clickthrough statistics is limited by a number of *presentation biases*, which strongly influence user click behavior. One of the most significant limitations of clickthrough data, is that clicks are based primarily on a document's *perceived* relevance [12], where a searcher guesses the page's relevance based on a short summary generated by the search engine. However, the "perceived" relevance may be inconsistent with the actual "intrinsic" relevance [12], where a searcher clicks on a result only to find out that it is not relevant. To address this problem, page *dwell time* (the time spent examining the result document) has been proposed as a measure of intrinsic document relevance [29, 10, 15, 27, 35, 8]. The main intuition is that "short" dwell time (typically, considered to be less than 30 seconds), indicates that a document is non-relevant. The most heavily studied scenario is that of a "bounce back", which happens when the searcher returned to the Search Engine Result Page (SERP) shortly after she clicked on a result, indicating low result relevance [33]. This heuristic and resulting metrics have been successfully adapted by the major search engines, and have undoubtedly improved search quality by detecting non-relevant or even detrimental results.

Unfortunately, the converse of the short dwell time rule is not true: a "long" page dwell time does not necessarily imply result relevance. In fact, a most frustrating scenario is when a searcher spends a long time searching for relevant information on a seemingly promising page that she clicked, but fails to find the needed information. Such a document is clearly non-relevant (and arguably one of the most detrimental to the searcher experience). Yet, based on dwell time alone, this document would be considered highly relevant, and remain high in the search ranking to frustrate future searchers.

To address this problem, we propose to use *post-click searcher behavior* to more precisely analyze how the searchers spend their time on the landing pages and the subsequently viewed documents, which would in turn allow for more accurate estimation of intrinsic document relevance. As an illustration, Figures 1(a-b) show the searchers' cursor movement on clicked result pages for the task of finding the phone number of the Verizon Wireless helpline for Massachusetts, where the user spends approximately 30 seconds examining each of the pages (i.e., both pages have almost equal dwell time). The color intensity in the figures indicates the amount of time the mouse cursor spent over the corresponding document regions, with the exact cursor coordinates indicated by the small crosses. The differences in the examination of a relevant page (Figure 1(a)) and a non-relevant page (Figure 1(b)) are striking. For the former, the searcher was carefully "reading" the text and using the mouse as a reading aid (examination of the page reveals that the answer of the search task indeed lies in the highlighted paragraph), while for the latter, the searcher appears to be "skimming" or "scanning" the page, without finding relevant information worth careful reading (indeed, the answer was not on the page). This example illustrates our underlying hypothesis: that page dwell time alone is not sufficient to distinguish between relevant and non-relevant pages, but post-click searcher behavior can provide the necessary additional evidence to distinguish the two.

WWW 2012,, April 16–20, 2012, Lyon, France
ACM 978-1-4503-1229-5/12/04.



(a) relevant (dwell time: 30s) *(b) non-relevant (dwell time: 30s)*

Figure 1: Cursor-based "Reading" examination heatmap of a relevant document (a) compared to "Scanning" of a non-relevant document (b), both with equal dwell time (30 seconds).

Specifically, we hypothesize that searcher interactions on landing pages such as cursor movements and scrolling can help more accurately interpret searcher viewing behavior, in turn, improve relevance estimation. That is, like eye movements, such interactions can reflect searcher attention. These interactions can be captured with Javascript code that is embedded in a browser Add-on (e.g., a search engine toolbar). This would allow estimating whether some parts of the landing page captured the searcher's attention and provide additional clues about the document relevance.

To test this hypothesis, we first identify patterns of examination and interaction behavior that correspond to viewing a relevant or non-relevant document (Section 3), and develop a novel model of inferring document relevance that incorporates rich Post-Click Behavior (PCB) such as cursor movements and scrolling that could capture these patterns (Section 4). The model is operationalized by converting these interactions into features, which can then be used as input to machine learning algorithms for tasks such as estimating personalized and aggregate document relevance, and improving result ranking (Section 5). In summary, our contributions include:

- Characterizing patterns of examination and interaction behavior that correspond to viewing a relevant or non-relevant document (Section 3).
- PCB, a novel model of relevance estimation that captures post-click behavior (Section 5).
- Empirical evidence that PCB is more effective than using dwell time information alone, both for estimating the explicit judgments of each user, as well as for ranking the 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 first research done in the area of implicit feed-back in information retrieval was that of Morita and Shinoda [29]. They conducted a study where participants was 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 [10] and [15]. Furthermore, Morita and Shinoda found only a very 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 [26] 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 [27] 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 [35] 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 [30] 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 clearly contains some relevance signal, numerous previous studies has found almost as many different interpretations of it with no clear consensus of the relationship to relevance of the document.

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 [10], click-through [15, 25], and exit type for a Web page [15]. On the other hand, mouse movements and

mouse clicks while viewing a document do appear to provide some correlation to user interest [10]. Furthermore, user behavior on the SERP, when combined with page dwell-time and session level information, can significantly improve result ranking in the aggregate (e.g., [2]), and can be further improved by personalizing these measures (e.g., [28]).

Other previous efforts focused on modeling more explicit user interactions on the page. Golovchinsky et al. [16] 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. [4] followed a similar idea but used the concept of a personal notebook where users could paste text passages worth remembering. On the basis of the text passages they built up term-based task profiles which were then used for re-ranking search result lists. Compared to a baseline ranking function not considering any feedback, the task-profile-based ranking performed significantly better. The previous two approaches both need more or less explicit and therefore rare user interactions (i.e., annotating, copying and pasting) to work properly. Buscher et al. [7] only rely on implicit data and determine which parts of a document have been read, skimmed, or skipped by interpreting eye movements. Read and skimmed parts were taken as relevant while skipped document parts were ignored. They report considerable improvements concerning re-ranking of result lists when including gaze-based feedback on the segment level compared to relevance feedback on the document level. Gyllstrom and Soules [20] follow a similar idea, but consider all text that has been visible on the screen for building up term-based task profiles. They use such profiles for task-based indexing of documents on the desktop and show that re-finding documents that way is more effective compared to simple desktop search.

Our work builds on previous research on connecting searcher examination patterns to user interest and document relevance. In particular, eye tracking studies have been helpful for understanding common patterns in search result examination (e.g., [25, 11]). To operationalize these insights, we exploit the coordination between the searcher gaze position and mouse movement over the search results, shown previously in references [31, 32, 19], as well as association between cursor movements over the search result pages and searcher intent [17, 18] and interests [23].

Most closely related to our work, Huang and White [23] found correlations between cursor hovering over some of the results on the Search Engine Result Page (SERP) and result relevance. Complementary to previous efforts, this paper is the first to analyze the examination patterns, and relevance, from *post-click* searcher behavior such as cursor movements on landing pages and subsequently viewed documents, and the first to develop a predictive model, PCB, that captures these patterns. As the rest of the paper demonstrates, PCB can provide significant improvements for estimating document relevance and consequently for improving search result rankings.

3. LANDING PAGE EXAMINATION

In this section, we describe the patterns of landing page examination and interaction that we identified. Overall, we observe two basic patterns of viewing, namely, "reading" and "scanning" (as illustrated in Figure 1). "Reading" tends to occur when relevant information (or seemingly relevant information) is found, and the searcher is consuming (or further verifying) the information. In contrast, "scanning" typically indicates that the searcher has not yet found the relevant information and is still in the process of searching. Typically, the viewing behavior is some mixture of these two basic components. Sometimes, the mixture is dominated by one of the two types. For example, Figure 1(a) is dominated by the "reading" behavior (suggested by the cursor heatmap overlaid on top of the answer of the search task [32]) while Figure 1(b) is dominated by the "scanning" behavior (suggested by the more vertically spread-out cursor distribution on the right of the screen on this page that does not contain the relevant information [32]).

At other times, the viewing behavior is more complex, especially, when the relevance of the document is not obvious (e.g., the document is long and contains a mix of relevant and irrelevant information). Figure 2(a) shows an example of viewing behavior on a long relevant landing page, while Figure 2(b) shows an example of viewing behavior on a irrelevant long page. The search task for both of the pages were "How many pixels must be dead on a MacBook before Apple will replace the laptop? Assume the laptop is still under warranty." and the dwell time on the two documents were roughly 70 seconds and 80 seconds, respectively. The two documents in this example are both from Apple's support forum and are much longer than the example documents in Figure 1. In such a case, using dwell time alone would suggest that the two are both relevant, and moreover the second document is slightly more relevant. However, the document examination patterns suggest that the two are quite different. The cursor movements on the first document are more focused on the left side, with clustering around the top posts, which suggests "reading" behavior (indeed, closer examination shows that the top posts contain relevant information). However, we do see that the pattern is more complex than what we have seen in Figure 1(a) – the cursor positions are more spread-out vertically and we do not observe extensive horizontal cursor movements. In contrast, on the non-relevant document in Figure 2(b), the searcher keeps the mouse still and scrolls – which indicates "scanning" behavior. Interestingly, here too the cursor positions are clustered on the left (indicating slowing down of cursor movements) over the top post, which may indicate "reading" behavior. Examination reveals that the page indeed contains on-topic information, that initially seems relevant, but does not contain the needed answer. Thus, in this example, the initial "reading" behavior is followed by a series of "scanning" before the searcher exits the page without finding her answer.

As these examples indicate, in addition to the variety in combining the two basic viewing patterns, the corresponding behavioral signals of these two patterns vary too. For example, when reading, searchers might keep the cursor still or use the cursor to mark the text (eye and cursor are coordinated only vertically) without actively moving the cursor horizontally (eye and cursor are coordinated both vertically and horizontally). Nevertheless, in both cases, the searcher tend to slow down the cursor movements or scrolling, especially, in the vertical direction. As for the "scanning" behavior, in contrast, the searcher tend to move the cursor and/or scroll faster as they are searching for the relevant information or sometimes also keeps the mouse still.

In summary, after examining many viewing sessions, we observe common patterns across all the post-click page examinations that correlate with document relevance, listed below:

- *Periods of horizontal reading indicate relevance*: The searchers are more likely to slow down and move mouse horizontally to read when the document is relevant, as opposed to only quickly scanning the document when it is non-relevant.

- *Focused attention indicates relevance*: searchers tend to focus on only one or a small number of areas for a relevant doc-

(a) relevant (dwell time: 70s) *(b) non-relevant (dwell time: 80s)*

Figure 2: An example of "Reading" a relevant long document (a) vs. "Scanning" a non-relevant long document (b).

ument, while distribute time more evenly throughout a non-relevant document. In contrast to the previous case, here, we mean if searchers exhibit "reading" behavior (e.g., slowing down) *multiple times*, it is more likely that she still did not find the right information that satisfies her need – for more complex task, or documents with denser text, such "reading" behavior are likely to be triggered.

- *Left-prevalence*: On relevant pages, searchers tend to keep the cursor towards the left half of the screen, where typically most of the content laid out on a Web page, to help reading or prepare to click on a link for more details.

- *"Scanning" followed by "reading" indicates relevance*: Often, a "scanning" behavior followed by focused, careful "reading" behavior at the end of the examination indicates relevance, while "reading" behavior in the beginning followed by "scanning"(i.e., the searcher is still not yet satisfied with what he or she has found so far), indicates non-relevance of the document.

- *"Skipping" indicates non-relevance*: Periods of reading or scanning, interspersed with periods of quick scrolling ("skipping" document sections) indicates lower relevance than continuous examination – searchers may become impatient, and accelerate "scanning" to an even faster pace.

These patterns can be captured by post-click behavioral signals such as sequences of cursor and scroll speeds and ranges. In the next section, we describe the features we designed to model these examination patterns, which can subsequently be used to better estimate document relevance.

4. POST-CLICK BEHAVIOR FEATURES

In this section, we describe our proposed *Post-Click Behavior (PCB)* features to capture the the page examination patterns that could indicate a difference in document relevance. In addition, we also include dwell time, task-level information (which is also shown to be useful in estimating document relevance in recent studies [22]), and the original search engine result ranking, as features in our model. The full list of PCB features and their brief descriptions are reported in Table 1, and expanded below.

4.1 Dwell Time

Dwell time, or document viewing time, has been previously used as the basic indicator of document relevance. 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 dwell time both as a baseline to compare against and as a feature in our full model.

4.2 Result Rank

The rank of search result is the belief in its relevance that the search engine holds, which is typically obtained by combining hundreds 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 document relevance, the rank would become uninformative. For the viewed documents in the search trail that were not ranked in a search engine result page, the rank of the landing page (i.e., the origin of the search trail the document was on) is used [1].

4.3 Cursor Movements

As suggested in the previous section, characteristics of cursor movements such as speed and range could indicate the searcher's reading behavior, and consequently the relevance of the document. For example, low speeds may indicate that the searcher was carefully "reading", while a long vertical range may indicate that the searcher found the document relevant and was willing to explore. We measure the number and frequency of the cursor movements, distance, speed, and the range the mouse cursor travels in pixels (both overall, and its horizontal and vertical components), as well

[1]The ranks are set to be 11 for a small portion of the documents whose ranking information is missing or cannot be recovered.

Group (30)	*Feature*	ρ
Dwell (1)	*dwell:* time of the page view in seconds	0.167**
Rank (1)	*rank:* the rank of the document or the rank of the origin (i.e., the landing page) of the search trail that the document is on if its rank is not available	-0.073
Cursor (14)	*cursorcnt:* num. of cursor movements	0.164**
	cursorfreq: cursorcnt/dwell	-0.082*
	dist: total overall distance the cursor traveled in pixels	-0.137**
	xdist: total distance the cursor traveled horizontally in pixels	0.101**
	ydist: total distance the cursor traveled horizontally in pixels	0.172**
	speed: dist/dwell	-0.101**
	xspeed: xdist/dwell	-0.143**
	yspeed: ydist/dwell	-0.124**
	xmin: minimal x coordinate	0.112**
	ymin: minimal y coordinatee	0.093*
	xmax: maximal x coordinate	0.067
	ymax: maximal y coordinate	0.243**
	xrange: xmax-xmin	-0.006
	yrange: ymax-ymin	0.172**
Scroll (5)	*scrlcnt:* num. of vertical scrolls	-0.008
	scrlfreq: scrlcnt/dwell	-0.206**
	scrldist: total vertical scroll distance	-0.092*
	scrlspeed: scrldist/dwell	-0.212**
	scrlmax: maximum scroll top	-0.026
AOI (3)	*dwell_aoi:* total time the cursor spent in the pre-defined Area of Interest (AOI)	0.227**
	cursorcnt_aoi: cursor count in AOI	0.189**
	cursorfreq_aoi: cursorcnt/dwell	-0.195**
Task (6)	*avg_dwell:* average dwell time of preceding page views in the task	0.081*
	querycnt: num. of preceding queries	-0.138**
	serpcnt: num. of preceding search engine result page (SERP) views	-0.142**
	clkcnt: num. of preceding clicks	-0.171**
	ctr: clkcnt/serpcnt	0.085*
	tasktime: total time elapsed in seconds since the task started	-0.046

Table 1: Feature descriptions and Pearson's correlations with relevance Levels (indicates statistical significance at p < .01 level; * indicates statistical significance at p < .05 level).**

as the minima and maxima of horizontal and vertical cursor coordinates.

4.4 Vertical Scrolling

Previous research (e.g., [10]) found that the amount a user scrolls correlates with the "interestingness" of a Web document in a non-Web search setting, while in a Web search scenario, another study [15] did not find a strong correlation between the amount of scrolling and the "satisfiability" of a clicked document. In this study, in addition to modeling the overall amount of scrolling, we propose to also model the frequency and speed of scrolling behavior, as well as the overall scroll distance and range in pixels. The intuition behind is to capture the searcher's examination patterns. For example, high frequency and speed of scrolling may indicate that the searcher was "scanning" or skipping parts of the document, while a moderate range of scrolling with low speeds may indicate that the searcher was "reading".

4.5 Interactions in the Areas of Interest (AOI)

It has been proposed that searchers are more willing to interact with the content when it is relevant. To capture this idea, we define an "Areas of Interest"(AOI) as the region in a document where the main content lies, and model the searcher behavior within the AOI. In particular, we measure the number and frequency of cursor movements within an AOI, in addition to these measures for the document as a whole. Since a typical Web page has its main content on the left half of the page, we define one AOI as the region of the document with the X-coordinates between 100 and 400 pixels, and the Y-coordinates larger than 100 pixels. More sophisticated estimation of AOI's can be done, but as we will show later, even this simple AOI appears to improve the correlation between the features and document relevance (Section 7.1).

4.6 Task/Session-level Context

As shown in the recent work [22], task-level information could be valuable for improving relevance estimation. The intuition is that a page viewed in a successful search task is likely to be more relevant while a page viewed in a unsuccessful task, is likely to be less relevant. To detect task success, we incorporate previously proposed features, such as the number of queries, number of clicks, click-through rate (CTR), average dwell time, overall task time, and the number of page views. These features have been shown to be effective in detecting success or frustration in previous studies [21, 14, 1] and are potentially useful in improving document relevance estimation [22].

4.7 User Normalization

Previous work has identified significant variation in behavior across different searchers (e.g., [27, 35, 19]). We propose three methods to normalize feature values for individual searchers. First, we could subtract the mean of the feature values for a user, from the original feature values (most common approach); Second, we could subtract the median feature values for the user, as it is typically more robust to outliers than the first approach; third, we could use z-score normalization, which transforms the original feature distribution into normal distribution by scaling the difference between the original value and mean by the standard deviation.

5. RELEVANCE ESTIMATION MODELS

We now describe the machine learning algorithms we used. We treat the relevance estimation as a regression problem, and experiment with two popular regression algorithms.

5.1 Ridge Regression (RR)

The first algorithm is Ridge Linear Regression, which is a variant of ordinary Multiple Linear Regression, whose goal is to circumvent the problem of predictors collinearity and overfitting. Furthermore, the M5's method is used to select attributes for use in the linear regression for each run. Specifically, the algorithm steps through the attributes and removes the one with the smallest standardized coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion. The advantages of using such a linear regressor lie in the easy interpretability and time-efficiency in training, which is potentially favorable in a large scale setting. And the disadvantage mainly lies in the less expressive power of the model, which does not capture the non-linear interaction among different features.

5.2 Bagging with Regression Trees (BRT)

The second algorithm is Bagging[6], which is a method for generating multiple versions of a predictor and using these to get an

aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. The single predictor or weak learner we used is C4.5 regression tree. The advantage of this non-linear regressor is, in contrast, the advanced expressiveness, which can help model the complex relationships among the features, and not surprisingly it suffers from longer training time and may not be applicable in certain large-scale scenarios.

6. EXPERIMENTAL SETUP

Next, we describe our experimental setup on estimating document relevance and re-ranking the documents.

6.1 Data

The data set we used for our experiments, which has hundreds of search tasks and explicit relevance judgments of visited Web pages, is from a user study conducted by researchers at the University of Massachusetts [14]. The usage data of the participants was tracked, specifically, containing the URLs the searchers visited, the fine-grained interactions with the browsed pages, such as clicks, cursor movements, and scrolling, the time-stamp of each page view and interaction is also recorded. The search tasks in the user study were designed to be representative of Web search and 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 should ensure that they do a good job on such search tasks, and as we show later in Section 7, our proposed techniques indeed improve relevance estimation and ranking for such difficult search tasks.

The original dataset is publicly available online [2]. Similarly, the processed data and source code for this paper is available at `http://ir.mathcs.emory.edu/data/WWW2012/`. Next, we describe the details of the user study and the collected data (additional information can be found along with the original dataset).

User study: The study relied on a modified version of the Lemur Query Log Toolbar [3] for Firefox browser. To begin a task, participants had to click a 'Start Task' button. This prompted them with the task and a brief questionnaire about how well they understood the task and the degree to which they felt they knew the answer. They were asked to use any of four search engines: Bing, Google, Yahoo!, or Ask.com and were allowed to switch at any time. Links to these appeared on the toolbar and were randomly reordered at the start of each task. Users were allowed to use tabs within Firefox.

Explicit Judgments: Each time the participants navigated away from a non-search page, they were asked the degree to which the page satisfied the task on a five point scale ("1" indicates the page "did not satisfy the information need at all" and "5" indicates that the page "completely satisfied the information need"), with an option to evaluate later.

We used this self-reported explicit judgment as our ground truth for document relevance. A total of 211 tasks were completed, feedback was provided for 463 queries and 694 visited pages. For our experiments, we studied the set of page views with dwell time at least one second and with at least one cursor coordinate recorded to exclude artificial URL visits (e.g., URL redirections) that are recorded in the dataset and focused on modeling the initial visit of a document in each session as subsequent visits of the same document typically exhibit larger variance in behavior and the dataset consists of only a very small portion of such subsequent page visits. As a result, our final dataset contains 666 page views with relevance judgments.

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

[3] `http://www.lemurproject.org/querylogtoolbar/`

6.2 Evaluation Metrics

Given a feature vector $\mathbf{x}$ of post-click page view, the explicit judgment of page relevance y, and a regression function $f(\mathbf{x})$ (where $(\mathbf{x}, y)$ is an instance of the test dataset D), we evaluated its performance on predicting the document relevance using the standard measure of correlation, and evaluated its performance on re-ranking documents using the standard measure of normalized discounted cumulative gain.

Correlation: Pearson's correlation $\rho_{f(.),S}$ between the document relevance predicted by $f(.)$ and true document relevance y across all instances in the test data D is given by:

$$\rho_{f(.),S} = \frac{\sum_{(\mathbf{x},y)\in D}(f(\mathbf{x}) - \mu_f(.))(y - \mu_y)}{(|D| - 1)\sigma_{f(.)}\sigma_{\mathbf{y}}}$$

where μ is the observed sample mean and σ is the observed sample standard deviation. This correlation coefficient is helpful for detecting the presence of informative predictions, even in the presence of shifting and scaling. The ideal value for correlation is 1.0, with a value of 0 showing no observed correlation.

Normalized Discounted Cumulative Gain at K ($\mathbf{NDCG_k}$): as a standard metric of search engine providers, 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 at penalizing the ranked list with highly relevant documents appearing at lower positions, with the graded relevance value reduced logarithmically proportional to the position of the result. $NDCG_k$ of 1.0 indicates a perfect ranking that is identical to $IDCG_k$ and smaller values indicates worse rankings. We first computed $NDCG_k$ 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 evaluated for various k values.

6.3 Methods Compared

We consider the following different methods for estimating document relevance, including methods using individual feature groups, combined feature groups, with or without user normalization, for both our linear regressor RR and non-linear regressor BRT.

DTR Baseline: we develop a strong baseline model that utilizes signals from click Dwell time, Task-level context, and the search engine original Ranking (*DTR*). This model is representative of the state-of-the-art methods using dwell time [35] and task-level information[22].

Post-Click Behavior (PCB): the full model with all the feature

groups combined, which include cursor movements, scrolling, interactions in areas of interest (AOI) and dwell time, task-level context and rank.

PCB with User Normalization (PCB_User): the full PCB model with user normalization for all feature groups, as described above.

Single Feature Group Runs: we evaluated models trained on the six individual feature groups, namely, dwell time, search engine original ranking, task-level context, cursor movements, scrolling, and interactions in the areas of interest (AOI). In particular, the dwell time, task-level context, and rank feature groups can be considered as three additional baselines to gauge the performance of our models. The three remaining behavioral feature groups are the proposed variants in modeling post-click interactions, and serve as the main building blocks of our full model.

Combined Feature Group Runs: we also evaluated PCB with each single individual feature group removed from the full model to test the contribution of different feature groups when other groups are presented. This is important, as some features in different groups could be correlated.

7. RESULTS

In this section, we describe and discuss our experimental results and findings. We start with analyzing the association between each individual feature and the explicit relevance judgments, and then move on to our results on relevance prediction and document re-ranking, where we evaluate each individual feature group and some combinations of the different groups.

7.1 Feature Association with Relevance

We now discuss the association between each individual feature and the explicit relevance judgements. Specifically, we computed Pearson's Correlation for each feature and conducted statistical significant testing. The results are summarized in Table 1, along with the descriptions of the features, significant associations are highlighted: * indicates significance at $p < .05$ level and ** indicates significance at $p < .01$ level. We organize the discussions by feature groups.

Dwell Time As we can see from Table 1, there is a moderate correlation of 0.167 between dwell time and document relevance, which is consistent with previous findings [35], since longer dwell time typically indicates searcher interests in the page. However, as we can see later, some other post-click behavioral signals are actually correlated better with document relevance, suggesting the potential of improving upon dwell time information.

Rank The correlation between search engine result ranking and document relevance is -0.073, which matches our intuition that smaller rank values correspond to higher relevance. However, the correlation is low and insignificant. One explanation is that all the visited documents on a search trail following a click typically share the same rank (as some of which were not ranked) but vary in their relevance levels. This assumption is supported by the observation of a higher though still insignificant ρ of -0.094 when the correlation is computed over only pages that were ranked by the search engines. This low correlation of the search engine result ranking with relevance reveals the difficulty of the search tasks in our dataset.

Cursor Movements As suggested in the previous section, characteristics of cursor movements are indicative of searcher's reading behavior. Interestingly, we do observe such tendency between the cursor movement features and the document relevance. Starting from the beginning of the list, the amount of cursor movements (i.e., *cursorcnt*) exhibits a similar level of correlation of 0.164 as dwell time, which makes sense, as the longer the time the searcher spent on a page the larger amount she might move the cursor.

A more interesting question then is, whether the cursor movements provide some additional information about document relevance – as we discuss later, cursor movements and dwell time provides complementary information – and based on the results in this section alone, we actually already observe stronger associations from some of the cursor features. For example, the maximal y coordinate of the cursor (i.e., *ymax*) exhibits a stronger correlation of 0.243 with relevance, which suggests that the further down the searcher moves the cursor the more likely she found the page to be relevant. This is consistent with the observation from our case studies (Section 3) – searchers tend to use mouse more actively and "read" when the page is relevant while the page is not relevant, keep mouse still and "scan", in which case, it is less likely that she would move mouse further down. Note that there is a difference between scrolling down and moving mouse down – as we can see from the table, the correlation between maximal scrolling and relevance is only a insignificant -0.026, we hypothesize that searchers tend to scroll when "scanning" and keep mouse still, while more likely to move the cursor to interact when interesting information is found.

Another interesting observation is about cursor movement speed: while overall the amount of cursor movement is correlated positively with document relevance, the speeds, both in vertical and horizontal directions, have negative correlation, which matches our observation and intuition: lower speed of cursor movements is indicative of "reading", which is more likely to happen when the page is relevant. As for horizontal movements, the distance cursor travels exhibits a significant positive correlation of 0.101. This feature captures the horizontal movement of reading aid behavior illustrated in Figure 1, the possible explanation of lower correlation based on our case studies is that the horizontal movement behavior happens less frequent than vertical moves, but when it does happen, it typically is a strong indicator of "reading" [32].

Vertical Scrolling In agreement with previous research [15], we do not observe a significant correlation between the amount of scrolling (i.e., *scrlcnt* and *scrlmax*) and relevance. However, interestingly, we do observe significant negative correlations of scrolling frequency and scrolling speed of -0.206 and -0.212 respectively, which well supports our hypothesis that high frequency and speed of scrolling indicate "scanning" behavior, which in turn, suggests lower document relevance.

Interactions in Areas of Interest (AOI) The intuition behind the AOI features is that searchers are more likely to interact with the content when it is relevant. Therefore, we specify the expected position of the main content of Web page as the AOI, hypothesizing that the interactions within the AOI are more indicative of document relevance. As we can see from Table 1, AOI features exhibits higher correlations as compared to their overall counterparts. For example, the correlation of AOI dwell time, which is the dwell time accumulated when the cursor is within the area of interest, increases substantially from correlation of 0.167 to 0.227 while the correlation of AOI cursor frequency increases even more significantly from -0.082 to -0.195.

Task/Session-level Context In agreement with previous work [22], we found that task-level information is indeed valuable in inferring document relevance. In particular, a document in a more successful search session is indeed more likely to be relevant, which is supported by statistically significant correlations between *CTR* and

relevance, as well as the average dwell time and relevance. In contrast, we find that a document is less likely to be relevant in a less successful search session, which is indicated by the significant negative correlations between relevance and features representing task length (e.g., query count and task time). This makes intuitive sense, since a long session typically indicates the more efforts searchers have to put in finding the information, a claim supported by previous studies [1, 14].

Next, we discuss our findings in predicting documents relevance, and compare the performance of each individual feature groups as well as different feature group combinations.

7.2 Predicting Document Relevance

Now we report our results and findings in predicting document relevance explicitly judged by the users. For training and testing, we used 10-fold cross-validation with 100 randomized experimental runs. We report the overall correlation aggregated over all the folds and runs (note that, each instance occurs only once in exactly one fold for each run). We evaluated the six single feature groups, different combinations of these groups, and the effects of adding user normalization information.

Single Feature Group Runs: The results of the single feature group runs are summarized in Table 2. As we can see, all the three post-click interaction feature groups outperform the three baseline feature groups using dwell time, task-level information and search engine ranks, as well as the stronger *DTR* baseline that combines these three groups of signals; but none of them is comparable with the full model *PCB*. This trend is consistent across both the linear ridge regressor (RR) and the non-linear bagging regressor (BRT) with only exception that *aoi* under-performs *DTR* when using BRT. Specifically, the correlation with relevance for the cursor feature group is the highest, followed by the scrolling feature group, aoi feature group, the task-level, dwell time and rank feature groups. Interestingly, BRT improves the performance of the cursor feature group over RR substantially. One possible interpretation is that the features within the cursor group have complex interactions with each other, which can not be successfully captured using a linear model such as RR.

Single Feature Group	*RR*	*BRT*
PCB	0.399*+	0.411*+
cursor	0.326*+	0.389*+
scroll	0.277+	0.268*+
aoi	0.261*+	0.177*
task	0.201*	0.146*
dwell	0.184*	0.136
rank	0.04	0.136
DTR	0.211	0.231

Table 2: Pearson's correlation between the predicted and actual document relevance for the single feature groups. The groups are listed in descending order of the BRT performance. (* indicates a significant improvement over all the worse-performing groups in the same column at p < .05 level, + indicates a significant improvement over the *DTR baseline* in the same column at p < .05 level)

Combined Feature Group Runs: The results are summarized in Table 3. As we can see, all the combined feature groups again outperform the *DTR* baseline that does not incorporate the post-click interaction features. For the ridge linear regression (RR) setting, the best performing model is the combination of all feature groups *PCB* and removing any one of the groups decreases the performance significantly. Among the six groups, the contributions of the cursor and scroll groups are the most significant while removing each of the other groups only results in decrease with a small margin. As for the non-linear bagging regression (BRT) setting, only the cursor, scroll, and rank groups contribute significantly when other groups are presented and the additive contribution from the ranking information is the least substantial among the three. The three remaining groups, namely, *dwell, task*, and *aoi*, do not seem to contribute additional information when the other groups are presented. One possible explanation is that the non-linear BRT regressor was able to capture the complex relationships among different features and induce the information carried by the features in dwell time, task-level context and AOI interactions, making it unnecessary to incorporate these features when other groups are presented, even though all the feature groups tend to be useful in combination when only a linear regressor such as RR is used.

Combined Feature Group	*RR*	*BRT*
PCB	0.399+	0.411+
no.cursor	0.326*+	0.336*+
no.scroll	0.353*+	0.379*+
no.aoi	0.394*+	0.412+
no.task	0.394*+	0.413+
no.dwell	0.395*+	0.414+
no.rank	0.393*+	0.409*+
DTR	0.211	0.231

Table 3: Pearson's correlation between the predicted and actual document relevance for the combined feature groups. The groups are listed in ascending order of the BRT performance. (* indicates a significant decrease in performance from *PCB* in the same column when removing the feature group at p < .05 level, + indicates a significant improvement over the *DTR* baseline in the same column at p < .05 level)

User Normalization: We further evaluated the effects of adding the user normalization information. The results are summarized in Table 4. As we can see, adding user information to our full model (*PCB_User*) further improves the performance in predicting document relevance, which was the best-performing model among all the other feature combinations, and as expected, the model also outperforms the *DTR* baseline. In particular, the improvement with a linear regressor was smaller compared to that of the non-linear bagging regressor. This result indicates the existence of variation in behavioral signals across different users, a claim supported by previous research [35, 19, 9]. However, as we have seen, even without the user information, the behavioral patterns seem sufficiently consistent to achieve improvement in estimation performance.

Next, we move on to discuss our results and findings on improving result ranking in aggregate using the estimated document relevance from our models.

7.3 Re-ranking

Next, we report our results on re-ranking documents using the estimated relevance from the regressors. For training and testing, we again used 10-fold cross-validation. We report $NDCG_k$ averaged over all the search tasks across different users. Specifically, we compare combined feature groups with one feature group removed at a time, the *DTR* baseline, and the full models *PCB* and

Combined Feature Group	*RR*	*BRT*
PCB_User	0.420*	0.447*
PCB	0.399*	0.411*
DTR	0.214	0.231

Table 4: Pearson's correlation between the predicted and actual document relevance when adding user normalization features (* indicates a significant improvement over the *DTR* baseline in the same column at p < .05 level)

PCB_User. As BRT generally performs better than RR, we use BRT for the rest of the experiments.

The feature ablation results are summarized in Table 5. The trend is the same as what we have observed in Table 3: cursor and scroll feature groups tend to contribute the most, while the rest of the groups contribute marginally when other groups are presented. One interesting difference in this setting is that for smaller K, the contribution of scroll features appears larger than that of the cursor features.

The results of our post-click behavior models, with and without user normalization (*PCB* and *PCB_User*) are reported in Figure 3. Both variants of the *PCB* model again outperform the *DTR* baseline, and adding user normalization features (*PCB_User*) provides additional moderate improvements in ranking, especially for smaller values of *K*.

Combined Feature Group	*K=10*	*K=20*
PCB	0.579	0.675
no.scroll	0.515 (-11.0%)	0.630 (-6.7%)
no.cursor	0.548 (-5.2%)	0.619 (-8.3%)
no.aoi	0.570 (-1.5%)	0.671 (-0.7%)
no.rank	0.576 (-0.5%)	0.669 (-0.9%)
no.dwell	0.578 (-0.1%)	0.677 (+0.2%)
no.task	0.587 (+1.5%)	0.681 (+0.8%)
DTR	0.515 (-10.9 %)	0.598 (-11.4 %)

Table 5: NDCG at K for the combined feature groups with one feature group removed at a time, the groups are listed in ascending order of NDCG@10.

Next, we evaluate the performance of PCB on the subset of documents that were ranked by the search engines (i.e., landing pages). The results are summarized in Figure 4. For the landing pages, *PCB* and *PCB_User* still consistently outperform the *DTR* baseline at all values of K, indicating that PCB predictions could be directly usable by a search engine for improving search ranking quality.

8. CONCLUSIONS

In this paper we introduced a new model for representing the searchers' post-click behavior (PCB) that captures not only dwell time and task-level information, but also fine-grained user interactions *after* clicking on a search result, such as cursor movements and scrolling. To our knowledge, PCB is the first successful attempt to exploit such "low-level" post-click behavioral signals to identify the basic patterns of "reading" and "scanning" behavior, as well as more complex combinations of these (Section 3), coupled with expressive features to capture these examination patterns automatically (Section 4).

Our experimental results show that these behavioral signals indeed correlate with searchers' explicit judgments of document relevance, and provide additional valuable information beyond dwell time and session-level information. Specifically, we found that the distance and range the cursor travels, as well as movement speed, especially its vertical component, are among the most predictive signals of document relevance; we also found that while the amount of scrolling is not itself strongly correlated with document relevance, the frequency and speed of scrolling are. In combination, these signals enable PCB to exhibit significant improvements of relevance estimation, as well as significant improvements in re-ranking the documents based on this relevance estimation. Finally, when user information is available (e.g., for long-term users of a search engine), adjusting the PCB model for each user's "normal" profile can further improve the prediction performance.

In summary, we have laid the groundwork for exploiting fine-grained post-click search behavior for document relevance estimation, identifying common page examination patterns and operationalizing our insights in a novel PCB model for effective relevance prediction. Together, our methods enabled substantial improvements of relevance estimation, and the resulting document ranking over and beyond dwell time alone.

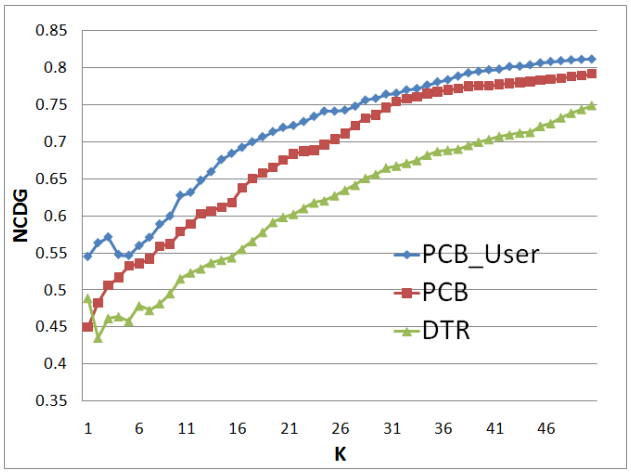


Figure 3: NDCG at K for the *DTR* baseline and our full models with (*PCB_User*) and without (*PCB*) user normalization features in re-ranking all the pages.

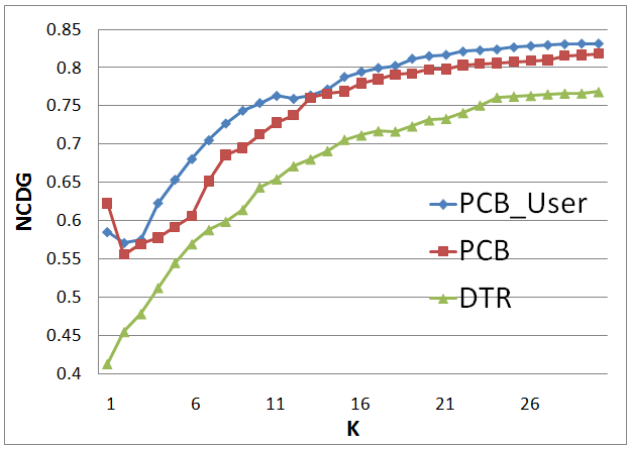


Figure 4: NDCG at K for the *DTR* baseline and our full model (*PCB*) in re-ranking only the landing pages.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant IIS-1018321, and by the Yahoo! Faculty Research Engagement Program. The authors also thank Henry Feild and Dmitry Lagun for assistance in data processing and valuable discussions.

9. REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR 2011)*, 2011.

[2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proc. of SIGIR*, 2006.

[3] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. of SIGIR*, 2006.

[4] 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, New York, NY, USA, 2008. ACM.

[5] N. J. Belkin. User modeling in information retrieval. *Tutorial at UM97*, 1997.

[6] L. Breiman and L. Breiman. Bagging predictors. In *Machine Learning*, pages 123–140, 1996.

[7] G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proc. of SIGIR*, 2008.

[8] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: a comparison to eye tracking. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 67–74, New York, NY, USA, 2009. ACM.

[9] G. Buscher, R. W. White, S. Dumais, and J. Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proc. of WSDM*, 2012.

[10] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, IUI '01, pages 33–40, New York, NY, USA, 2001. ACM.

[11] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proc. of CHI*, 2007.

[12] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 181–190, New York, NY, USA, 2010. ACM.

[13] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, 2008.

[14] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR'10*, pages 34–41, 2010.

[15] 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.

[16] G. Golovchinsky, M. N. Price, and B. N. Schilit. From reading to retrieval: freeform ink annotations as queries. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 19–25, New York, NY, USA, 1999. ACM.

[17] Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 707–708, New York, NY, USA, 2008. ACM.

[18] Q. Guo and E. Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 130–137, New York, NY, USA, 2010. ACM.

[19] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI EA '10: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pages 3601–3606, New York, NY, USA, 2010. ACM.

[20] K. Gyllstrom and C. Soules. Seeing is retrieving: building information context from what the user sees. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, pages 189–198, New York, NY, USA, 2008. ACM.

[21] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *WSDM'10*, pages 221–230, 2010.

[22] A. Hassan, Y. Song, and L.-w. He. A task level user satisfaction metric and its application on improving relevance estimation. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, CIKM '11, New York, NY, USA, 2011. ACM.

[23] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1225–1234, New York, NY, USA, 2011. ACM.

[24] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 41–48, New York, NY, USA, 2000. ACM.

[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] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 408–409, New York, NY, USA, 2001. ACM.

[27] D. Kelly and N. J. Belkin. Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 377–384, New York, NY, USA, 2004. ACM.

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

[29] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 272–281, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

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

[31] K. Rodden and X. Fu. Exploring how mouse movements relate to eye movements on web search results pages. In *Web Information Seeking and Interaction Workshop*, 2006.

[32] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *Proc. of CHI*, 2008.

[33] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proc. of KDD*, 2009.

[34] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 315–323, New York, NY, USA, 1998. ACM.

[35] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 297–306, New York, NY, USA, 2006. ACM.