Characterizing Multi-Click Search Behavior and the Risks and Opportunities of Changing Results during Use

Chia-Jung Lee
University of Massachusetts
Amherst, MA, USA
cjlee@cs.umass.edu

Jaime Teevan
Microsoft
Redmond, WA, USA
teevan@microsoft.com

Sebastian de la Chica
Microsoft
Redmond, WA, USA
sedelach@microsoft.com

ABSTRACT
Although searchers often click on more than one result following a query, little is known about how they interact with search results after their first click. Using large scale query log analysis, we characterize what people do when they return to a result page after having visited an initial result. We find that the initial click provides insight into the searcher's subsequent behavior, with short initial dwell times suggesting more future interaction and later clicks occurring close in rank to the first. Although users think of a search result list as static, when people return to a result list following a click there is the opportunity for the list to change, potentially providing additional relevant content. Such change, however, can be confusing, leading to increased abandonment and slower subsequent clicks. We explore the risks and opportunities of changing search results during use, observing, for example, that when results change above a user's initial click that user is less likely to find new content, whereas changes below correlate with increased subsequent interaction. Our results can be used to improve people's search experience during the course of a single query by seamlessly providing new, more relevant content as the user interacts with a search result page, helping them find what they are looking for without having to issue a new query.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Search process*.

Keywords
Log analysis, web search, result list change, search dynamics.

1. INTRODUCTION
Many queries involve multiple search result clicks. Previous research found that for queries where at least one search result was clicked, 46% contained additional clicks [12]. Since 2003 this number appears to have grown, with multi-click queries now representing 59% of the clicked queries in our dataset. There are many reasons why searchers might click multiple results. They may, for example, have trouble finding what they are looking for and thus visit multiple results with limited success, or they may have a need that cannot be satisfied by a single result [8].

Despite the prevalence of multiple clicks after a query, much of the research to understand search result interaction has focused on user behavior surrounding the first click. An accurate picture of behavior when many results are clicked could positively impact everything from the retrieval models people build to how search engines are evaluated. This paper presents the first study that we are aware of that characterizes what people do within a single query when they return to a search result page after having visited an initial result. We find that people's initial interaction with the result page, including what they click and how long they dwell, significantly impacts their future interactions with those results.

While searchers think of search result lists as static, the results for a single query actually often change over time [14, 19], even after very short intervals such as the time between when a user visits a clicked search result and when that user returns to identify a second result to click. Previous research has shown that change interferes with a person's ability to interact with the results during repeat queries [25]. This paper looks at the impact of short term search result change on user behavior during multi-click queries. We find that when changes occur during the course of a single query, they interfere with the searcher's ability to find new information, leading to increased abandonment and slower clicks.

However, search result change presents not just a risk, but also an opportunity for the search engine to provide new, more relevant information without additional input from the user. While changes to a search result list sometimes happens as a result of unexpected instability (e.g., concurrent indexing [14]), change can also happen as the result of intentionally designed features. For example, it is not always possible for a search engine to identify the most relevant content immediately after a query is issued. The implicit feedback users provide as they search can be incorporated in real time to produce a better ranking [15, 27, 28]. New content may also become available as the web changes [1], or search engines may want to take more than a few hundred milliseconds to process complex queries [26]. The ability to provide some results initially, and then change the results as new information becomes available could enable search engines to significantly and seamlessly improve the search experience. In our analysis we observe that there are cases where change leads to greater satisfaction, and explore one way to take advantage of this opportunity to positively impact millions of users.

The goal of this paper is to provide in-depth picture of the relationship between the first and subsequent clicks following a query, with a focus on instances where the result page changes in between. After a discussion of related work, we describe the approach we use to understand multi-click behavior. We then characterize how people interact with results after their first click, and look at the impact of change on this behavior. We conclude with a

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

HTTP://DX.DOI.ORG/10.1145/2600428.2609588

discussion of how our findings can be used to provide new content as a person searches, and implement and explore an example.

2. RELATED WORK

Understanding people's search behavior is critical to improving the search experience, and substantial effort has been invested to this end. Large scale search logs provide an invaluable picture that can be used to estimate things like search success and result relevance. Previous research has formalized the task of relevance learning as a click modeling problem using assumptions about general user behavior. Position bias (where top ranked documents attract more attention, even when less relevant) is a well-known example [12]. The examination hypothesis [6] suggests that a document is relevant only if it is examined and clicked. Based on this hypothesis, many extensions have developed sophisticated ways to model user behavior [5, 29]. It is often assumed that users examine results from top to bottom without skipping, as is the case with the cascade model [6]. While this assumption is generally effective, people's search behavior can be more complicated. This motivates subsequent work [4, 10, 11] that emphasizes modeling multiple clicks after a single query. Recent studies go beyond the examination hypothesis to consider the impact of lower ranked documents on clicks [22].

While existing research shows how characterizations of user behavior can be used to improve search, very little is understood about multi-click behavior beyond the fact that it exists. Previous research suggests 46% of clicked queries have more than one search result click [12]. Despite the fact that multi-click behavior is not well understood, it has been explored as a way to improve ranking and relevance estimates. As a form of implicit feedback, Agichtein et al. [2] used post-click behavior (among other features) to optimize the ranker effectiveness, and Zhong et al. [30] proposed a Bayesian approach to incorporate post-click features for estimating document relevance. These models have been shown to improve document ranking, but relatively little can be learned from them about how and why users interact with search results and how search result presentation impacts user behavior.

A particularly unique aspect of our analysis is that we look at what happens when search results change as the user interacts with them. Although users are typically not aware of it, results regularly change [14, 19]. Change can occur unintentionally, as an artifact of server side variation, such as the contents of the cache or which back end index is hit [14]. It can also reflect changes to the underlying web [1], with new content identified and changes to existing content impacting ranking. Or it can be a result of personalization and contextualization, with new, more relevant content identified via implicit relevance feedback and provided to the users as they interact with search results. For example, SurfCanyon dynamically updates the search result list as users interact with it [15], and White et al. explored providing users with a dynamic list of relevant sentences as they searched [27].

Researchers of human-computer interaction have long known that interface instability can cause problems, even when change seems beneficial. For example, dynamic menus were developed to help people access menu items more quickly than traditional menus by bubbling commonly accessed items to the top. Rather than decreasing access time, however, dynamic menus slow users down as commonly sought items no longer appear where expected [21]. Similar problems result from instability for search results [17]. For example, White et al. [27] tried to help people search by dynamically re-ranking lists of relevant sentences using implicit feedback, and found that people did not perform as well with a dynamic list as they did when it was static. This is probably because, as Selberg and Etzioni [19] state, "Unstable search engine results are counter-intuitive for the average user, leading to potential confusion and frustration when trying to reproduce the results of previous searches." Teevan et al. [25] present the only log study that we are aware of on the impact of search result change on user behavior. They find that searchers take significantly longer to click on a repeat search result following change. We extend this work by providing a detailed look at interaction with results that change within a single query, rather than across sessions.

Because searchers may value having new information presented in the course of a single query, several solutions have tried to address the fact that dynamic search results can cause problems. For example, SurfCanyon attempts to avoid confusion while providing real time implicit relevance feedback by highlighting new results in a separate location [15]. However, this approach calls out changes at a time when most users are merely focused on finding what they are looking for. The Re:Search Engine tries to avoid this by merging new content into an existing search result list [24]. While the solution shows promise, it has only been studied in a laboratory setting on a small scale. In addition to using log analysis to characterize how people interact with change, we propose providing users with new content when they return to a result page while maintaining stability in the results they have seen, and run a preliminary test of this approach with millions of users.

In summary, the work presented in this paper extends existing models of search result interaction by focusing specifically on how people behave when they return to a search result page after their first search result click. Because search results can change, even as they are being used, we look carefully at the impact of such changes, and suggest and test one potential solution.

3. APPROACH

We now describe how the analysis presented in this paper was performed. We describe the dataset, formalize the problem, and define the measures we study, including behavior-based measures of search success and measures of search result change.

3.1 Dataset

To understand how people interact with search results for multi-click queries, we analyzed logs collected by Microsoft's Bing search engine. We sampled two months of log data from 2012 for users in the United States English language locale. The sample was filtered to remove bots, spam, and outliers (e.g., queries followed by more than 20 clicks or result lists with more than 14 search results). We also removed adult and navigational queries because these query types have very unique behavior following the first click. Navigational queries, for example, are typically followed by only one click. For each query, the logs contain information about when the query was issued and the URL and rank of any clicked results. Using this we extracted the subset of instances where a query was issued, a result was clicked, and the user returned to the result page following the click. The resulting dataset contains 8,863,684 queries and 17,154,920 clicks from 1,658,931 distinct users.

The dataset also includes 17,727,368 impressions of the results displayed to the user, half representing the result list seen before clicking, and half representing the list seen when returning from an initial click. The results provided by all major search engines change over time [14, 19], and change can occur even during the course of a single query. In such cases the results shown after the user returns to a result page following a click are different from those displayed prior. Changes may arise from instability (including the dynamic nature of the web and the complex architectures of search engines [14]) or be intentional. For example, in the logs we analyze in this paper Bing displayed, by design, eight results to the user following their initial query, and twelve when the user returned. In each case, the results were ranked according the best information available to the search engine at that instant. Changes like those observed in our logs are prevalent for top search engines [14, 19], although most users are not aware of them.

3.2 Problem Specification

Using this dataset, we characterize the behavior of users who clicked a search result following a query, visited the clicked result for some period of time, and then returned to the search result page. When the search result page is presented to the user for a second time, we study a wide range of characteristics of the interactions that the users might have with it.

To formulate the problem, we specify a general scenario tuple T:

$$T = \ < (Q,\ S_1, C_{S_1}),\ (\delta_{S_1 \to S_2}) >$$

The first part of the tuple, $(Q,\ S_1, C_{S_1})$, represents the fact that a user issued a query Q, viewed a search result page S_1 in response to Q, and then clicked the result C_{S_1}. Our dataset further requires that the user then used the back button or some other means (e.g., the refresh button) to return to the search result page. When this happens, a search result page for Q is once again presented to the user, which we call S_2. While S_1 and S_2 may be the same, there can also be differences between them. These differences are denoted as the second part of the tuple, $\delta_{S_1 \to S_2}$. All subsequent user interactions with the results for Q other than the first click occur with S_2 since after the first click Bing instructs the browser to keep the search result page in its cache.

Given the scenario tuple T, we characterize user behavior with the second search result page (S_2) by asking:

1. Which factors from T impact user behavior with S_2 and how? Which lead to the most user satisfaction?
2. How do changes to the result page ($\delta_{S_1 \to S_2}$) impact users? Do certain changes improve or degrade overall user experience?

In Section 4 we address the first question, using the first part of the scenario tuple to study the impact of the user's issued query and initial click on their subsequent behavior. In Section 5 we address the second question, using the second part of the tuple to characterize how the system appears to influence a user's interaction with S_2 by changing or holding stable the search result list.

3.3 Defining Measures

In our analysis we use several behavior-based measures of search success and search result change. Standard statistical analysis including confidence interval and z-test were conducted on these measures where appropriate and when the two means derived from two populations were compared.

3.3.1 Measures of Search Success

We look at four common measures of search success: the number of clicks, click satisfaction, click position, and time to click.

Number of clicks One way to understand a user's search experience is to look at the number of clicks that a user makes following a query. In particular, researchers have explored this using *abandonment*, which is a measure of how likely a searcher is to not click on any result at all following a query. While abandonment can indicate a positive search interaction when the searcher finds what they are looking for directly within a search result page [16], the absence of a click is generally taken as an indication that the user has failed to find relevant content [4].

$$p(Abandon) = \frac{\#(\{queries, abandon\})}{\#(\{queries\})}$$

All of the queries in our sample are filtered to have at least one click, meaning abandonment from the first result page, S_1, is zero. Because we are interested in users' interactions with S_2 when they return following a click, we study abandonment of S_2. This behavior is unlikely to indicate a positive experience because any relevant inline content was probably seen prior to the initial click.

Satisfaction Another way to measure a user's success during a search is to look at time spent on the visited result pages. Previous research has found that implicit signals such as clicks, time, and end user action are good predictors of satisfaction [8], with a dwell time of 30 seconds on a result commonly used to indicate satisfaction. We refer to clicks with dwell times of 30 seconds or longer as *SAT clicks*, and those with shorter dwell times as *NSAT clicks*. Due to the limitation of event-based logging, it is impossible to calculate the dwell time of the last click in a search session because there is no subsequent event. For precision, we only consider clicks where the dwell time can be calculated accurately.

A little over half (59%) of the initial clicks on S_1 in our subsample are SAT clicks. Occasionally, for simplicity, we will refer to users who had a SAT initial click (C_{S_1}) as *SAT users*, since they come into our analysis of S_2 already satisfied, and users who had a bad start (or an NSAT initial click) as *NSAT users*. Regardless of their initial experience, users who return to a result list after an initial click are probably trying to find additional relevant information. To measure how often this happens successfully, we look at whether any subsequent clicks for that query are SAT clicks. If at least one click is a SAT click, we call it a *SAT return*. The ratio of the number of SAT returns to the number of *NSAT returns* provides a picture of how many more times the return experience is satisfactory versus unsatisfactory. The larger the value, the more satisfied users are. We refer to this *satisfaction ratio* as R_{SAT}:

$$R_{SAT} = \frac{\#(\{queries, Return_{SAT}\})}{\#(\{queries, Return_{NSAT}\})}$$

Click Position The position of a clicked result can also provide insight into the search experience. Typically, the higher a result click is in a search result list, the better the list is considered to be [13]. People also trust that relevant results are highly ranked, and thus have a positional bias towards clicking higher [9]. Consistent with this previous work, the first clicks for a majority (51%) of the queries in our sample are on the first result. Commonly understood assumptions about click position, however, do not necessarily hold true for the second and subsequent clicks, since the user is likely to be oriented to the search results in a different way.

All users in our sample clicked one initial result on S_1, and zero or more results on S_2. We define three possible changes in click position between two consecutive clicks: the user can click higher in a result list than they originally did (*Up*), click the same position in both lists (*Stay*), or click lower the second time (*Down*). Thus the set of possible click patterns is $C = \{Up, Stay, Down\}$. The position of the initial click affects the possible subsequent behavior. For example, *Up* clicks can only occur when the first click is not on the first result in S_1, and *Down* can only occur when the initial click is not the lowest ranked result in S_2. We measure the probability of each pattern occurring in the logs:

$$p(C = c) = \frac{\#(\{queries, c\})}{\#(\{queries\})}, c \in C$$

Time to Click The time between when a user is first presented with a search result list and when that user clicks a result for the

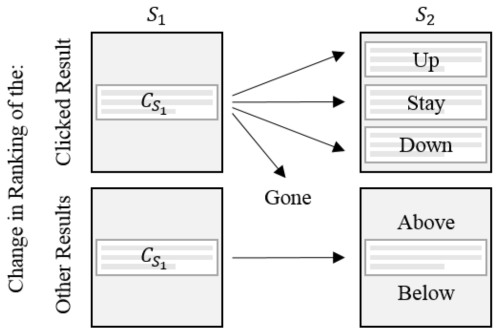

Figure 1. The types of search change studied, including instances where the clicked result changes rank (top) and where it remains static but the results around it change (bottom).

first time provides an indication of how hard it is to locate a result to click. We refer to the time it takes a user to make their first click as t_{S_1}, since it occurs on S_1, and the time it takes to make their second click as t_{S_2}, since it represents their first click on S_2.

3.3.2 Measures of Search Result List Change

In addition to changes in search result interaction before and after the first click, we also measure changes to the results presented. We study the impact of two different types of change, illustrated in Figure 1. In the first (top of Figure 1) we look at instances where the position of the search result initially clicked changes between when the result page is first presented (S_1) and when it is next presented (S_2). In the second (bottom of Figure 1), we look at instances where the other results on the result page change given the initial click remained in the same position. We parameterize $\delta_{S_1 \to S_2}$ in terms of the two result pages displayed to the user for the query Q (S_1 and S_2) and the initial click C_{S_1}.

Change in Ranking of the Clicked Result A positional change of the first result clicked can be important, as previous research suggests that people are particularly likely to remember where the results they clicked appear in a search result list [23]. When a user clicks on a search result and then returns to the result list to look for additional results, the result that was clicked initially (C_{S_1}) can appear in the same place (which we refer to as *Stay*), or be ranked closer to the top (which we refer to as *Up*) or lower (*Down*). It can also disappear entirely (*Gone*). These changes are represented as the top set of changes in Figure 1. All of these types of change occurred in our dataset; for 62% of the queries we sampled, the initially clicked result stayed in the same position, for 14% it moved up, for 23% it moved down, and for 1% it disappeared.

Change in Ranking of the Other Results We also look at the impact of changes that occur between S_1 and S_2 that do not impact the position of the clicked result. These are represented in the bottom part of Figure 1. To avoid confounding such changes with changes to the initially click result's position, we only look at the 62% instances where the clicked result stays in exactly the same place (referred to as *Stay* earlier). The other results can either change above that click (*Above*) or below it (*Below*). We consider a change to have occurred if any results occur in a different position from where they occurred in S_1 or do not appear at all. Eighteen percent of the queries had a change above and 94% below. Because changes to the result immediately preceding or following the initial click may be particularly noticeable, we also look specifically at changes to these (referred to as *Above1* or *Below1*).

The position of the initial click affects the types of changes that can be observed. For example, the initial click is very likely to be on the first search result, and in these cases it is impossible for the search results above where the click occurred to change. Likewise, if the initial click is on the last search result, results below can only change if new results are added to the result list. We use the entire dataset when analyzing the effect of changes in position to the initial click, but only use the instances that can be defined for analysis for changes to the rest of the search result page.

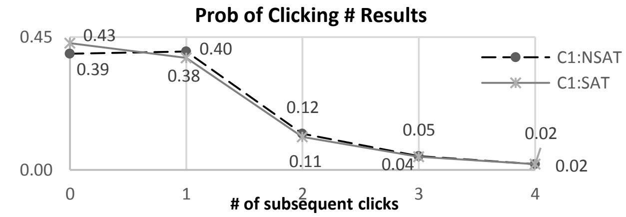

Figure 2. The probability a user will click a certain number of results on S_2 after an initial click. Users who were satisfied with their initial click are less likely to click again.

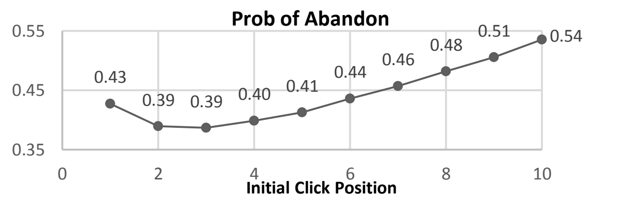

Figure 3. The probability of abandoning upon returning as a function of different initial click positions. The lower the initial click, the more likely the user will not click again.

4. INTERACTION AFTER A CLICK

We begin our analysis by looking at how characteristics of a user's initial interaction with a result list relates to their subsequent interactions. Regardless of any changes that may have occurred to the result ranking, we explore how many results people click when they return, whether the results they click satisfy their need, the position of their clicks, and the time it takes to make a click.

4.1 Number of Clicks

The median number of results users clicked following their initial click was one, meaning the median number of clicks for queries in our sample was two. This is higher than typically observed in the literature because we only analyzed queries that had at least one click and for which the user had returned to the search result page following that click. Abandonment of S_2 (or clicking no results upon returning) indicates that the user has failed to find new relevant content. The probability of abandoning the search result following the first click but after returning to the result list was 0.41.

Users who were satisfied with their initial click seemed to put less effort into finding additional relevant results upon returning. Figure 2 plots the probability that users would click a certain number of results on S_2, broken down by whether the initial click (C_{S_1}) was satisfactory (SAT) or not (NSAT). SAT users abandoned with a probability of 0.43, or 11% more than NSAT users (0.39), with 99% confidence interval (CI) from 0.0391 to 0.0409 for the difference between the two means. Users who were unsatisfied initially also tended to click more results than SAT users. The median number of clicks for NSAT users was one, and zero for SAT users. This may imply that satisfied users have significantly less ($p < .0001$) motivation to find additional relevant content because they already found something useful.

Subsequent click behavior also appears to be impacted by the position of the initial click, as can be seen in Figure 3. The proba-

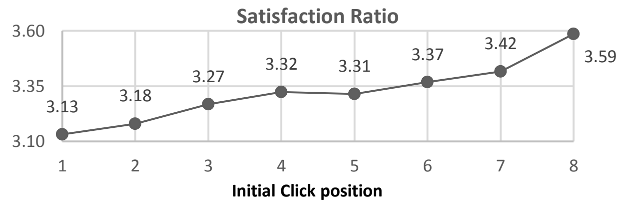

Figure 4. Satisfaction ratio given initial click position. Users who click lower are more likely to be satisfied when returning.

bility of abandoning S_2 generally grows as the first click moves lower, from 0.39 when the second result is the first clicked to 0.54 when the tenth result is. The difference in abandonment between position 2 and 10 for SAT and NSAT users is significantly different with a p-value < .0001 and 99% CI (0.1396, 0.1604) for the two mean differences. A lower initial click is usually considered indicative of lower result quality, and this may be why users are less likely to find new relevant content upon returning. However, the probability of abandonment is also relatively higher (0.43) when the first click is on the first result. These queries may be ones where the user is particularly satisfied and thus is less likely to put significant effort into continuing their search.

4.2 Satisfaction

We also look at how satisfied users were with the results they found upon returning. Approximately 64% of the first clicks on S_2 (i.e., second clicks overall) were SAT clicks. However, any click on S_2 may result in satisfaction, not just the first. We observe that for 76% of queries that have subsequent clicks, at least one was a SAT click. As was the case for general click behavior, two factors emerge as being particularly likely to correlate with one or more SAT clicks on S_2: the user's satisfaction with their initial click, and that click's position in the result list. Regarding the first point, although users who were satisfied with their first click tended to be more likely to abandon the results after their first click, they also seemed to be satisfied with newly clicked results more often than NSAT users, assuming they clicked on something. Specifically, we find that SAT users had a satisfaction ratio (R_{SAT}) of 4.75, compared to 2.10 for NSAT users. In these cases the result quality may be high, or perhaps the user or task easily satisfied.

Secondly, although users were more likely to abandon their query when their initial click was ranked lower, they were also more likely to be satisfied by subsequent clicks. Figure 4 shows that lower initial click positions had higher satisfaction ratios. As we will show in Section 4.3, a lower initial click also tended to be followed by higher subsequent clicks, and it may be these higher ranked results were indeed more relevant even though they were initially skipped. Likewise, users who start out clicking high are more likely to follow up with lower ranked clicks, which are presumably less relevant and thus less likely to be satisfactory.

Click Pos Δ	All	Initial Click		Initial Click Position				
		NSAT	SAT	1	2	3	4	5
Up	18.6%	20.2%	17.4%	0.0%	26.7%	39.2%	43.7%	***47.1%***
Stay	16.0%	13.2%	18.1%	20.6%	14.8%	9.7%	9.2%	8.0%
Down	***65.4%***	***66.7%***	***64.5%***	***79.4%***	***58.6%***	***51.0%***	***47.1%***	44.9%

Table 1. The likelihood that the position of the second click was above, the same as, or below the first click. Users tend to move down the result list, except when the initial click was low.

4.3 Click Position

We now look at the position of the clicked results as a function of when the click occurred. Figure 5 shows the positional probability distribution for the first 10 clicks following a query, with Ci representing how likely the i^{th} click was to occur at each position. Note that only the first click (C1=C_{S_1}) was on S_1; all subsequent clicks (C2 to C10) took place on S_2. As expected, the first click is most likely to be on the first result. For the subsequent clicks, however, the peak of the i^{th} click is at position i. This is consistent with a general trend of progressing down the result set linearly, as observed previously via analysis of search [3] and gaze logs [7, 9, 13]. However, the top positions (positions 1 to 3) become relatively popular again compared to other positions among later clicks (see C6-10). It appears that users sometimes return to the beginning of the list after having actively clicked on many results.

Table 1 summarizes how a user's initial interactions with the result list impacted whether they clicked higher or lower in S_2 than in S_1. As suggested by the progression of clicks in Figure 5, users were most likely to move down the result list after returning to it, with 65.4% of all second clicks being lower than the first. However, 16.0% of all users clicked at the same position (sometimes on a new result, if the initial result changes rank), and 18.6% clicked higher. For subsequent pairs of consecutive clicks, when they occurred, there was a tendency for $p(c = Up)$ to increase and $p(c = Down)$ to decrease, especially for later clicks.

When this analysis is broken down by whether the user was satisfied by their initial click or not, we see that satisfied users were 1.37 times more likely to stay on the same result ($p(c = Stay)$) than NSAT users (18.1% v. 12.2%, p < .0001). It appears short-time re-finding behavior [25] is particularly common for satisfied users, perhaps because the clicked result is indeed the current best for the user. We also see in Table 1 that the initial position of the first click affects the landing position of the second click. Users starting with a lower initial click (e.g., at position 5) choose to click higher results more often than lower results. While clicking results from top to bottom is usually believed to be more natural, a reverse click order (i.e., $c = Up$) may indicate a difficult search.

We also analyzed how far users moved in the result list between clicks. Follow-up clicks tended to occur close in rank to the previous click, with a median distance of one between two consecutive

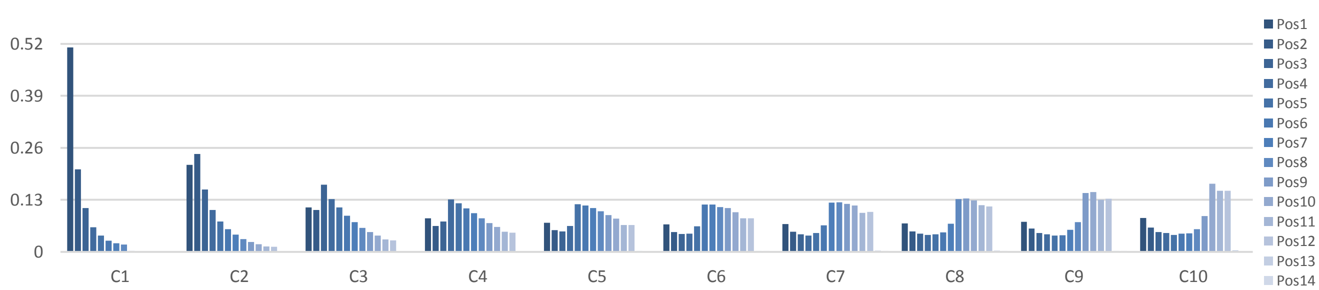

Figure 5. The click position distribution for the first through tenth clicks. C1 (or C_{S_1}) is the initial click on S_1 and the rest are on S_2. The i^{th} click is most likely to be on the i^{th} result.

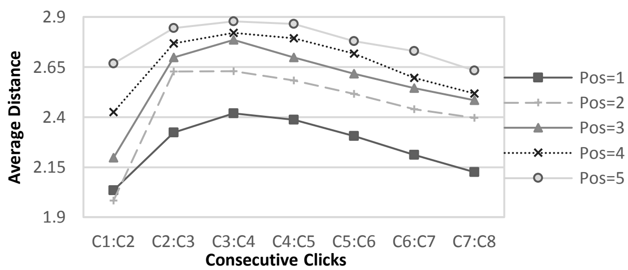

Figure 6. The distance between two consecutive clicks for different initial positions. C*i*:C*j* represent the *i*th and *j*th click. Users take larger steps following lower initial clicks.

clicks. However, clicks that occurred low in the result list were more likely to be followed by a large move in position. Figure 6 shows the average absolute value of the positional difference between consecutive clicks, with different lines representing different initial click positions. It may be that hard queries, where users click low ranked results, cause users to expend more effort scanning for relevant results. We also observe that, as reflected by the distance between two clicks, users appear to search conservatively in the beginning, jumping around in the middle, and finally focus for later clicks. This is consistent across different initial positions.

4.4 Time to Click

The time it takes for a user to click a result tells us how quickly that user is able find what they are looking for. We observe that it took longer to make the first click (t_{S_1} = 18.31 seconds, median 9.66) than the second (t_{S_2} = 13.86, median 3.04). This suggests that users learned something about the search result list during their initial interaction. Consistent with this, we observe that users who spent more time inspecting the search result list prior to their first click were more likely to make their second click relatively faster than their first. Figure 7 shows t_{S_2} as a function of t_{S_1}. In general (All), t_{S_2} grows with t_{S_1} but at a slower speed.

We further break the data down by the user's satisfaction with their initial click. Users took 28.6% longer to make their first click when they found a satisfactory result than an unsatisfactory one (median 10.75 seconds v. 8.36). It may be that spending more time reading prior to clicking helps users find better results. It may also be that these are slower users in general, and it takes more time for them to both click and return. As shown in Figure 7, initially SAT users also spent more time reading results before clicking a second time. This may be part of the reason why these users are also usually more satisfied with their subsequent clicks.

Next we look at the impact of the initial click positions on the time to the second click, shown in Figure 8 for initial click positions of one through five. Not surprisingly, the time to first click is usually longer for lower initial positions, as users at least need to locate the result before clicking it. For example, the median of t_{S_1} for Pos=1 is 6.75 seconds and for Pos=5 is 19.23 seconds. However, a higher initial click also seems to lead to a longer time to click the second result, perhaps because users have spent less time inspecting the other results in the list. The trend reverses when t_{S_1} is longer than approximately 14 seconds. These may represent hard queries where the user inspects all results before clicking.

4.5 Summary of Interaction after a Click

We showed that a user's interaction with a result page following a click is strongly influenced by features of their initial interaction. Users who appeared satisfied with the first result they found were less likely to identify new content to visit, but more likely to be

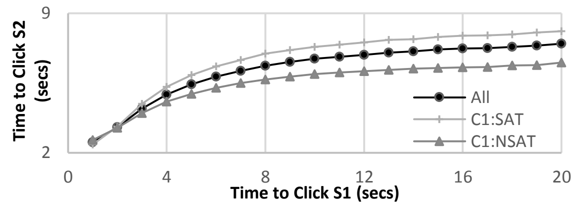

Figure 7. Time to click S_2 as a function of time to click S_1. The second click tends to occur faster, particularly for NSAT users.

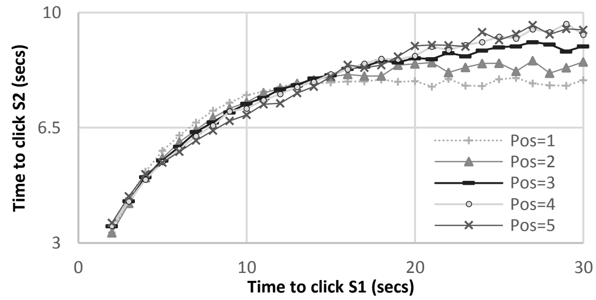

Figure 8. Time to click S_2 as a function of time to click S_1. Users who click on a top ranked result quickly need more time to make their second click than those who click lower.

satisfied with the new content if they did. We confirm that users in general click results from top to bottom of a ranked list, but observe that top ranked positions regain their relative popularity for later clicks. The user's initial click position can affect the distance between consecutive clicks, with lower initial clicks resulting in a larger gap between two consecutive clicks. People's second click tended to occur faster than the first, and satisfied users usually spent more time reading results before clicking.

5. WHEN SEARCH RESULTS CHANGE

In addition to changes in search result interaction before and after the first click, there can also be changes to the underlying search results that are presented. We now look at different types of change correlate with post-click behavior.

5.1 Number of Clicks

We begin by looking at the relationship of change to the number of clicks a user made following their initial click, focusing on abandonment, and observe that change often preceded high abandonment. Table 2 shows the probability of abandonment as a function of whether the initial click, C_{S_1}, moved up in the result list, stayed in the same place, moved down, or disappeared entire-

Initial Click	Up	Stay	Down	Gone
All	0.432	0.425	0.375	***0.492***
NSAT	0.414	0.401	0.359	***0.483***
SAT	0.445	0.441	0.386	***0.498***

Table 2. The probability of abandonment by whether the result initially clicked moved up, down, stayed in the same place, or disappeared. Users abandoned most when it disappeared.

Initial	Above		Above1		Below1		Below	
Click	Δ	Static	Δ	Static	Δ	Static	Δ	Static
All	***0.414***	0.396	***0.409***	0.397	0.395	***0.422***	0.423	***0.431***
NSAT	***0.399***	0.374	***0.391***	0.374	0.379	***0.398***	0.397	***0.411***
SAT	***0.425***	0.413	***0.423***	0.414	0.406	***0.438***	0.439	***0.445***

Table 3. The probability of abandonment for changes in the search result list above or below initial click, given the clicked result's position remains the same. Users abandon more when results above change, and less when results below change.

Initial Click	Up	Stay	Down	Gone
NSAT	2.00	2.08	2.20	***2.31***
SAT	4.65	***4.78***	4.75	4.61

Table 4. The satisfaction ratio by whether the clicked result moved up, stayed, moved down, or disappeared. While SAT users like it to remain static, NSAT users prefer it removed.

Initial Click	Above		Above1		Below1		Below	
	Δ	Static	Δ	Static	Δ	Static	Δ	Static
NSAT	***2.30***	2.21	***2.25***	2.22	***2.16***	2.07	***2.09***	1.99
SAT	***4.93***	***4.93***	4.85	***4.93***	***4.85***	4.77	***4.79***	4.61

Table 5. The satisfaction ratio for changes in the result list above or below the first click. Users tend to be more satisfied when results change, although users who were satisfied with the first result they found want the results above it to remain static.

ly. The probability of abandonment is particularly high when C_{S_1} disappeared from the list. For example, 0.492 for *Gone* is significantly higher than 0.425 for *Stay* with a p-value < .0001 and 99% CI (0.0628, 0.0712). It may be that when users return to a result page they expect to see the link they followed as a colored link, and its absence could be confusing. On the other hand, if C_{S_1} is ranked lower in S_2 than it was in S_1, we observe a lower abandonment probability than if it stayed in the same place. These trends are consistent when behavior was partitioned by whether the user was satisfied with the initial result they found or not.

We also looked at the abandonment rate when the result list changed but the initially clicked result remained static (Table 3). Changing results above the initial click (Above) led to higher abandonment, while changing results below the initial click (Below) led to lower abandonment. This was also true when the result immediately before (Above1) or after (Below1) the clicked result changed. All pairwise comparisons between different groups of users suggest that change in result ranking significantly impact user response (p < .0001). As we have seen that people often progress through the result page, it may be that changing results that have already been viewed causes confusion. Changing the results below the initial click, on the other hand, appears beneficial. We observe similar behavior when breaking the data down by users' initial satisfaction.

5.2 Satisfaction

The increase in abandonment following change suggests most change can interfere with a search, particularly when the clicked result disappears or change occurs high in the result list. However, when we look at user satisfaction, we see that change can sometimes help the user find new relevant content, particularly if they were unable to find what they were looking for initially. Table 4 shows the satisfaction ratio when changes occur to the ranking of the initial click. For users who were not satisfied with their initial click, moving an NSAT result up the list correlated with the least subsequent satisfaction, while removing it correlated with the highest. Promoting an unsatisfying result harms the user experience, while moving it down or removing it improves the user experience as long as the user does not abandon the search. In contrast, users who found a result that satisfied them on their first click were most likely to be satisfied on subsequent clicks if the result stayed in the same place. Consistent with what we observed for abandonment, removing the link a satisfied user liked provided the worst experience.

Table 5 shows the satisfaction ratio when the search result list changed but the initially clicked result remained in the same place. As we saw with abandonment, result changes that happened below the initial click appear to be beneficial, resulting in a higher

Change in Position of Click	Change in Rank of Initially Clicked Result			
	Up	Stay	Down	Gone
Up	37.64%	15.02%	16.71%	32.51%
Stay	17.25%	14.63%	18.86%	16.08%
Down	***45.12%***	***70.35%***	***64.44%***	***51.41%***

Table 6. The change in click position as a function of how the rank of the first clicked result changes. Users tend to progress down the page, but are more likely to move up the page when the results moves up or disappears.

Change in Position of Click	Above		Below	
	Δ	Static	Δ	Static
Up	46.26%	38.65%	14.78%	16.58%
Stay	4.66%	6.07%	14.53%	13.62%
Down	***49.07%***	***55.28%***	***70.69%***	***69.80%***

Table 7. The change in click position as a function of how the search results changed around the initial click. Users tend to progress down the page, but are more likely to move up when there has been change above their initial click.

satisfaction ratio for both SAT and NSAT users. However, unlike abandonment (where users were more likely to abandon if the results above the click changed), the satisfaction ratio for users not satisfied with their initial click went up even with changes above their click. Change high in the result list may help unsatisfied searchers, as long as they do not abandon their search.

5.3 Click Position

In addition to relating to whether users abandon or find relevant content upon returning to a result page, change also correlates with where users focus their attention, as evidenced by how changes impact the position of their second click. When the result initially clicked moves up or the results above the click change we see that the next click occurs higher in the list. Table 6 shows how the position of a user's second click changes compared to the position of the first given a change in rank of the initially clicked result. Consistent with what we observed in Section 4.3, users are most likely to progress down the result page with their clicks when the initially clicked result remains in the same position or moves down in rank. However, if the initially clicked result moves up in rank or disappears from the list, the user is significantly (p < .0001) more likely to click higher in the result list for the second click (37.64%) than if the result stayed in the same place (15.02%) or moved down (16.71%). It may be that users orient themselves around their previous click while progressing through a result page.

We also analyzed how changes above and below the clicked result relate to changes in the position of the user's subsequent clicks (Table 7). Changes above the initial click appeared to attract users' attention, significantly (p < .0001) increasing the likelihood they clicked results above (46.26%) compared to when those results remained static (38.65%). Changes below the initial click, on the other hand, had a smaller impact on click position. Results for Above1 and Below1 are very similar and thus omitted.

5.4 Time to Click

Our analysis also suggests that changing results may slow users down as they try to find new results to click upon returning. To illustrate this, we plot the time to the second click (t_{S_2}) as a function of the time to the first click (t_{S_1}) in Figure 9, broken down by whether the user's initial click changed rank or stayed in the same position. In all cases, people were able to click a second result fastest when their search result was shown in the same place. Consistent with what we observed in our earlier analyses, when

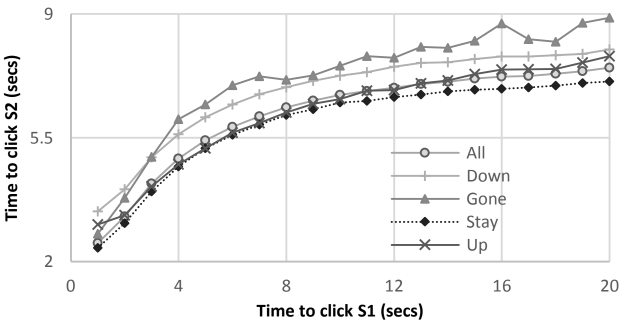

Figure 9. Time to click S_2 as a function of time to click S_1, broken down by how the first clicked result moved. The second click is fastest if it is in the same position.

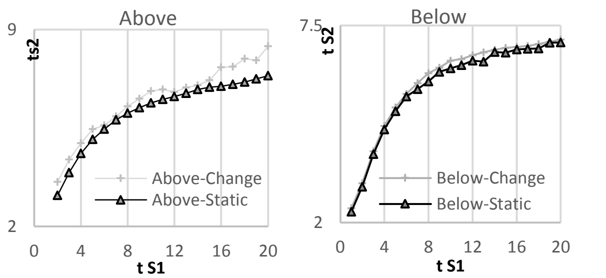

Figure 10. Time to click S_2 as a function of time to click S_1, broken down by whether the results above or below the initial click changed. The second click is fastest when the list is static.

the result disappears entirely it correlated with a particularly large delay. However, while a clicked result moving up suggests increased abandonment and decreased satisfaction, it also delays the second click the least. It may be that since people tend to remember clicked results as having been ranked higher than they were [23], result lists with this type of change appear very similar.

Figure 10 shows the same graph for instances where the initial click remained in the same position and the results above and below changed. In both cases change delayed t_{S_2} as compared to the static case, although changing the results below did so to a lesser extent. This is consistent with our previous findings that suggest lower changes are less disruptive.

5.5 Summary of When Results Change

We have seen conflicting evidence as to whether search result change during a single query benefits or harms the user experience, suggesting there are both risks and opportunities to providing dynamic results. For users who were satisfied with their initial click, changing the result page appears primarily to cause harm; these users mostly preferred the page to be static, except for changes to results below their initial click. In contrast, users who were not satisfied by their initial click appear more likely to benefit from change. Even seemingly significant search result changes, such as the removal of the first clicked result, sometimes improved these users' satisfaction. Although change is potentially beneficial under certain conditions, it should be introduced carefully because along with the increased satisfaction often comes the risk of increased abandonment and time to click.

6. DISCUSSION

The analysis presented in this paper paints a rich picture of multi-click search behavior, particularly in the context of search result change. These findings can be used to model user behavior better and improve the search experience. In general, we observe that relevance alone is not the only criteria that should be considered by search engines for multi-click queries. Instead, search engines must account for the user's previous experience with the search results. In this section, we explore ways search engines might do this to build an accurate picture of multi-click behavior, support fast comprehension of the search result list, keep users engaged after their first click, and introduce new, relevant content in a seamless manner over the course of a single query. We then present a simple example that highlights the potential for these findings to positively impact millions of users.

A number of the measures we study could be valuable for modeling user behavior and improving ranking and relevance evaluation for multi-click queries. For example, models that assume a linear progression through the result list [6] appear to be roughly accurate for the first few clicks, but could be improved to assume an increased likelihood of returning to earlier results for later clicks. Rankers could also use more complex features of a searcher's initial interactions with a result page than previously explored [2, 30], such as click rank and dwell time, to optimize their estimates of document relevance.

Accounting for people's initial interactions with a search result page is important because people use what they learn during their first encounter when they return. We observed, for example, that second clicks were typically faster than first clicks. To help users quickly understand search content, a search engine could offer a summary or visual representation of the results. It could also help orient users in the result page when they return, marking visited content and highlighting important changes, doing for a single query what Qvarfordt et al. explored for a session [18]. Different approaches could be used to keep users engaged as a function of their initial experience. For example, users who spend only a short time on the search result page initially may not have constructed a rich picture of the results and thus need more orientation support than consistency upon returning. Likewise, users who spend only a short time visiting a search result may not want that search result stressed or promoted when they return but rather want new content drawn to their attention. On the other hand, users who are satisfied with the first results they find are less likely to continue clicking when they return. A search engine could instead provide these users with query suggestions or information related to the clicked result to support further exploration on the same topic.

Our results also reveal an opportunity to provide new, relevant content to searchers when they return to a result list. Thus far efforts to contextualize results have focused on using information from the initial queries in a session to improve the ranking of subsequent queries [20]. Our findings suggest it may also be possible to identify new content for users without their ever having to issue a new query. This could be done using implicit feedback from the user's initial interactions (e.g., dwell time and click position), or by taking more than a few hundred milliseconds to process the initial query [26]. However, there appears to be a risk to capitalizing on this opportunity, in that changing the result ranking during a search may cause confusion. When ranking results mid-query, a search engine must account for the user's initial experiences. Clicked results appear to be used for orientation, and thus should probably be included in subsequent result lists instead of displaced by new, more relevant results. The most relevant new content should not naively be ranked at the top of the list, but instead placed where the user will attend to it (e.g., immediately below the previously clicked result). Satisfied users appear less tolerant of change, so the largest changes should be reserved for when a user's initial experience is unsatisfactory.

List Δ	Above		Above1		Below1		Below	
Type	Δ	Static	Δ	Static	Δ	Static	Δ	Static
Orig	***0.41***	0.40	***0.41***	0.40	0.39	***0.42***	0.42	***0.43***
Stable	***0.45***	0.40	***0.44***	0.40	***0.44***	0.42	0.41	***0.48***

Table 8. The probability of abandonment for changes in search results above or below initial click in the original dynamic dataset and the stabilized dataset.

6.1 Example: Stability with New Results

As an example of how a search engine might use the analysis presented in this paper to intentionally provide new content during a multi-click query in a seamless manner, we conducted an initial exploration into a simple approach. We describe the approach we implemented, and discuss what we learned about the complexities of controlling change in the process.

6.1.1 Approach

In our analysis up to this point, we have studied search results that change during the course of a query without taking into account the user's initial interactions with the results. While some of this change was the result of uncontrolled instability, the search engine also chose, by design, to display a longer result list when a user returned. Due to changing conditions, new results often appeared early in the list and the initial result ordering changed. Given the importance of stability, we implemented an approach that continued to display the same new content but in a stabilized manner. The results from S_1 were held constant when a user returned to S_2, and four to six new results were appended at the end of the list. We hypothesized these new results would be seen if needed but not disrupt the user's search experience. We refer to this method as the *stabilized* approach, and discuss it in the context of the *original*, completely dynamic approach.

Using the same approach to data collection described in Section 3.1, we collected log data for 9,883,375 queries with stabilized results from a two month period in the year 2013. As before, the dataset was restricted to instances where users had clicked a result following their query and returned to the result list. However, because of the enforced stability, the result clicked initially rarely moved; only 6% of the time in total did it move up, down, or disappear, as compared to 76% of the time in our earlier analysis. Although we aimed for 100% stability among the results from S_1, this was not always possible due to factors outside of our control related to operating in a large scale production environment. Given the initial click remained in the same position, results changed above that click 4% of the time (again, due to factors outside of our control) and below 99% of the time (by design).

Six months separate the collection of the stabilized and original datasets. Changes in user base, underlying ranker, task, and even season can impact user behavior, so the two datasets are not directly comparable. Here we provide some initial observations of the differences in how user behavior correlates with change within each individual dataset, but we are unable to directly compare the two datasets. The goal of this discussion is to provide preliminary insight into the expected and unexpected ways that stabilization changed the user experience to the extent possible.

6.1.2 Observations

Our analysis of the new dataset suggests that showing new content in a controlled manner may benefit users for multi-click queries. However, our observations of how people interact with the stabilized results also highlight an unexpected edge case where change appears particularly confusing: users whose first click is on the last result appear to have a particularly unsuccessful return experience if new results were appended below the clicked result.

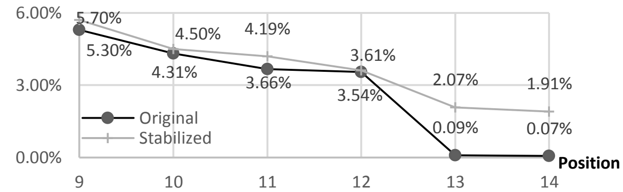

Figure 11. The position distribution of the subsequent clicks on S_2 (focused on positions 9 to 14) in original and stabilized datasets given results changed below initial click.

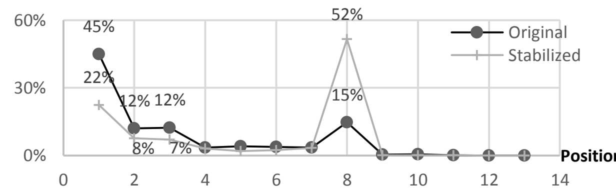

Figure 12. The position distribution of the initial click on S_1 in original and stabilized datasets given results changed one below initial click.

In general, we observe that users abandon the re-ranked search result list less (with a 2.5% drop in the probability of abandonment) and find new content more (with a 3.6% increase in the satisfaction ratio) in the stabilized dataset than in the original dataset. Table 8 shows the probability of abandonment of the stabilized dataset broken down by whether the results changed above the initial click or below. As a point of reference, it also includes the probability of abandonment of the original dataset. For both the original and stabilized rankings, users were more likely to abandon their search after the first click when the results above that click changed compared with when they remained static. This seemed particularly true for the stabilized experience. Users were 13% more likely to abandon the query when the results above the click changed compared to when they were static in the stabilized condition, and only 3% more likely to abandon when the results changed in the original condition. It may be that users resisted change more when most results were the same. Because changes above the click were rare in the stabilized case, the increase in abandonment had minimal impact on overall abandonment.

Table 8 also shows that for both datasets the probability of abandonment is lower when the results below the initial click changed (Below). There is a 15% decrease in abandonment when results below change in the stabilized condition, and only a 3% decrease in the original condition. We further observe that users were more likely to click on the appended results in the stabilized case than they were to click on the low ranked results when the initial results were not held static, as shown in Figure 11.

However, the most noticeable change with the stabilized dataset is that there was a negative impact when the result immediately below the clicked result changed (Below1). There is a 5% increase in abandonment when the immediate result changed in the stabilized condition, whereas in the original condition changing in fact helped reduce abandonment by 8%. In the stabilized dataset, change immediately below the user's initial click primarily occurred when users clicked the result at position eight (i.e., the last result of S_1), as shown in Figure 12. In the original, more dynamic dataset, instances of change immediately below the initial click were more widely distributed. This negative impact thus may be because users who click the last result in the list are surprised to see additional content appended below that result.

This is but a simple initial exploration into how the controlled introduction of new content might positively impact the user experience during a single query. Our findings suggest that stabilizing results can have a positive impact, but may also make some types of change more detrimental. Given this and our earlier analysis, we believe there are many further opportunities to contextually enforce stability or provide new content while people search.

7. CONCLUSION

This paper explored how a person interacts with the search result page after their initial click. By analyzing the Bing query logs, we showed that a user's initial search result click can provide important insight into that user's subsequent interactions with the result page. For example, a short initial dwell time correlates with increased future interaction, perhaps because someone who does not find what they are looking for is more motivated to look for results in the following steps. On the other hand, if a user appears satisfied with their initial click but returns to the result page regardless, they are usually happier with their subsequent clicks than others. We confirmed that users tend to move down a result list as they search, but observed that top positions can regain popularity as a search progresses. We also saw that searchers are generally faster when selecting the second result to click than the first, but can take longer if they only spent a short time inspecting the result list prior to their first click.

Although search engine users think of query results as static, when a searcher returns to a search result following a click there is an opportunity for the results to change. Such changes may hinder a user's ability to find what they are looking for, as reflected by an increased abandonment probability. However, some changes may enhance overall satisfaction if the user does not abandon the search task. Behavioral responses to change vary based on the user's initial experience with the result list. Initially satisfied users react positively to minimal change, while users who failed to locate a good result initially benefit more from changes. Although altering search results may sometimes be helpful, it appears that users have to spend extra time adjusting to the new content.

We discussed several ways these results could be used, and explored one way these results can be used to provide new content during a single query by maintaining a static search result list and appending additional results at the end. We found that this invites clicks on the appended results but highlights challenges when change does occur. We discuss ways these findings can be used, including proactively adjusting results for users who are frustrated by their initially clicked result while maintaining stability for others. Our results can be used to improve people's search experience during a single query by providing new, more relevant content as the user interacts with a search result page, allowing users to find what they are looking without having to issue a new query.

8. REFERENCES

[1] Adar, E., Teevan, J., Dumais, S. T. & Elsas, J. L. The web changes everything: Understanding the dynamics of web content. *WSDM 2009.*

[2] Agichtein, E., Brill, E. & Dumais, S. T. Improving web search ranking by incorporating user behavior information. *SIGIR 2006.*

[3] Baeza-Yates, R., Hurtado, C., Mendoza, M. & Dupret, G. Modeling user search behavior. *Latin American Web Conference 2005.*

[4] Chapelle, O. & Zhang, Y. A dynamic Bayesian network click model for web search ranking. *WWW 2009.*

[5] Chen, W., Wang, D., Zhang, Y., Chen, Z., Singla, A & Yang, Q. A noise-aware click model for web search. *WWW 2012.*

[6] Craswell, N., Zoeter, O., Taylor, M. & Ramsey, B. An experimental comparison of click position-bias models. *WSDM 2008.*

[7] Dumais, S. T., Buscher, G. & Cutrell, E. Individual differences in gaze patterns for web search. *IIiX 2010*.

[8] Fox, S., Karnawat, K., Mydland, M., Dumais, S.T. & White, T. Evaluating implicit measures to improve web search. *TOIS*, 23(2), 2005.

[9] Granka, L. A., Joachims, T. & Gay, G. Eye-tracking analysis of user behavior in WWW search. *SIGIR 2004.*

[10] Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y.-M. & Faloutsos, C. Click chain model in web search. *WWW 2009.*

[11] Guo, F., Liu, C. & Wang, Y.-M. Efficient multiple-click models in web search. *WSDM 2009.*

[12] Jansen, B. J. & Spink, A. An analysis of web information seeking and use: documents retrieved versus documents viewed. *ICOMP 2003*.

[13] Joachims, T., Granka, L. A., Pan, B., Hembrooke, H. & Gay, G. Accurately interpreting click-through data as implicit feedback. *SIGIR 2005.*

[14] Kim, J. & Carvalho, V. R. An analysis of time-instability in web search results. *ECIR 2011.*

[15] Kim, J. Y., Cramer, M., Teevan, J. & Lagun, D. Understanding how people interact with web search results that change in real-time using implicit feedback. *CIKM 2013.*

[16] Li, J., Huffman, S. & Tokuda, A. Good Abandonment in mobile and PC internet search. *SIGIR 2009.*

[17] Obendorf, H., Weinreich, H., Herder, E. & Mayer, M. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. *CHI 2007*.

[18] Qvarfordt, P., Golovchinsky, G., Dunnigan, T. et al. E. Looking ahead: Query preview in exploratory search. *SIGIR 2013*.

[19] Selberg, E. & Etzioni, O. On the instability of web search engines. *RIAO 2000*.

[20] Shokouhi, M., White, R. Bennett, P. et al. Fighting search engine amnesia: Reranking repeated results. *SIGIR 2013.*

[21] Somberg, B. L. A comparison of rule-based and positionally constant arrangements of computer menu items. *CHI 1987.*

[22] Srikant, R., Basu, S., Wang, N. & Pregibon, D. User browsing models: Relevance versus examination. *KDD 2010.*

[23] Teevan, J. How people recall, recognize, and reuse search results. *TOIS*, 26(4), 2008.

[24] Teevan, J. The Re:Search Engine: Helping people return to information on the web. *UIST 2007*.

[25] Teevan, J., Adar, E., Jones, R. & Potts, M. Information re-retrieval: Repeat queries in Yahoo's logs. *SIGIR 2007.*

[26] Teevan, J., Collins-Thompson, K., White, R. W., Dumais, S. T. & Kim, Y. Slow search: Information retrieval without time constraints. *HCIR 2013*.

[27] White, R. W., Ruthven, I. & Jose, J. M. Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. *SIGIR 2002.*

[28] White, R. W., Ruthven, I., Jose, J. M. & van Rijsbergen, C. J. Evaluating implicit feedback models using searcher simulations. *TOIS*, 23(3), 2005.

[29] Zhang, Y., Chen, W., Wang, D. & Yang, Q. User-click modeling for understanding and predicting search-behavior. *KDD 2011*.

[30] Zhong, F., Wang, D., Wang, G., Chen, W., Zhang, Y., Chen, Z. & Wang, H. Incorporating post-click behaviors into a click model. *SIGIR 2010.*