No Clicks, No Problem: Using Cursor Movements to Understand and Improve Search

Jeff Huang
University of Washington
Seattle, WA 98195 USA
chi@jeffhuang.com

Ryen W. White
Microsoft Research
Redmond, WA 98052 USA
ryenw@microsoft.com

Susan Dumais
Microsoft Research
Redmond, WA 98052 USA
sdumais@microsoft.com

ABSTRACT

Understanding how people interact with search engines is important in improving search quality. Web search engines typically analyze queries and clicked results, but these actions provide limited signals regarding search interaction. Laboratory studies often use richer methods such as gaze tracking, but this is impractical at Web scale. In this paper, we examine mouse cursor behavior on search engine results pages (SERPs), including not only clicks but also cursor movements and hovers over different page regions. We: (i) report an eye-tracking study showing that cursor position is closely related to eye gaze, especially on SERPs; (ii) present a scalable approach to capture cursor movements, and an analysis of search result examination behavior evident in these large-scale cursor data; and (iii) describe two applications (estimating search result relevance and distinguishing good from bad abandonment) that demonstrate the value of capturing cursor data. Our findings help us better understand how searchers use cursors on SERPs and can help design more effective search systems. Our scalable cursor tracking method may also be useful in non-search settings.

Author Keywords

Cursor movements, clicks, implicit feedback, Web search.

ACM Classification Keywords

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

General Terms

Experimentation, Human Factors, Measurement.

INTRODUCTION

Understanding how people interact with Web sites is important in improving site design and the quality of services offered. The Web provides unprecedented opportunities to evaluate alternative design, interaction, and algorithmic methods at scale and *in situ* with actual customers doing their own tasks in their own environments [19]. Such studies typically involve measuring clicks which can be obtained easily at scale. However, they fail to capture behaviors that do not lead to clicks (e.g., which items are attended to, in what order, etc.) or subjective impressions. Gaze-tracking studies with participants present in the laboratory can provide more detailed insights but on a smaller scale. In this paper we consider how mouse movements, which can be collected remotely on a large scale, can be used to understand richer patterns of behavior.

We focus on understanding cursor activities in Web search behavior. People conduct Web searches to satisfy information needs. Their interaction with search engines begins by issuing a search query, then reviewing the search engine results page (SERP) to determine which, if any, results may satisfy their need. In doing so, they may move their mouse cursor around the page, hovering over and possibly clicking on hyperlinks. Small-scale laboratory studies have observed participants making many uses of the cursor on SERPs beyond hyperlink clicking [1,21,25]. These uses include moving the cursor as a reading aid, using it to mark interesting results, using it to interact with controls on the screen (e.g., buttons, scroll bars), or simply positioning the cursor so that it does not occlude Web page content. However, studying such behaviors in small-scale laboratory settings is limited in terms of what inferences can be made.

Tracking mouse cursor movements *at scale* can provide a rich new source of behavioral information to understand, model, and satisfy information needs. Recent research has shown that cursor movements correlate with eye gaze [6,13,25,26], and may therefore be an effective indicator of user attention. We believe that cursor data, like click data [18], can provide signals that reveal searcher intent and may be useful in improving the search experience. Cursor data can be used to complement click data in several ways. First, cursor data can be captured for uncommon queries where strong indicators of relevance such as result clicks may occur less frequently or not at all. For example, analyzing click logs for a query that has been issued several times but never clicked may provide limited relevance information, but cursor behavior on the SERP associated with the query may provide insight about relevance. Second, in cases of so-called *good abandonment* [20], where the content on the SERP satisfies the user's information need directly, a search result click may be unnecessary. Thus the lack of a click should not always be interpreted as a search failure. Cursor behavior may help in distinguishing between good and bad search abandonment.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

The research questions that we ask are: (i) to what extent does gaze correlate with cursor behavior on SERPs and non-SERPs? (ii) what does cursor behavior reveal about search engine users' result examination strategies, and how does this relate to search result clicks and prior eye-tracking research? and (iii) can we demonstrate useful applications of large-scale cursor data? Answers to these questions help us determine the utility of cursor tracking at scale, and ultimately inform search system design and improve the experience for users of search engines.

RELATED WORK

One line of related research has explored the use of cursor movements, clicks, and gaze as *implicit indicators of interest* on Web pages. In early work, Goecks and Shavlik modified a Web browser to record themselves browsing hundreds of Web pages [11]. They found that a neural network could predict variables such as the amount of cursor activity on the SERP, which they considered surrogate measurements of user interest. Claypool et al. [7] developed the "curious browser," a custom Web browser that recorded activity from 75 students browsing over 2,500 Web pages. They found that cursor travel time was a positive indicator of a Web page's relevance, but could only differentiate highly irrelevant Web pages. Surprisingly, they also found that the number of mouse clicks on a page did not correlate with its relevance. Hijikata [15] used client-side logging to monitor five subjects browsing a total of 120 Web pages. They recorded actions such as text tracing and link pointing using the cursor. The findings showed that these behaviors were good indicators for interesting regions of the Web page, around one-and-a-half times more effective than rudimentary term matching between the query and regions of the page. Shapira et al. [27] developed a special Web browser and recorded cursor activity from a small number of company employees browsing the Web. They found that the ratio of mouse movement to reading time was a better indicator of page quality than cursor travel distance and overall length of time that users spend on a page.

In the search domain, Guo and Agichtein [12] captured mouse movements using a modified browser toolbar and found differences in cursor travel distances between informational and navigational queries. Furthermore, a decision tree could classify the query type using cursor movements more accurately than using clicks. Guo and Agichtein also used interactions such as cursor movement, hovers, and scrolling to accurately infer search intent and interest in search results [13]. They focused on automatically identifying a searcher's research or purchase intent based on features of the interaction. Buscher et al. investigated the use of gaze tracking to predict salient regions of Web pages [2] and the use of visual attention as implicit relevance feedback to personalize search [4].

Another line of research examined the *relationship between eye gaze and cursor positions*. An early study by Chen et al. [6] measured this relationship in Web browsing by recording 100 gaze and cursor positions from five subjects browsing the Web. They showed that the distance between gaze and cursor was markedly shorter in regions of encountered pages to which users attended. Liu and Chung [21] recorded cursor activity from 28 students browsing the Web. They noticed patterns of cursor behaviors, including reading by tracing text. Their algorithms were capable of predicting users' cursor behaviors with 79% accuracy.

More recent work has focused on the relationship between cursor and gaze on search tasks. In a study involving 32 subjects performing 16 search tasks each [25,26], Rodden et al. identified a strong alignment between cursor and gaze positions. They found that the distance between cursor and gaze positions was longer along the x-axis than the y-axis, and was generally shorter when the cursor was placed over the search results. Rodden et al. also observed four general types of mouse behaviors: neglecting the cursor while reading, using the cursor as a reading aid to follow text (either horizontally or vertically), and using the cursor to mark interesting results. Guo and Agichtein [14] reported similar findings in a smaller study of ten subjects performing 20 search tasks each. Like Rodden et al., Guo and Agichtein noticed that distances along the x-axis tended to be longer than the distances along the y-axis. They could predict with 77% accuracy when gaze and cursor were strongly aligned using cursor features.

The research presented in this paper extends previous work in a number of ways. Our analysis of the cursor-gaze relationship (Study 1) involves more search tasks than prior studies, compares SERP and post-SERP Web pages, and confirms earlier results with a large study using the same SERP layout that we use in the remainder of the paper. More importantly, we develop a scalable approach to capturing cursor data that enables us to analyze real user activity in a natural setting for more than 360 thousand searches from an estimated 22 thousand searchers (Study 2). Finally, using two case studies, we show how cursor data can supplement click data on two search-related problems.

STUDY 1: GAZE-CURSOR RELATIONSHIP

We begin by replicating and extending prior laboratory experiments on the relationship between gaze and cursor activity using the same SERP layout deployed in our large-scale cursor study (Study 2, see Figure 2). Study 1 also involves more tasks and participants than prior laboratory studies, and measures the relationship between gaze and cursor position on SERP and on non-SERP pages.

Data

We used a Tobii x50 eye tracker with 50Hz tracking frequency and 0.5° visual angle on a 1280 × 1024 resolution 17 inch monitor (96.42dpi) and 1040 × 996 resolution Internet Explorer 7 browser. Cursor and gaze coordinates were collected in an eye-tracking study of 38 participants (21 female, 17 male) performing Web searches. Participants were recruited from a user study pool. They ranged in age between 26 and 60 years (mean = 45.5, σ = 8.2), and had a wide variety of backgrounds and professions.

Each participant completed 32 search tasks on the same search engine, with the same SERP layout template, as used for the large-scale cursor study described in the next section (see Figure 2). Half of the tasks were navigational (i.e., they had to find a specific Web page) and half were informational (i.e., they had to find factual information). Each task started with a description of what participants should look for on the Web. Gaze and cursor positions were recorded for each SERP as well as subsequent Web pages (i.e., pages visited after clicking on a search result). In total, we collected data for 1,210 search tasks, 1,336,647 gaze positions, and 87,227 cursor positions. Gaze-specific findings on this data set, unrelated to cursor behavior, have been reported by others [5,10]. Those researchers granted us access to their data so that we could examine the relationship between gaze and cursor behaviors.

Gaze and cursor positions were extracted from the eye-tracking logs. In our data, the gaze positions were recorded approximately every 20ms, whereas cursor positions were recorded approximately every 100ms. Since cursor and gaze events did not always have identical timestamps, a gaze position was interpolated for every cursor position. Interpolation was performed by calculating gaze x and y coordinates weighted by the coordinates of the nearest gaze coordinates before and after the cursor position. For example, the interpolated x-coordinate for eye gaze is calculated as $x_0 + (x_1 - x_0)((t_c - t_0)/(t_1 - t_0))$ where t_c is the time for the corresponding cursor position, x_0 is the gaze's x-coordinate preceding the cursor position, recorded at time t_0, and x_1 is the gaze's x-coordinate following the cursor position, recorded at time t_1. To reduce noise, cursor positions were only captured if they occurred between gaze positions that were at most 100ms apart.

Findings

Figure 1 shows the frequency distribution for different values of Δx (distance between cursor x-coordinate and gaze x-coordinate), Δy (distance between cursor y-coordinate and gaze y-coordinate), and Euclidean distance between cursor and gaze coordinates, i.e., $\sqrt{(\Delta x)^2 + (\Delta y)^2}$. The solid lines in Figure 1 show the distances for SERP pages. As can be seen, cursor and gaze positions are quite similar for both x and y values, their deltas peaking near 0, when the gaze and cursor positions are in the same place. The mean Euclidean distance between cursor and gaze is 178px (σ = 139px) and the median is 143px. The most common offset for the cursor is +3px (to the right) for the x-coordinate and +29px (lower) for the y-coordinate. That is, the cursor is most likely to be just below where the user is focusing with their eyes. We also observed that the differences are greater in the x than y direction (average 50px in the x direction and 7px in the y direction), similar to other studies [14, 25]. Possible explanations for the difference between Δx and Δy include: (i) users may place the cursor to the left or right of their gaze to prevent it from obscuring the text as they read up or down, and (ii) computer screens are usually wider, offering more horizontal space for the cursor.

The dotted lines in Figure 1 represent post-SERP landing pages. Distances between the gaze and cursor on the landing pages were greater than those on the SERP (215px vs. 178px), perhaps due to greater variance in the layout and the content of those pages, as has already been suggested by earlier gaze analysis [2]. Thus the cursor is a better proxy for user attention on the SERP than post-SERP pages. Monitoring cursor behavior on SERPs may help estimate which results or features users attend to and when, and we now turn to a large-scale study of this.

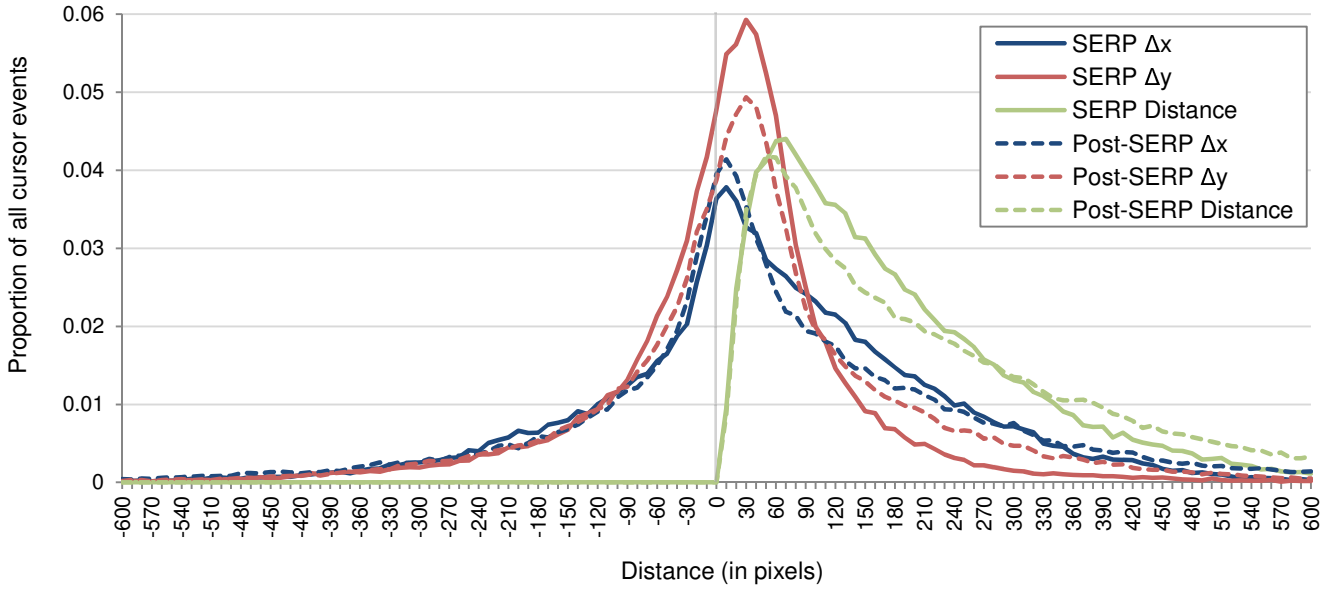

Figure 1. Δx, Δy, and Euclidean distance plotted in a frequency distribution for SERP and post-SERP pages. Solid lines represent these distances gathered on the SERP, while dashed lines represented distances gathered on post-SERP pages (landing pages).

STUDY 2: LARGE-SCALE CURSOR TRACKING STUDY

Following on from the eye-tracking study, we instrumented cursor tracking on the SERP of the Bing search engine, deployed as an internal flight within Microsoft. Cursor tracking at scale involves careful instrumentation of the SERP to address issues with page load latencies associated with the cursor capture script, and the need to remotely record the large volumes of data generated from cursor behavior. We now describe the method that we devised to capture cursor movement data on SERPs at scale.

Method

We wanted to collect cursor data without requiring additional installation. To do this, we instrumented the search results page using client-side JavaScript embedded within the HTML source for the results page. The embedded script had a total size of approximately 750 bytes of compressed JavaScript, which had little effect on the page load time. The script recorded users' cursor interaction within the Web page's borders relative to the top-left corner of the page. Since cursor tracking was relative to the document, we captured cursor alignment to SERP content regardless of how the user got to that position (e.g., by scrolling, or keyboard). Therefore this approach did not constrain other behaviors such as scrolling or keyboard input.

In previous cursor tracking studies, cursor position was recorded at particular time intervals, such as every 50 milliseconds (ms) [13] or every 100ms [25]. This is impractical at a large scale because of the large amount of data to transfer from the user's computer to the server. One alternative is to record events only when there is activity, but this is still problematic because even a single mouse movement can trigger many mouse movement events. We devised a different approach by only recording cursor positions after a movement delay. From experimentation, we found that recording cursor positions only after a 40ms pause provided a reasonable tradeoff between data quantity and granularity of the recorded events. This approach recorded sufficient key points of cursor movement, e.g., when the user changed directions in moving or at endpoints before and after a move; occasionally, points within a longer movement were also captured if the user hesitated while moving. All mouse clicks were recorded since they were less frequent. The events were buffered and sent to a remote server every two seconds and also when the user navigated away from the SERP through clicking on a hyperlink or closing the tab or browser; this was typically 1-3 kilobytes of data. The pseudo-code below summarizes this logic.

```
onCursorMove:
  loc = getCursorPos()
  wait(40 milliseconds)
  if loc == getCursorPos(): // cursor stable for 40ms
    buffer.add(time,loc,getRegion(loc),"position")
onCursorClick:
  buffer.add(time,loc,getRegion(loc),"click")
onTick, onPageClose:
  send(buffer)
  clear(buffer)
```

A server-side process aggregated data from multiple pageviews belonging to the same query (e.g., from returning to SERP using the browser "back" button or viewing multiple result pages), to facilitate query-level in addition to pageview-level analysis. All analysis presented in this paper is at the query level. Table 1 describes the fields present in each record. We identify *regions* that the cursor hovers over using attributes in the HTML, and use two such regions in subsequent analyses (result rank, link id).

Table 1. Fields in data recorded by cursor tracking script.

Field	Description
Event	Cursor move or click
Cursor Position	x- and y-coordinates of the cursor
Timestamp	Time that the event occurred
Region	Result rank or link id
QueryId	Unique identifier for each query
CookieId	Unique identifier for each cookie
Query	Text of the issued query
Result URL	URL of clicked result (if any)

The large volume of data collected using the method described in this section allowed us to examine a number of aspects of how searchers use their cursors on SERPs. For this purpose, we use the query-level data, comprising all clicks and cursor movements for a query instance. In addition to the location of cursor positions, we summarize the total amount of cursor activity for a query using *cursor trails* (i.e., complete contiguous sequences of cursor movements on the SERP). As we show later, these trails are useful in situations where no clicks are observed.

Data were accumulated from a random sample of Microsoft employees' searches on the commercial Web search engine used between May 12, 2010 and June 6, 2010. In total, we recorded 7,500,429 cursor events from 366,473 queries made by 21,936 unique cookies; the actual number of users may be fewer since multiple cookies could belong to a single user. Although we realize that employees of our organization may not be representative of the general Web searcher population in some respects, e.g., they were more technical, we believe that their interaction patterns can provide useful insights on how SERPs are examined.

We now summarize our results on general cursor activity, evidence of search result examination patterns, and the relationship between click and cursor hover activity. We then present two applications demonstrating the potential utility of gathering cursor data at scale.

General Cursor Activity

We begin by determining where on the SERP users click and move their mouse cursors. This offers some initial insight into differences between click and movement data.

Figure 2 shows heatmaps for clicks and cursor movement activity for the same query aggregated over all instances of the query *[lost finale explanation]* (in reference to the final episode of the US television series "Lost") observed 25 times from 22 different users in our data. Heavy interaction

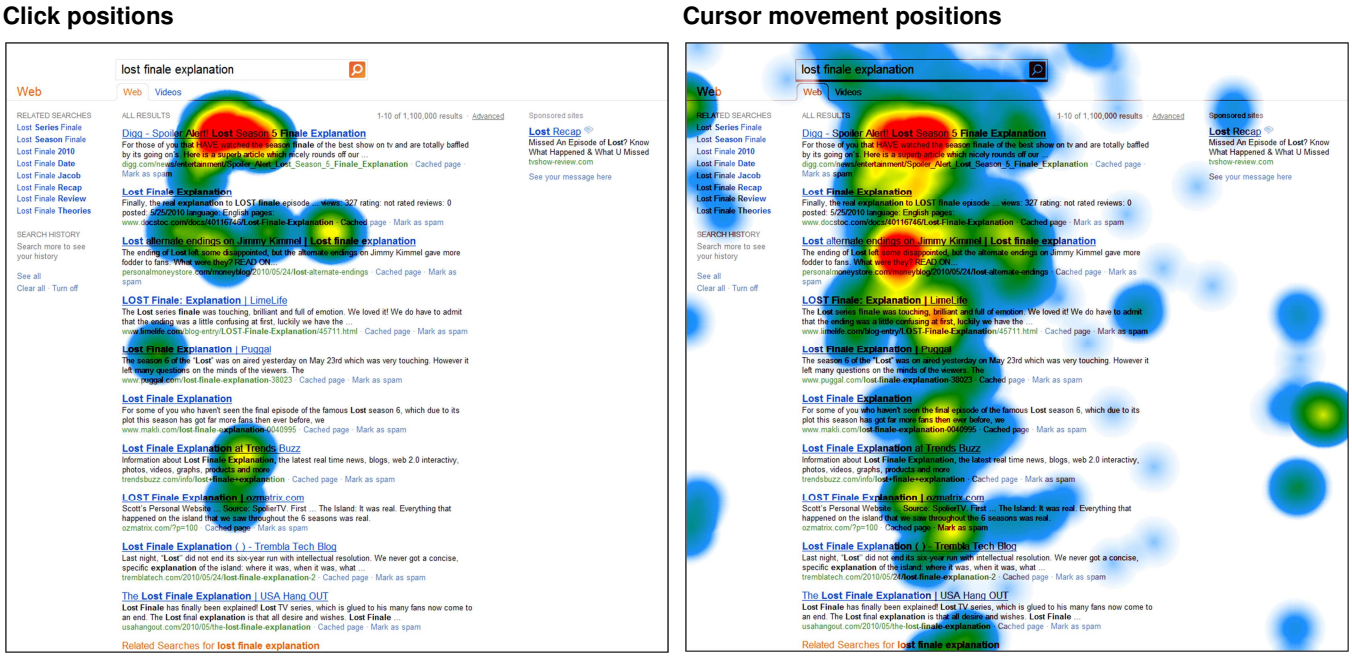

Figure 2. Heatmaps of all click positions (left) and recorded cursor positions (right) for the query *[lost finale explanation]*. Heavy interaction occurs in red/orange/yellow areas, moderate interaction in green areas, light interaction in blue areas.

occurs in red/orange/yellow areas, moderate interaction in green areas, and light interaction in blue areas. Most of the clicks occur on results 1, 3 and 7, and this is also seen in the cursor activity. However, there are some interesting differences as well. For example, there is considerable cursor activity on result 4 even though it is not clicked. The cursor heatmap also shows some activity on query suggestions (on the left rail) and advertisements (on the right rail) although there are no clicks on these regions. Across all queries, cursor positions are more broadly distributed over the SERP than clicks. Thus cursor movement can provide a more complete picture of interactions with elements on the SERP. Such information may be useful to search engine designers in making decisions about what content or features to show on search result pages.

Search Result Examination

In addition to monitoring general cursor movement activity on the SERP, we can also summarize cursor movements that reflect how people examine the search results. Previous work on gaze tracking demonstrated differences in the length of time that users spend reading each of the results based on its position in the ranked list [9]. In a similar way, we were interested in whether the time participants spent hovering over the search results was related to the position in the ranked list. To reduce noise caused by unintentional hovering, we removed hovers of less than 100ms in duration. In Figure 3 we present a graph of the average time spent hovering over each search result title (shown as bars; corresponding scale shown on the left side), and the average time taken to reach each result title in the ranked list (shown as circles connected with lines; corresponding scale on the right side). Error bars denote the standard error of the mean (SEM).

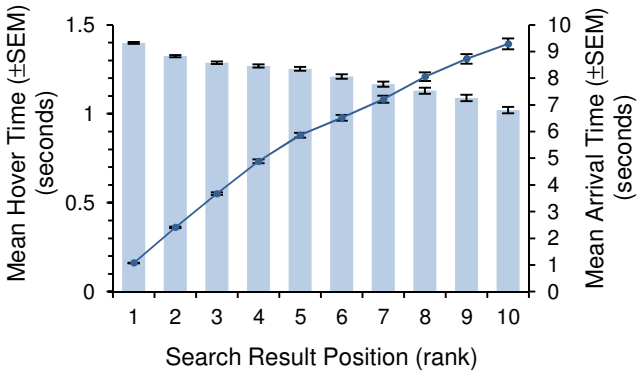

Figure 3. Mean title hover duration (bars) and mean time for cursor to arrive at each result (circles).

The figure shows that time spent hovering on the results decreases linearly with rank and that the arrival time increases linearly with rank. The results are generally similar to gaze tracking findings reported in previous literature [5,9,18]. Hover time decreases with rank as was previously reported; however, cursor hover time drops off less sharply than gaze duration. This difference may be because we miss some of the rapid skimming behavior on low ranks that has been observed previously [5,9,18] since we only recorded hovers after a 40ms pause (to reduce data payload) and filtered out hovers of 100ms or less (to reduce cases of capturing accidental hovers). As expected, search results that are lower ranked are entered later than higher ranked results due to the typical top-to-bottom scanning behavior [8]. The arrival time is approximately linear, suggesting that users examine each search result for a similar amount of time.

We also examined which results were hovered on before clicking on a result, re-querying, or clicking query sugges-

tions or advertisements. This provides further information about how searchers are using their cursor during result examination and again allows us to compare our findings with prior eye-tracking research from Cutrell and Guan [9]. Figure 4 summarizes our findings. This figure shows the mean number of search results hovered on before a click as blue lines, and clicks as red circles. The data are broken down by result position (1-10), and separately for clicks on query suggestions, clicks on ads, and re-queries.

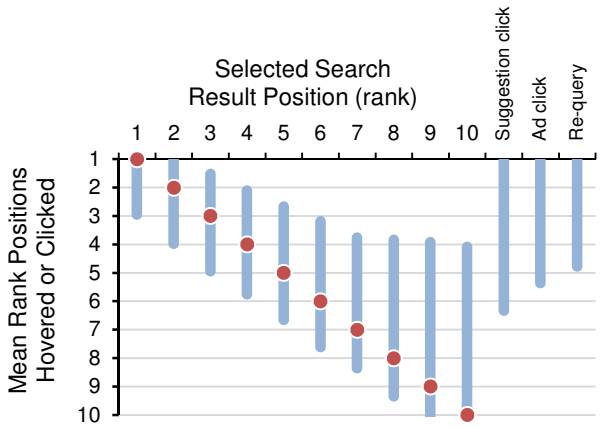

Figure 4. Mean number of search results hovered over before users clicked on a result (above and below that result). Result clicks are red circles, result hovers are blue lines.

Figure 4 shows that prior to clicking on a search result, people consider the surrounding search results. For example, before clicking on result 1, people also hover on results 2 and 3 on average; when they click on result 2 they also hover on results 1, 3, and 4; etc. The findings are similar to those reported by Cutrell and Guan [9], but differ in that the search result hovers do not appear to extend as far *above* the clicked search result in cases where a result is clicked on far down the list (in positions 6–10). This may be because queries where low-ranked clicks are observed may have clearly irrelevant results in top ranks, and by excluding hovers of less than 100ms we miss rapid skims over such irrelevant results.

The findings also show that users consider many results prior to turning their attention to the additional SERP features: on average six results in the case of query suggestions, five results in the case of advertisements, and around four results prior to re-querying. This behavior is similar to that reported in [9], at least in terms of re-querying, which is examined in both studies. Cutrell and Guan do report inspection further down the list (up to rank position 8) prior to re-querying, whereas we find that users hover on around 4 results. One explanation for the difference is that the cursor does not track well with eye movements in situations where users rapidly skim low-ranked search results. An alternative explanation is that in naturalistic non-laboratory settings, users may only consider the top-ranked search results prior to trying another query by clicking on a query suggestion or re-querying.

In the next section we compare the distributions of search results clicks and search result hovers.

Comparing Click and Hover Distributions

Prior studies have presented data on click distribution [18,24] or gaze distribution for the search results [5,18]. These distributions tell us how much attention is given to each result because of its rank and other features such as snippet content [9]. Some theoretical models of behavior depend on accurate models of these distributions, e.g., [16] assumes the frequency with which users review a search result is a power law of its rank, while [28] assumes the frequency with which a search result is clicked follows a geometric distribution of its rank.

In this experiment, we show a cursor hover distribution, and compare it with the corresponding click distribution. Figure 5 shows both the number and proportion of cursor hovers and clicks that occur on each of the top ten search result links. Bars representing absolute counts are scaled by the primary y-axis (on the left), e.g., there are approximately 240,000 occurrences of the cursor hovering on the first search result. Circles representing percentages are scaled by the secondary y-axis (on the right), e.g., 50% of result clicks occur on the first search result.

As is typical in SERP interactions, users interact more with top-ranked search results since they are in a more accessible location and are generally more relevant. However, Buscher et al. [5] reported that the distribution of clicks does not always reflect the relative distribution of visual attention (measured by gaze in their study). Similarly, we find that hovers are more evenly distributed across the top-ten results than clicks, and the *hover rate* is higher than clickthrough rate for all ranks beyond the first position. There are proportionally more clicks than attention on the top-ranked result, which is consistent with previously-reported bias towards selecting one of the top organic results [18,22]. This suggests that for lower-ranked search results, result hovers may correlate with clicks more than at top ranks.

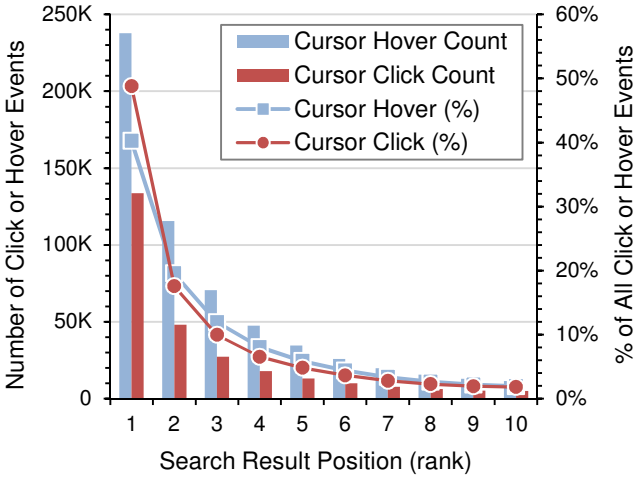

Figure 5. Frequencies and percentages of cursor hovers and clicks occurring on the search results. Percentages reflect the proportion of hover or click events over all ten results.

Unclicked Hovers

Finally, we are interested in seeing if hovering over a result but not clicking on it can be a useful signal of user interest. To examine this, we define an *unclicked hover* as an instance of the cursor hovering over a link but not clicking that link before being moved to another location on the page. Table 2 shows the number of unclicked hovers on a search result and the percentage of times that it was subsequently clicked by the same user. Result clicks can occur without an unclicked hover when the user does not hover over the result for at least 100ms and go to another location on the page before coming back and clicking the result.

Table 2. Percentage of unclicked hovers for which the hovered search result was eventually clicked.

# unclicked hovers	0	1	2	3	4	5
Result clicked	7.0%	16.7%	19.0%	22.4%	23.3%	25.2%

When there are no unclicked hovers, the result is not very likely to be clicked (only 7% of the time). Observing one or more unclicked hovers dramatically increases the likelihood of a result click, perhaps because it demonstrates that the user has attended to it. The more unclicked hovers the more likely the user will ultimately return to the result and click it. The Pearson correlation between the number of unclicked hovers and the percentage eventually clicked is strong ($r = 0.84$), when considering up to 10 unclicked hovers. Thus the number of unclicked hovers on a result may help predict result clickthrough or perhaps result relevance.

Segmenting the unclicked hovers by the search result rank shows that result rank significantly affects unclicked hover behavior. Figure 6 shows the proportion of each result rank that is eventually clicked after an unclicked hover.

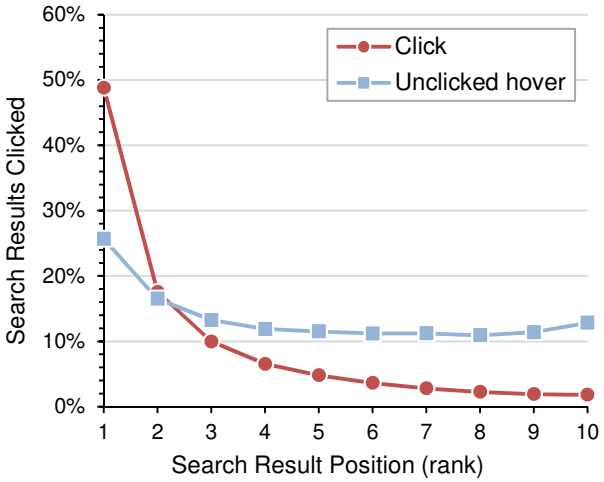

Figure 6. Proportion of search results that are eventually clicked after an unclicked hover, plotted against the click distribution from Figure 5.

The blue squares show that a search result is eventually clicked after an unclicked hover around 25% of the time for the top-ranked result and less than 15% for low-ranked results. However, when we consider that low ranked results typically have a low clickthrough rate, an unclicked hover on a low ranked result may actually provide implicit feedback that the result is relevant. To show this, we overlay the click distribution on the chart to compare the probability that an unclicked hover results in a later click (blue squares) with the original probability that the search result will be clicked (red circles). We see that whether an unclicked hover is a positive or negative indicator depends on result rank. To quantify the degree of this effect we compute the phi correlation (φ) across all queries. For the first search result, the presence of unclicked hovers negatively correlates with result clicks ($\varphi = -0.47$), but for results at lower ranks, unclicked hovers positively correlate with clicks ($\varphi = 0.59$).

In this section we examined characteristics of cursor behavior during search result examination using data from a large-scale study involving more than 20 thousand people and 360 thousand queries. We now turn to applications of this data for improving our understanding of search.

APPLICATIONS OF CURSOR DATA

There are a broad range of possible applications of large volumes of cursor tracking data, from query classification to search interface enhancements. In this section we present two applications mentioned earlier in this paper: estimating search result relevance and distinguishing good abandonment from bad abandonment. The first application makes use of features from search result hovers, whereas the second uses features derived from cursor trails.

Estimating Search Result Relevance

One useful application of cursor movement data is for estimating search result relevance. At scale, these data could be used as an additional data source to train search engine ranking algorithms and boost retrieval performance. We conducted a study in which we gathered human relevance judgments for query-URL pairs, and examined the correlation between features of the cursor movements and the human relevance judgments. In addition, we examined the value that cursor movements provide compared with search result clicks, the more traditional source of behavioral data used to estimate search result relevance.

We obtained human relevance judgments for thousands of queries as part of an ongoing evaluation of search engine quality. Trained judges assigned relevance labels on a five-point scale—*Bad*, *Fair*, *Good*, *Excellent*, and *Perfect*—to top-ranked pooled Web search results for each query. This provided hundreds of relevance judgments for each query. We intersected the judgment data with our cursor data, resulting in 1,290 query-result URL pairs for which we had both explicit relevance judgments and cursor activity. These pairs formed the basis of our analysis. We computed the following features for each pair:

- **Clickthrough rate:** Fraction of the times that URL was clicked when the query was issued (and URL returned).
- **Hover rate:** Fraction of times that URL was hovered over when the query was issued (and URL returned).

- **Number of unclicked hovers:** Median number of times for which the query was issued and the URL is hovered on but not clicked, per the earlier definition. We selected the number of unclicked hovers as a feature because we found that it was correlated with clickthrough in our previous analysis.
- **Maximum hover time:** The maximum time that the user spent hovering over the result per SERP instance. We take the maximum as this indicates the point where the user was most interested in the result.

As stated earlier, the clickthrough rate is commonly used to estimate the relevance of a URL to a query from behavioral data [18], and is included in this analysis as a baseline.

We computed the Pearson correlations between each feature and the human relevance judgments (represented numerically as a five-point scale ranging from 0 to 4 inclusive) independently and in combination using linear regression. Table 3 summarizes the findings, grouped by whether results were clicked for the query. All correlations and differences between correlations are significant at $p < 0.02$ using Fisher's z' transformations where appropriate.

The results of this analysis show that the use of cursor tracking data can improve estimates of search result relevance. Result hover features correlate better with human relevance judgments than clickthrough rates (0.46 vs. 0.42), and they lead to an improved model when combined with clickthrough (0.49 vs. 0.42). In addition, even when there are no clicks for a query, hover features show a reasonable correlation with human judgments (0.28). This is particularly important since many queries occur infrequently, resulting in little or no clickthrough data. Further analysis on the impact of query-URL popularity shows that hover features provide most value over clickthrough rate when query-URLs are less popular. There are large and significant increases in the correlation for query-URL pairs with fewer than five instances in our cursor data (0.45 hover vs. 0.35 click) and small and not significant for pairs with five or more instances (0.59 hover vs. 0.58 click). Thus cursor data appears to be especially useful when click data is less plentiful, which allows relevance estimates to be made for a much larger range of queries.

The correlations between human judgments and unclicked hovers and hover time are interesting as well. For clicked queries, unclicked hovers and hover time are negatively correlated with relevance judgments. This appears to contradict previous findings which suggested that hesitation over a result is a positive indicator of relevance [21,23]. This may be because clicks often occur on top-ranked results, where unclicked hovers are negatively correlated with clickthrough (as shown in Figure 6). For unclicked queries, we find small positive correlations between judgments and all measures. Unclicked queries have fewer relevant results, leading to more exploration lower in ranked list (where unclicked hovers are positively correlated with clicks).

Table 3. Correlations between click and hover features and relevance judgments for queries with and without clicks.

Result clicks or no clicks	Feature source	Correlation with human relevance judgments
Clicks (N=1194)	Clickthrough rate (c)	0.42
	Hover rate (h)	0.46
	Unclicked hovers (u)	-0.26
	Max hover time (d)	-0.15
	Combined[1]	**0.49**
No clicks (N=96)	Hover rate	0.23
	Unclicked hovers	0.06
	Max hover time	0.17
	Combined[2]	**0.28**

[1] $y = 2.25 - 0.1c + 1.38h - 0.08u - 0.12d$; [2] $y = 0.36 + 0.80h + 0.22u + 0.30d$

In this section we showed that the correlation between explicit relevance judgments and search activity increases when cursor actions are added to clicks, especially when clicks are infrequent or unavailable.

Differentiating Between Good and Bad Abandonment

A second application of cursor information is to distinguish between different types of abandonment. Abandonment occurs when searchers visit the search engine result page, but do not click. As noted in previous research [20], abandonment can suggest that users are dissatisfied with the search results (bad abandonment) or that they have found the answer directly on the SERP (good abandonment). For example, for queries like *[Vancouver weather]* or *[WMT stock price]*, answers are typically shown on the SERP so there is no need for people to click through to other pages. We now examine whether features of SERP cursor behavior can distinguish between good and bad abandonment.

As reported in [20], it may be straightforward to estimate good or bad abandonment for queries where search engines offer special interface treatment (e.g., weather updates or stock quotes). A more challenging scenario is determining whether observed abandonment for other queries is good or bad. To study this we focused on selected queries from our log data that were short questions (ending in a question mark) which could be answered by SERP snippets. A similar query class was also studied in earlier abandonment research [20]. To identify examples of likely good abandonment in such cases, we performed some hand labeling.

To determine whether these short questions were answered, one of the authors reviewed the SERPs returned to users and identified whether an answer appeared in the snippet text of results. Judgments were made for results which were hovered on for at least 100ms, indicating that they had been attended to but not clicked on. Of the 859 queries for which the SERPs were visually inspected, 184 (21%) contained the answer in the snippet content and hence were identified as likely examples of good abandonment. The remaining 675 queries were classified as bad abandonment.

We computed summary measures that reflect how the cursor was used on the SERPs. Specifically, we looked at cursor trail length, cursor movement time, and cursor speed for each SERP, defined as follows:

- **Cursor trail length:** Total distance (in pixels) traveled by the cursor on the SERP.
- **Movement time:** Total time (in seconds) for which the cursor was being moved on the SERP.
- **Cursor speed:** The average cursor speed (in pixels per second) as a function of trail length and movement time.

Table 4 shows the mean (M) and SEM for each measure.

Table 4. Features of cursor trails for queries associated with likely good and bad abandonment.

Feature	Abandonment Type			
	Good		*Bad*	
	M	SEM	M	SEM
Cursor trail length (px)	1084	98	1521	71
Movement time (secs)	10.3	0.9	12.8	0.6
Cursor speed (px/sec)	104	9	125	5
Number of queries	184		675	

As can be seen from the table, our preliminary analysis reveals differences in trail length, movement time, and the speed with which users moved their mouse cursor in good and bad abandonment queries. Cursor trails were shorter in cases where good abandonment was likely, compared to instances of bad abandonment. Searchers also spent less time moving the cursor, and moved the cursor more slowly when answers were in the snippet (good abandonment). All differences between the measures for good and bad abandonment were significant using independent measures t-tests (trail length: $t(857) = 2.58$, $p = .01$; movement time: $t(857) = 2.20$, $p = .03$; cursor speed: $t(857) = 2.17$, $p = .03$). It appears that when the answer appears on the SERP, users need to consider fewer results, and move the cursor more slowly as they examine snippet content in detail. These findings show that features of cursor trails, such as length, duration, and speed, are different for good and bad abandonment. Further research will examine the extent to which these differences in feature values can be used to predict good and bad abandonment in operational settings for a broader set of queries.

We now discuss the implications of these and other findings presented in this paper for the design of search systems.

DISCUSSION AND IMPLICATIONS

In this paper we explored how cursor data, which can be easily collected at scale, can be used to augment more commonly-used click measures to provide a richer picture of how searchers interact with search result pages. In Study 1 we extended previous work on the relationship between gaze and cursor using a large laboratory study. In Study 2 we developed and deployed a system to collect much larger-scale cursor data, and summarized search result examination patterns evident in cursor activities. Finally, we presented two applications of these cursor data to improve estimates of search result relevance and distinguish good from bad search abandonment. We believe that these results are quite promising, but can be extended in several ways.

Our analyses report aggregate data averaged over all users and all queries. Some laboratory studies have shown differences in search strategies for different people [2,10]. We would like to examine the extent to which such individual differences are also reflected in large-scale cursor behavior. Additionally, page layouts for SERPs are constantly changing and differ between queries. We would like to better understand how SERP features such as advertisements, query suggestions, or snippet presentation methods, as well as general component layout, influence cursor behavior. There is also a need to study the effect of different methods of scrolling (e.g., via the scrollbar or mouse scroll wheels) on the cursor tracking data. Finally, we would like to extend our research to other search engines and a broader range of users from outside Microsoft Corporation.

We demonstrated two applications of cursor tracking data at scale: estimating search result relevance and distinguishing good from bad search abandonment. In both cases, we showed clear value from leveraging large amounts of cursor tracking data. We showed that cursor features were useful for queries with result clicks (where cursor data augment click data) and for queries without result clicks (where cursor data can be a reasonable substitute). We believe that cursor features may also be used for other search-related tasks such as query classification or search ranking, and for a variety of other Web-based applications. .

Cursor data has qualitative uses as well. Usability tools that use cursor behavior (e.g., [1]) may be useful to search quality analysts. For aggregate analysis of mouse movements, heatmaps (such as those in Figure 2) can show where the interaction took place for different SERP features or queries. Heatmaps allow analysts to quickly see aggregate behavior across multiple query sessions or queries. This may be useful for determining whether users notice new features and how cursor behavior changes following their introduction.

We are also interested in continuing to explore methods for summarizing cursor activity that incorporate interesting characteristics of search behavior but can also be collected at large scale. Our current approach of only recording movements after a 40ms pause precludes analysis of cursor metrics such as acceleration. There may be alternative approaches for sampling cursor movement such as identifying sub-movement boundaries [17]. Exploring efficient methods to tune the trade-off between performance and data granularity is an important next step.

CONCLUSIONS

We described a study of mouse cursor behavior on SERPs combining eye-tracking findings with analysis of large-

scale cursor data. We: (i) show that cursor and gaze position are correlated especially on search result pages, confirming and extending previous findings, (ii) develop a scalable approach to capturing cursor movements at scale and analyze search behavior evident in these data; and (iii) demonstrate the value of cursor features in two applications (estimating search result relevance and distinguishing good and bad search abandonment). Our study adds to the general understanding of how users examine search results, which is typically difficult to study in naturalistic settings on a large-scale, and demonstrates utility of these data in search-related applications. Future work will explore enhancements to the applications presented and investigate the applicability of our general approach to recording cursor movements at scale in settings beyond search.

ACKNOWLEDGMENTS

We thank Danny Bain, Craig Miller, Sarvesh Nagpal, and other members of the Bing team, for help with the development and deployment of the cursor tracking code.

REFERENCES

1. E. Arroyo, T. Selker, and W. Wei. Usability tool for analysis of web designs using mouse tracks. *Ext. Abstracts CHI '06*, 484–489.
2. A. Aula, P. Majaranta and K-J. Raiha. 2005. Eye-tracking reveals personal styles for search result evaluation. *Proc. INTERACT '05*, 1058–1061.
3. G. Buscher, E. Cutrell., and M.R. Morris. What do you see when you're surfing? Using eye tracking to predict salient regions of web pages. *Proc. CHI '09*, 21–30.
4. G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. *Ext. Abstracts CHI '08*, 2291–2996.
5. G. Buscher, S. Dumais, and E. Cutrell. The good, the bad, and the random: An eye-tracking study of ad quality in web search. *Proc. SIGIR '10*, 42–49.
6. M.C. Chen, J.R. Anderson, and M.H. Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. *Ext. Abstracts CHI '01*, 281–282.
7. M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. *Proc. IUI '01*, 33–40.
8. N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. *Proc. WSDM '08*, 87–94.
9. E. Cutrell and Z. Guan. What are you looking for?: An eye-tracking study of information usage in web search. *Proc. CHI '07*, 407–416.
10. S. Dumais, G. Buscher, and E. Cutrell. Individual differences in gaze patterns for web search. *Proc. IIiX '10*, 185–194.
11. J. Goecks and J. Shavlik. Learning users' interests by unobtrusively observing their normal behavior. *Proc. IUI '00*, 129–132.
12. Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. *Proc. SIGIR '10*, 707–708.
13. Q. Guo and E. Agichtein. Ready to buy or just browsing? Detecting web searcher goals from interaction data. *Proc. SIGIR '10*, 130–137.
14. Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. *Ext. Abstracts CHI '10*, 3601–3606.
15. Y. Hijikata. Implicit user profiling for on demand relevance feedback. *Proc. IUI '04*, 198–205.
16. J. Huang and A. Kazeykina. Optimal strategies for reviewing search results. *Proc. AAAI '10*, 1321–1326.
17. R.J. Jagacinski, D.W. Repperger, M.S. Moran, S.L. Ward, and B. Glass. Fitts' law and the microstructure of rapid discrete movements. *J. Exp. Psychol. [Hum. Percept.]*, 1980, 6(2), 309–320.
18. 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. Inform. Syst., 25(2)*, 2007.
19. R. Kohavi, R. Longbotham, D. Sommerfield, and R.M. Henne. Controlled experiments on the Web: Survey and practical guide. *Data Mining and Knowledge Discovery, 18(1)*, 2009, 140–181.
20. J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. *Proc. SIGIR '09*, 43–50.
21. C. Liu and C. Chung. Detecting mouse movement with repeated visit patterns for retrieving noticed knowledge components on web pages. *IEICE Trans. Inform. & Syst.*, 2007, E90-D(10), 1687–1696.
22. L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay. The influence of task and gender on search and evaluation behavior using Google. *Inform. Process. Manage.*, *42(4)*, 2006, 1123–1131.
23. F. Mueller and A. Lockerd. Cheese: Tracking mouse movement activity on websites, a tool for user modeling. *Ext. Abstracts CHI '01*, 279–280.
24. G. Pass, A. Chowdhury, and C. Torgeson. 2006. A picture of search. *Proc. InfoScale '06*, 1.
25. K. Rodden and X. Fu. Exploring how mouse movements relate to eye movements on web search results pages. *Workshop on Web Information Seeking and Interaction at SIGIR '07*, 29–32.
26. K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. *Ext. Abstracts CHI '08*, 2997–3002.
27. B. Shapira, M. Taieb-Maimon, and A. Moskowitz. Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. *Proc. SAC '06*, 1118–1119.
28. K. Wang, T. Walker, and Z. Zheng. PSkip: Estimating relevance ranking quality from web search clickthrough data. *Proc. KDD '09*, 1355–1364.