Leaving So Soon? Understanding and Predicting Web Search Abandonment Rationales

Abdigani Diriye[1], Ryen W. White[2], Georg Buscher[2], and Susan T. Dumais[2]
[1]University College London Interaction Centre, University College London, UK, WC1E 6BT
[2]Microsoft Corporation, One Microsoft Way, Redmond, WA, USA 98052
a.diriye@ucl.ac.uk, {ryenw, georgbu, sdumais}@microsoft.com

ABSTRACT
Users of search engines often abandon their searches. Despite the high frequency of Web search abandonment and its importance to Web search engines, little is known about *why* searchers abandon beyond that it can be for good or bad reasons. In this paper, we extend previous work by studying search abandonment using both a retrospective survey and an in-situ method that captures abandonment rationales at abandonment time. We show that although satisfaction is a common motivator for abandonment, one-in-five abandonment instances does not relate to satisfaction. We also studied the automatic prediction of the underlying reason for observed abandonment. We used features of the query and the results, interaction with the result page (e.g., cursor movements, scrolling, clicks), and the full search session. We show that our classifiers can learn to accurately predict the reasons for observed search abandonment. Such accurate predictions help search providers estimate user satisfaction for queries without clicks, affording a more complete understanding of search engine performance.

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

Keywords
Web search abandonment; Abandonment rationales.

1. INTRODUCTION
Search engine result page (SERP) abandonment is a type of search abandonment that occurs frequently and happens when users do not click on any of the results returned for a query [22][29][30]. Since clicks are absent in abandoned queries, it is difficult to understand why searchers are abandoning. They may have obtained the information they sought directly on the SERP, they may be dissatisfied with the results and failed to find any results worth clicking, or there may be other reasons for abandonment (e.g., accidental closure of the browser window). Since research in this important area is limited, a comprehensive analysis of the reasons behind SERP abandonment is both timely and necessary.

To understand the reasons for SERP abandonment and to enable search engines to estimate search success, we need to be able to automatically determine abandonment rationales. Features of interaction mined from large-scale logs have been shown to be useful in understanding searchers' satisfaction with search results [1][5][11][16]. However when modeling SERP abandonment, the absence of search result clickthrough data means that this important signal cannot be used to make inferences regarding the cause of abandonment. We therefore need to study whether there is evidence in the query, the results returned, and/or search interaction behavior beyond hyperlink clicks that might help predict people's motivations for abandonment.

We extend previous work in this area in the following ways:

- We capture the reasons for abandonment in-situ from the abandoning searchers rather than from third-party judges who must estimate the search intent and abandonment rationale [22]. We show that the reasons for abandonment are more nuanced than previously understood.
- We capture all search-related behaviors for a larger number of users than in previous work [22][29][30]. The scale allows us to study abandonment across a broader range of different search intents and the detailed logging allows us to examine how the abandonment occurred and its relationship with other behaviors in the full search session.
- We develop and study classifiers to predict abandonment rationales given data gathered from SERP content and search interaction during abandonment. We show that these classifiers can accurately predict SERP abandonment rationales given features of the SERP as well as within-session interaction.

We began our studies of abandonment rationales by employing a restrospective survey, and elicited abandonment rationales directly from searchers based on their recall of recent abandonment events. We then used the survey findings to inform the design of a Web browser plugin that captured abandonment rationales in-situ, allowing us to obtain the reasons for SERP abandonment as it happened and at the same time gather data about SERP content and searcher interaction that may be useful for predicting why people abandoned. The plugin was deployed within Microsoft Corporation for a period of four weeks and was adopted by over 2,500 people. Using data collected by the plugin, we show that we can train classifiers capable of reliably predicting why a user abandoned a SERP given their interaction behavior on the SERPs plus their search flow during the session.

The remainder of this paper is structured as follows. Section 2 presents related work on abandonment and the use of interaction features (in particular mouse cursor behavior since it does not involve link clicks) to infer SERP satisfaction. Section 3 describes the findings of a retrospective survey that we used to elicit a range of abandonment rationales from users. Section 4 presents our study methodology including the browser plugin deployed to capture abandonment rationales in-situ and the data collection methods. Section 5 provides an overview of the explanations gathered in-situ, and the prediction experiments are described in Section 6. In Section 7 we discuss the findings of our study and their implications. We conclude in Section 8.

2. RELATED WORK
The click-centric nature of Web search has made the use of hyperlink clicks a popular approach for studying user search behavior and search goals. Traditionally, clickthrough data has been inter-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

preted as implicit user feedback about the relevance of search results. Clicks and dwell times on search result pages are often interpreted as a positive signal that the user is satisfied—to some degree—with the result. Clicks have been used to improve a variety of search-related tasks, e.g., to predict search success [11][16], compare alternative search algorithms [5], and learn ranking functions [25]. Conversely, the absence of clicks is interpreted as a negative signal of the quality of the results, and some efforts have been made to reduce query abandonment [28]. Gaze tracking provides a much richer understanding of patterns and sequence of user attention on search result pages [8][23], and might be useful in differentiating between good and bad abandonment. However, gaze tracking is difficult to instrument outside of laboratory settings and thus is impractical at Web scale.

More recently, researchers have used cursor movements as an alternative to gaze tracking. Several researchers have examined the relationship between eye gaze and cursor positions during search tasks [13][17][26]. Rodden et al. [26] identified a strong alignment between cursor and gaze positions. They also observed different types of cursor behaviors: (i) neglecting the cursor while reading, (ii) using the cursor as a reading aid to follow text (either horizontally or vertically), and (iii) using the cursor to mark interesting results. Guo and Agichtein [13] examined the relationship between search intent and cursor movements through a browser toolbar. In a follow-up study, they conducted two shopping tasks, and found that interaction features improved accuracy in discriminating between research and purchase intents [14].

Li et al. [22] were the first to distinguish between good and bad abandonment in search and the need to augment click behavior to understand abandonment. They defined good abandonment as an abandoned query for which the searcher's information need was successfully addressed by the SERP, without needing to clickthrough to additional pages. This can happen when the answer to a user's query is in the snippet text or, increasingly, when search engines provide specific kinds of answers to try to meet the user's information need (e.g., weather, flight status information, stock quotes, etc.). Li et al. compared abandonment for desktop and mobile search in three different locales (U.S., Japan, and China) and developed ground truth data using editorial judgments. They defined "potentially" good abandonment as queries that had a dominant information need that could be answered by a SERP, and "likely" good abandonment by examining such SERPs to see whether the answers were on the page. Likely good abandonment queries included those like [weather Oregon] or [1 USD in GBP] and had answers in either the result snippet or in a dedicated inline answer element on the SERP. Li et al. manually labeled potential good abandonment in small set of 400-1000 abandoned queries from the Google search engine. They did so by only considering the query and not the search results returned by Google. From that analysis, they estimated that 19-32% of abandoned Web searches conducted on desktop computers could be related to satisfaction (i.e., those queries that were classified as "yes" or "maybe" with respect to potential good abandonment).

Stamou and Efthimiadis studied searches without clicks by surveying a small sample of volunteer searchers [29][30]. In one study [29], they recruited 38 graduate students and asked them to complete an external survey (running in a separate Web browser window) for each Web search they performed in a single day. They identified two reasons why queries are abandoned: for intentional causes such as spell-checking, finding an answer or keeping abreast of the latest information; and unintentional causes such as irrelevant results, no search results, search was interrupted, and repetition of already-examined results. Stamou and Efthimiadis found that 13% of the study queries had no clicks, split evenly between intentional and unintentional causes. Analysis of participant responses revealed that the most common reason for unintentional abandonment was dissatisfaction with the search results, whereas the explanations provided for intentional abandonment were more evenly distributed between different explanations. However, since participants were not prompted to complete the survey, the distribution of explanations gathered may not be fully representative of all task types (e.g., participants may be less likely to remember the survey when engaged in intensive tasks).

In a follow-up study [30], Stamou and Efthimiadis address the lack of searcher prompting by using a browser plugin, deployed to a group of six participants. They studied two types of SERP inactivity: pre-determined (user planned on finding an answer on the SERP without clicking) and post-determined (user planned on visiting a result when they queried, but decided not to once the results were returned). Stamou and Efthimiadis found that just over a quarter of the queries (27%) were abandoned because of a predetermined intention, and that half of the post-determined abandoned queries were negatively affected by the information on the SERP, representing dissatisfied or bad abandonment.

Castillo et al. [6] highlighted the value of studying search behavior as a proxy to understanding how "tenacious" searchers were in finding inline answers for SERPs. Chilton and Teevan [7] studied repeated behaviors to understand abandoned SERPs containing inline answers. Huang et al. [17] recently described a scalable method for collecting cursor interaction patterns on SERPs. In one of their experiments they sought to distinguish good from bad abandonment. They examined one category of abandoned queries, namely short questions that contained answers in the result snippets on the SERP (what Li et al. [22] called the "answer" category). For these queries, Huang et al. found differences in cursor trail length, movement time, and cursor speed depending on whether the answer was present in the result snippets (good abandonment) or was not present in the result snippets (bad abandonment). Good abandonment was associated with shorter trails, less movement and slower cursor speed. However, they did not show whether cursor data can be applied to *predicting* whether abandonment is good or bad or applied to a wider range of query types. Beyond the SERP, White and Dumais [32] studied more extreme abandonment, where people switch away from the search engine.

We extend the previous work described in this section in a number of ways. First, we capture abandonment rationales in-situ from the abandoning users, rather than retrospectively via third-party judges. Second, we record more extensive log data than other studies across a greater number of users, providing access to broad range of abandonment intents and affording other analysis such as the nature of abandonment occurrences which have not previously been studied. Finally, we develop classifiers to accurately predict the reasons for abandonment and explore features that may be useful in distinguishing between abandonment types.

We begin by describing our initial explorations of the reasons for SERP abandonment, which informs the design of the in-situ survey and the analysis that we perform. Our first step was to distribute a survey asking people to recall their last abandonment episode and provide details of the motivation for it.

3. EXPLORING QUERY ABANDONMENT

To obtain an initial set of candidate explanations for why users abandoned their searches we used a retrospective survey.

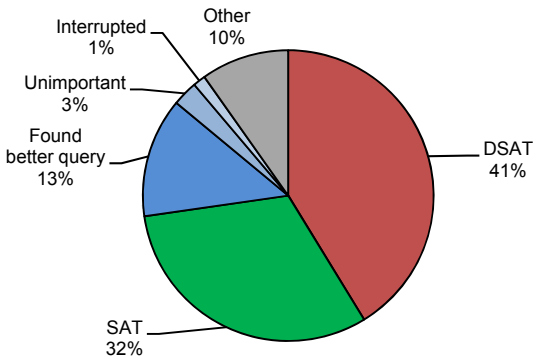

Figure 1. Reasons for SERP abandonment.

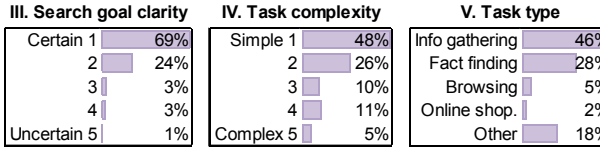

Figure 2. Survey responses for search goal clarity, task complexity, and task type.

3.1 Survey Methodology

An invitation to complete an online survey was distributed via email to a random sample of 3,000 employees from within Microsoft Corporation's campus in Redmond, WA. The sample comprised employees in a range of technical and non-technical roles. In completing the survey, respondents were asked to recall one recent example of a query they issued to a search engine where they did not click on any SERP hyperlink. Given this point of reference, the survey asked participants about:

I. Their motivation for not clicking on any link on the SERP. The reasons provided to participants were as follows:
   a. Dissatisfied with results (DSAT)
   b. Found the information that they sought directly on the SERP (SAT)
   c. Better query came to mind that more accurately represented their information need, before examining the search results
   d. Interrupted (e.g., by someone, by other task)
   e. Search was not sufficiently important
   f. Accidental (e.g., computer crashed, accidentally closed Web browser tab), and
   g. Other (participants were asked to specify). This option was included in case one of the reasons above did not adequately capture their rationale.

II. Their level of satisfaction with the search results (five-point scale ranging from *satisfied* to *dissatisfied*)

III. The clarity of their search goal (five-point scale ranging from *clear* to *unclear*)

IV. The complexity of their search task (five-point scale ranging from *simple* to *complex*), and

V. The kind of task they were performing (based on Kellar et al.'s goal classification [20]). Response options were:
   a. Information gathering
   b. Fact finding
   c. Browsing
   d. Online shopping, and
   e. Other (participants were asked to specify).

The nature of the search task has also been shown to effect SERP abandonment [30] and how searchers examine the SERP more generally [3][8]. We therefore requested information about task types in question V to better understand the nature of the tasks that people were performing when they abandoned SERPs.

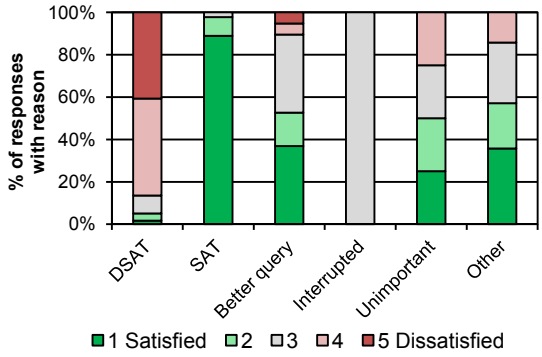

Figure 3. Satisfaction scores for different SERP abandonment reasons.

3.2 Retrospective Survey Results

Overall, we received responses from 186 survey participants. Figure 1 provides an overview of the reasons for abandonment provided by survey respondents. Our analysis of the responses showed that dissatisfaction (*DSAT*) with the search results returned by the search engine was the primary reason for SERP abandonment (41%). This was followed by satisfied abandonment (*SAT*) at 32% of responses. Interestingly, there was a large fraction of abandonment cases (27%) which the participants neither rated as clear SAT nor clear DSAT. Most of these were cases where the participants abandoned because they decided on a better query before they examined the search results (13%), sometimes they did not pursue the search any further because it was not important enough (3%), and rarely did they state that they got interrupted (1%). The *Other* category (10% of all responses) contained reasons with insufficient frequency to warrant their own category (e.g., "I figured it out for myself"). *Other* was also occasionally used even though there was a response option dedicated to the participants' explanation. For example, the *Accidental* category did not appear directly in any of our survey responses, but some of the responses for *Other* implied that the reason was accidental: "I lost network connection" or "I unintentionally closed the tab."

Participants' responses to questions III-V are summarized in the histograms in Figure 2. The survey revealed that 93% of the participants had a clear understanding of what they were looking for, 74% of the tasks were simple, and most of them (74%) involved information gathering or fact-finding tasks. These statistics align well with previous research on understanding user goals in Web search, independent of abandonment [20][27]. It appears that there is nothing remarkable about the types of tasks for which people abandon SERPs, at least in terms of the questions that we asked.

To better understand the relationship between abandonment rationales and participants' level of satisfaction with the search results provided the search engine we cross-tabulated their responses. Figure 3 provides a breakdown of user satisfaction (question II in the bulleted list of survey questions shown above) by the different reasons for abandonment. From the figure we can see that while dissatisfaction is most prominent for DSAT and satisfaction is most prominent for SAT, the remaining abandonment reasons are to a large extent characterized by neutral to positive satisfaction. When *Interrupted* was offered as a reason for abandonment (n=2), the level of satisfaction was entirely neutral, suggesting, as expected, that interruptions were not caused by the SERP.

The retrospective survey provided us with some insight into the reasons for query abandonment and the approximate frequencies with which each explanation happened (or at least could be recollected by participants). This information was useful to us in making decisions about which response options to offer in the plugin that we deployed to monitor abandonment rationales in-situ. We now describe the methodology that we used to monitor abandonment rationales and search interactions at abandonment time.

4. IN-SITU STUDY METHODOLOGY

In selecting an in-situ methodology for our study, we also considered a log-based analysis or lab-based experiments. The high degree of naturalism and ecological validity afforded by the plugin made it more attractive than the other methods. Log-based studies capture the behavior, but not the rationale for it. Lab-based studies may capture rationales, but may also expose people to artificial conditions and may lead to unnatural patterns of search behavior. Since one of the goals of this research is to develop an abandonment classifier that could be applied in real settings, capturing explanations in-situ in a natural setting was important for us.

To this end, we developed and deployed a Web browser plugin called *AbandonTracker* that surfaces a survey in a popup window to the searcher asking for an explanation whenever SERP abandonment is detected. In the remainder of this section, we describe the plugin, including important and transferable design decisions that we made, and its deployment within our organization. The first step is to define precisely what we regard as abandonment.

4.1 Definition of Abandonment

Intuitively, SERP abandonment occurs when the searcher does not click on any of the links on the SERP. Li et al. [22] provide a definition of abandonment by requiring "*a query that is not followed by any click or any further query within a 24-hour period.*" We agree broadly with this definition, but refine it in two important ways to consider: (i) the nature of the click and (ii) the nature of the trigger event (used to determine when and how the SERP abandonment occurs). Hence, we define abandonment in Web search as a situation where the following conditions are met:

1. **No clicks on results:** There are no hyperlink clicks on any results or advertisements, including results returned by the ranking algorithm and direct answers inserted into the results for topics such as news and weather. Note that we include advertisements in our definition of abandonment since we believe that they may also satisfy searcher needs in a similar way to search results. If a SERP click is on a related search, spelling suggestion, query alteration, or navigational link offered by the search engine (e.g., changing the scope of the search from "Web" to "Images") then we also regard that as abandonment since clicking on these links takes the user to another SERP.
2. **Trigger event occurs:** In addition to defining what counts as abandonment, we also need to define the point in time that the determination of no clicks (condition 1) should be made. To do this we define a set of abandonment *triggers* comprising one of the following actions:
   a. Manual requery: Explicitly issue a new query via a search box on the SERP, or a search box present in a Web browser or toolbar.
   b. Tab closed: Close the Web browser tab.
   c. Manual URL entry: Type URL directly into the Web browser address bar, and, thus, leaving the current page.
   d. Change search scope: Click on a navigation link to change to a different search vertical e.g., Web → Images.
   e. Click spelling suggestion

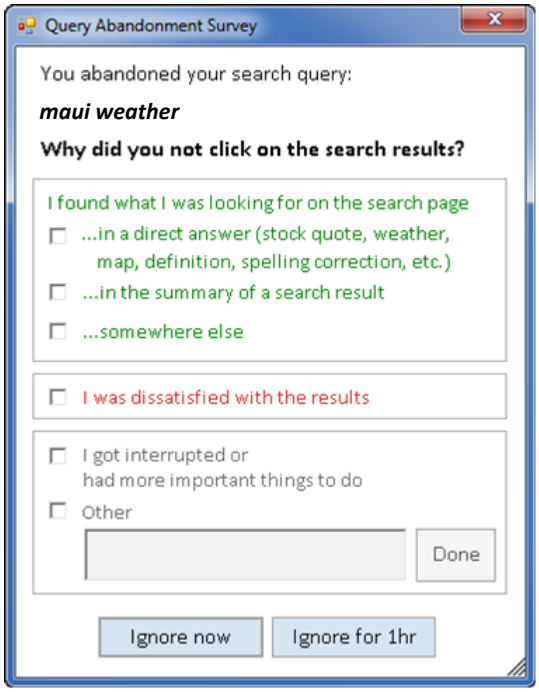

Figure 4. Screenshot of the AbandonTracker survey for the query [maui weather].

   f. Click query suggestions or query alterations: Click the related searches provided by the search engine or recourse links to reverse alterations made to the query by the search engine.
   g. Timeout: There are no clicks on the SERP for a period of 30 minutes from the time that the SERP loaded.

When the above two requirements have been met, the *AbandonTracker* system displays a popup survey, requesting that users indicate why they abandoned the search results. The survey appears on top of the SERP before the next Web page begins loading.

4.2 AbandonTracker Implementation

The AbandonTracker plugin was developed for the Internet Explorer Web browser. To remove the effect of variations in search engine quality and SERP layouts we focused on abandonment on a single search engine. In this subsection we describe the popup survey shown to participants and the data gathered by the plugin.

4.2.1 Popup Survey

Whenever a query is abandoned per our definition in the previous section a popup is shown on top of the browser window. See Figure 4 for an example of the popup survey for the query [maui weather]. The survey prompts the user to enter the reason for abandoning their query. It presents the abandoned query string to help the user identify the query that is being referred to (particularly useful in cases where many SERPs are being viewed in different browser tabs). The survey also offers four broad abandonment rationales from which the participant can select the appropriate response based on the findings from the retrospective survey. These are *SAT*, *DSAT*, *Interrupted or Unimportant*, and *Other*. Participants who selected the *Other* category could optionally specify their reason in the text area at the bottom of the survey.

Since the survey interrupted searchers directly with a popup, we wanted to keep it compact and easy to complete quickly. Therefore we elected not to include the option *a better query came to*

mind, since it could overlap with DSAT and require effort from users to distinguish between the explanations. Since all trigger events were automatically recorded (including manual requery), this explanation could be reconstructed using some combination of the manual requery trigger event, the time between the initial query and the requery, and by analyzing the *Other* reasons provided by users. Since we believed that it would be useful to know where on the SERP users found their information and how often they did so, we also offered multiple explanations for the source of the information leading to SAT abandonment.

In deploying the plugin to participants, we were concerned that given the frequency of SERP abandonment, the popup might appear too often, interrupting searchers from their primary task, and potentially irritating them to the point where they uninstall the plugin. We addressed this concern in three ways: Firstly, we introduced two "ignore" buttons in the survey, one to ignore the current instance of SERP abandonment, and another button to ignore all SERP abandonments for the next hour. Secondly, we implemented a trigger controller mechanism that suppressed the popup for 50% of all SERP abandonments on a per user basis. Thirdly, the popup survey would show up for a maximum frequency of 10 times per day per user to reduce the overall per-user effort.

4.2.2 Data Gathering

In addition to participant responses to the questions in the popup survey, AbandonTracker also records Web interaction data. For each user, it records the unique plugin identifier, all URLs that the participant visited, timestamps, unique identifiers for browser tabs and browser instances, and stores this information in a remote database. We also recorded the titles, the URLs, and the snippets of the top 10 search results, the presence/absence of other SERP features such as direct answers, and interactions with the SERP, including hyperlink clicks as appropriate.

In addition to the standard click logs of the search engine, we also captured a number of cursor actions on the result page using a scalable methodology similar to that described by Buscher et al. [4]. JavaScript-based logging functions were embedded into the HTML source code of the SERP. The x- and y-coordinates of the mouse were recorded every 250 milliseconds if the mouse had moved at least eight pixels (approximately half a line of text on the SERP) since it was last polled, non-hyperlink clicks (e.g., clicks in SERP whitespace to hide popups, right-clicks to print or refresh), scrolling, text selection, focus gain and loss events of the browser window, as well as bounding boxes of several areas of interest (AOIs) on the SERP (e.g., top and bottom search boxes, mainline results and its contained result entities) and the browser's viewport size. Combining these data sources enabled us to develop a rich picture of how searchers engaged with the SERP. Features were extracted from data gathered by the plugin as well as the search engine to build the predictive models described later.

4.3 AbandonTracker Deployment

The AbandonTracker was deployed to employees of a large technology company. Participants were recruited via an email invitation to a random sample of 5,000 employees. The invitation was sent to employees with a variety of backgrounds and job roles, ranging from software engineers to patent attorneys and administrators. In total, over 2,500 members of the organization installed the plugin. A number of steps were taken to ensure participants' privacy such as not storing personally identifiable information directly and not recording requests to intranet and secured servers. To motivate our participants to keep the plugin installed for the duration of the study, we randomly selected two participants per week of the study who had the plugin installed and awarded them each a 50 USD gift card. Awards were not tied to the amount of feedback that participants provided to avoid unduly biasing their search behavior with monetary incentives.

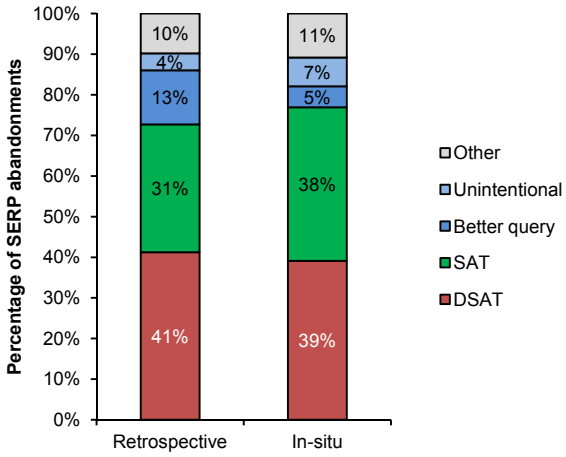

Figure 5. Distributions of the different reasons for query abandonment in retrospective and in-situ survey data.

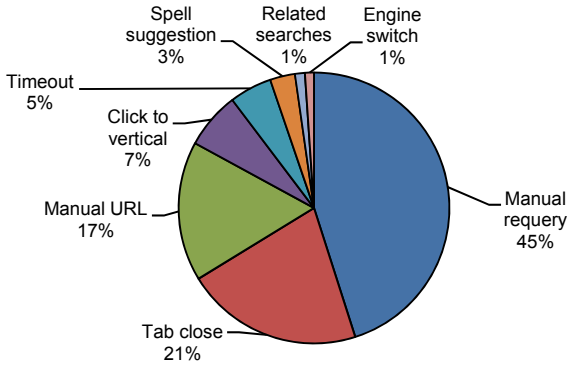

Figure 6. Abandonment trigger conditions showing how searchers abandoned their queries.

5. IN-SITU ABANDONMENT ANALYSIS

We now present an analysis of the data gathered by AbandonTracker. We discuss the characteristics of the data, describing overall usage statistics and characterizing the motivation behind observed abandonment instances and the ways in which people abandon (i.e., the *why* and the *how* of SERP abandonment). We also describe a comparison of the in-situ data with that gathered via the retrospective survey. No previous study of abandonment has explored these important issues in this depth or at this scale.

5.1 Overview of In-Situ Data

The study ran for 30 days in late 2011. We discarded data gathered on the first two days from our analysis since these were when we sent out the recruitment emails (and when 93% of our users installed the plugin), and we did not want initial testing of the plugin to affect data quality. We report results for 928 participants who provided feedback for at least one abandoned query. The other users either did not abandon, did not provide feedback, or did not use the search engine that we focus on. During the study, those participants visited 739,505 URLs in 172,887 distinct browser tabs, 39,606 of which were queries to the search engine we study. About 22% of these visited SERPs were abandoned per our definition of SERP abandonment provided in the previous

section. In half of these abandoned SERPs, the survey popup window was suppressed by the system. Of the remaining 50% of abandonment instances where the popup was shown, 59% of the popups were explicitly ignored by participants. The dataset was further processed to remove queries related to users testing the plugin functionality (e.g., [test], [hello world]). In the end, we gathered 1,799 abandonment instances that we analyzed further.

5.2 Explaining Abandonment In-Situ

Figure 5 (right) shows that there is a fairly even split between the *SAT* and *DSAT* reasons for abandonment. Findings showed that when satisfied, the answer to an abandoned query was eight times more likely to come from a dedicated inline answer on the SERP (such as weather, stock quote, etc.) than from result summaries. However, *SAT* and *DSAT* explanations still only occupied around 80% of the reasons for why people abandoned. The other 20% of queries deserves special attention since it has not been considered in previous studies. For simplicity we created a superclass called *Unintentional* comprising the *Interrupted*, *Unimportant*, and *Accidental* classes. This superclass comprised 7% of abandonment instances. Since we did not record the *better query* option directly on the in-situ survey, we reviewed the comments provided in response to *Other* where the trigger was manual requery and identified those corresponding to our definition of *better query* (e.g., "I was dissatisfied with my query", "didn't enter the right keywords"). Overall, 5% of the responses indicated that the abandonment rationale was related to the participant deciding to build a better query. The remaining queries were in the *Other* class and comprised reasons such as being directed to the incorrect engine vertical (e.g., "I wanted images"), seeking an intranet site, change of criteria, or undirected searching (e.g., "just vaguely browsing").

For comparison, we also show the distribution of explanations from the retrospective survey (Figure 5 left). Previous work on search engine switching [15] showed that there can be noteworthy differences in explanations gathered from in-situ and retrospective methods which might be suggestive of cognitive biases in the types of events that people recollect. We wanted to see whether the same trend was observed in our data. From Figure 5 we can see that the distributions of explanations are fairly similar. The main difference is in the fraction of SAT instances of abandonment (31% in the retrospective survey versus 38% in the in-situ plugin). Similar reductions in SAT for retrospective versus in-situ analysis have been observed in previous work of a similar nature [15] and may in part be related to people's tendency to more readily recall negative events in retrospective studies [21].

5.3 Abandonment Triggers

Understanding *how* people abandon (the so-called *trigger*) is important for applications of abandonment, since to model abandonment, we need to know how to capture it. Li et al. [22] set their trigger as no click or query in the 24 hours after the abandonment. However, given that people use search engines frequently (and often more than once per day), the 24-hour requirement is not sufficient for a complete analysis of abandonment; not to mention that they did not study *how* people abandoned SERPs. Figure 6 shows that there are a range of abandonment trigger conditions. Interestingly, the most frequent way in which searchers abandoned a SERP was by manually entering the new query in its search box or in the Web browser (45% of abandonments). Closing the Web browser tab in which the SERP was displayed (21%) or manually entering a URL in the Web browser's address bar (17%) were the main two other abandonment triggers. Manual URL entry and tab closure are both suggestive of task termination, whereas manual re-query may also reflect task continuation, depending on the nature of the query. The other trigger types are less popular (all 7% comprising of instances or less) and cover many different events with a range of possible explanations.

To better understand the relationship between explanations and triggers we study all explanation-trigger pairs. Figure 7 provides a breakdown of abandonment reasons by trigger conditions. Note that by definition, the better query explanation is only available for manual requery. The distribution of reasons differs depending on how the query was abandoned. A number of key insights can be made from these results. First, if the participant closed the Web browser tab, the session timed out, or they entered a URL, they were 2-3 times as likely to be satisfied as when manually re-querying. This suggests a link between behavior and abandonment rationales (e.g., tab closed→SAT, manual-requery→DSAT). We further explore the behavior-rationale relationship in the next section as we turn attention to predicting abandonment rationales using only the information a search engine can principally record.

6. PREDICTING SERP ABANDONMENT

Although prior work has explored some characteristics of abandonment in a limited capacity (and more limited than this study) [22][29][30], to our knowledge there has been no work on predicting explanations for observed abandonment instances. Using the 1,799 labeled abandonments from the in-situ study as ground truth, we built and tested classifiers to predict abandonment rationales. This section describes the features that we generated, the classifiers used, the evaluation metrics, the models that were compared in the study, the learning procedure, and the findings.

The four explanation labels from earlier were used corresponding to the main answer alternatives from the survey popup (Figure 4): *SAT*, *DSAT*, *Unintentional*, and *Other*. Although we hand-labeled the *better query* class, to avoid bias from this labeling we only used the original user feedback for prediction. We frame the prediction task as binary, predicting a single label vs. all other labels. We focus on binary prediction since it effectively models an application scenario that is likely to be of great interest to search providers (e.g., generate a list of DSAT abandonments for further inspection). We examine the features that are the most important in predicting abandonment rationales. We also study the features that distinguish SAT and DSAT abandonment since these provide useful insight into behavioral differences between the classes.

6.1 Feature Generation

Our predictions are made retrospectively once an instance of abandonment is observed. Around 2,000 features are generated for the abandoned SERP, the previous and next SERP, and the full session. We limited feature generation to the immediately preceding and succeeding SERPs both to maintain a manageable feature space and because those SERP interactions may provide the best insight about abandonment rationale for the current SERP (e.g., the next SERP may be for a refined query, suggesting DSAT).

For each abandoned SERP in our data, we extracted many features of the query, the associated SERP of interest, and the search session for the task of predicting the rationale for the observed SERP abandonment. The features were divided into five classes: (i) session, (ii) query, (iii) result, (iv) hyperlink-click and dwell, and (v) cursor. We now describe each class in more detail.

Search Session Features: Search sessions begin with a query to a search engine, and terminate following 30 minutes of inactivity between successive actions [31]. Session features for each query include the numbers of abandonments observed in the session, the total number of queries issued, whether the abandoned query was re-issued again in the session, the position of the abandoned query

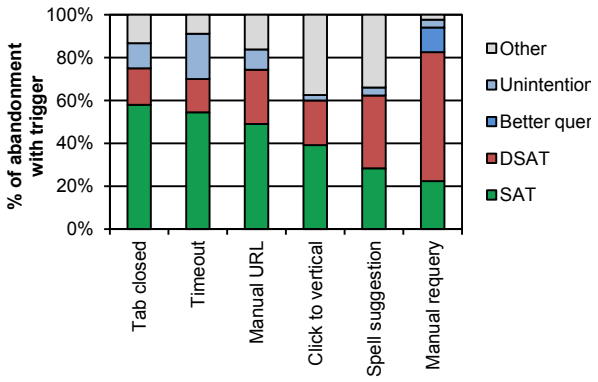

Figure 7. Abandonment reasons broken down by trigger conditions. Legend order = order in columns.

of interest with respect to all queries in the session and the position of the query of interest with respect to all abandonments.

Query Features: These include features of the query string itself (length in characters and length in tokens), and historic features such as the query frequency in the logs of a commercial search engine and the average query SERP result clickthrough rate. Historic feature values were derived from a year of search engine query-click logs from 2010. Other features included the similarity between successive queries in the search session, measured in different ways including overlap and cosine similarity.

Result Features: We computed features of the SERP returned for the query. These features included the number of results that were returned, the number of advertisements and their position on the SERP (to help capture whether the query had commercial intent), whether related queries or spelling suggestions were shown, the average length of result URLs, and whether there was any special treatment for the query such as special "direct answers" for queries with directly-serviceable needs (e.g., [weather in maui hawaii]). Since abandonment may occur because searchers find information in the titles and snippets of returned search results, we also computed features of the cosine similarity between the query string and each result title, and the cosine similarity between the query string and each result snippet. In addition to what was shown on the SERP, we also created features reflecting the match between the query and each of the top-ranked search results via the score assigned by the search engine's ranking algorithm.

Hyperlink Click and SERP Dwell Features (Engagement): Although abandoned SERPs by definition do not contain clicks on the results or advertisements, there may still be clicks on other regions of the page (e.g., related searches). We computed around 300 features of user clicks on various SERP components (e.g., the number of clicks on search results and the number of clicks on the search box) and overall dwell time on the result page. Note that when we compute these features for the previous and next SERPs they may include clicks on search results and advertisements.

Cursor Activity Features: Features of cursor interaction with the SERP can reveal patterns and preferences that cannot be observed through clickthrough behavior and can be captured at scale [17]. As described earlier, we captured cursor movements on the abandoned pages and the previous and post abandonment SERP. The cursor-related features that we computed included:

Trails: These features are derived from the cursor movement trails on the SERP and include trail length, trail speed, trail time, total number of cursor movements, and summary values (average, median, standard deviation) for single cursor movement distances and cursor dwell times in the same position, etc. We also created features for the total number of mouse movements and the total number of times that the cursor changed direction.

Hovers: We recorded total hover time on the SERP. Since we recorded the coordinates of areas of interest we were also able to associate cursor movements with particular SERP elements. This allowed us to represent the total hover time on inline answers (e.g., stock quotes, headlines, etc.), total time hovering in the lower and upper search boxes on the result page, in the left rail (where search support such as query suggestions and search history would usually be shown), in the right rail (where advertisements would usually be shown), and in the algorithmic results. We also computed the total time that the cursor was idle on the SERP.

Result Inspection Patterns and Reading: In a similar way to [3], this summarizes how users inspected the results, including the total number of search results that users hovered over, the average result hover position, and the fraction of the top ten results that were hovered. We also created features of the linearity of scanning over results and evidence of the user reading with the mouse.

Non-hyperlink Clicks: The total number of non-hyperlink clicks in various AOIs on the SERP, including the number of clicks in the upper and lower search box, the left and right rails, the algorithmic results, and across all SERP regions.

Scrolling: Including the total number of scroll events, the number of times the user scrolled up and down, the total scroll distance (in pixels), the maximum scroll height (in pixels) referring to the y-coordinate at the top of the viewport relative to the SERP, and time between SERP load and scrolling.

Other: Including whether the user clicked on the search box (suggesting that they were going to re-query), the number of text selections (total and unique results), and the number of hover previews (total and unique results). Hover previews are an interface feature that provides more information about a search result when requested by hovering on its caption.

6.2 Classifiers

We experimented with a variety of algorithms to predict abandonment rationales, using the feature sets described in the previous section. We found that multiple additive regression trees (MART) [12] was the best-performing classifier. Both L1 and L2 regularization were used to avoid overfitting predictions to the training set (90% of the sampled set determined via cross validation). L1 selects only effective features, and L2 penalizes extreme feature weights. The effectiveness of these two regularizations was demonstrated theoretically and empirically [24], showing that classification with regularization can be effective even if there are exponentially as many features as training examples.

6.3 Evaluation Metrics

In evaluating the performance of our predictions, we measure precision (the fraction of predicted instances that were correctly predicted) and recall (the fraction of all true instances that were correctly predicted). In this paper we report on the F_β measure, with β set to 0.5 rather than β=1.0. $F_{0.5}$ assigns twice as much weight to precision than to recall. High precision is very important in application scenarios for a predictor of abandonment rationales. In an online scenario, we would want to be highly confident before adapting the search experience based on abandonment rationale predictions. In an offline scenario, such as studying dissatisfaction in log data, we need to obtain a set of dissatisfied abandonment instances for further analysis. Since there are many abandonment events in logs, we do not need to classify all of them (have high

Table 1. Binary prediction performance ($F_{0.5}$) for each feature class and task. Significant differences from the marginal are marked using bold = $p < .05$, underlined bold = $p < .01$. For *SAT* and *DSAT*, significant differences from the Answer Presence baseline are marked with ○ = $p < .05$, ● = $p < .01$.

Feature Class	SAT	DSAT	Uninten.	Other
All	**0.6303**●	**0.7097**●	0.0472	**0.4516**
All.NoCursor	**0.6146**●	**0.6950**●	0.0401	**0.4217**
All.NoInteraction	**0.6169**●	**0.6723**●	0.0453	**0.2534**
Session	**0.4574**●	**0.5484**●	0.0754	0.1917
Click+Dwell	**0.5054**●	**0.6163**●	0.0508	**0.4193**
Cursor	**0.5390**	**0.6017**●	0.0618	0.1879
Result	**0.5843**●	**0.6557**●	0.0387	0.1894
Query	**0.5523**	**0.6137**●	0.0673	**0.2259**
Marginal	0.4322●	0.4440●	0.0828	0.1902
Answer Presence	0.5443	0.5124	n/a	n/a

Table 2. Top five features by evidential weight for the prediction of *SAT* vs. other classes and *DSAT* vs. other classes.

	Class	Features	Weight	r
SAT vs All	*Query*	CosineSimToNextQuery	1.000	–0.309
	Result	SERPHasAnswer@Pos1	0.683	+0.241
	Cursor	HoverTimeInTopSearchBox	0.479	–0.076
	Result	MinRankerScore	0.442	+0.187
	Cursor	TotalDwellTimeOnAOIs	0.410	+0.053
DSAT vs All	*Query*	CosineSimToNextQuery	1.000	+0.358
	Cursor	NonHyperlinkClickCount	0.565	+0.203
	Result	AvgRankerScore	0.550	–0.007
	Click+Dwell	TimeToClick_nextSERP	0.478	+0.014
	Cursor	ClickCountInTopSearchBox	0.474	+0.249

recall) as long as we can precisely label some. Note that we experimented with $F_{1.0}$ and the trends in results are the same as $F_{0.5}$.

6.4 Methods Compared

We compare the effectiveness of different feature sets for performing the predictions. We also trained the binary classifiers on varying sets of the features described in Section 6.1. In addition to analyzing the performance of each class individually, we also consider the following three feature combinations:

- **All:** Classifiers trained on all features.
- **All.NoInteraction:** Binary classifiers trained on all features except those that capture post-query interaction behavior on the SERP such as *Click+Dwell*, or *Cursor*. This helps us understand the importance of SERP interactions in predicting the reasons why people abandon. SERP interactions have been used in previous work to estimate satisfaction [1][10][13].
- **All.NoCursor:** Classifiers trained on all features other than *Cursor*. The cursor features represent a new source of behavioral information and we were particularly interested in their contribution to the overall prediction performance of the model.

We used two baselines in our experiments:

- **Marginal:** The marginal distribution across each label.
- **Answer Presence:** The presence or absence of direct answers on the result page. Direct answers are special elements more targeted at task completion / answering information needs on the SERP than normal algorithmic results [7]. Examples of such answers are weather reports and stock quotes. Many direct answers are designed to encourage satisfied abandonment, especially when presented at or near the top of the search result list. This baseline predicts SAT if the SERP has a direct answer element in the top-three result positions, and DSAT otherwise. This baseline is much stronger since it is based on an operational system (which others could replicate) and it is query-dependent. Note that we do not compute these values for the *Unintentional* or *Other* classes as there is no clear mapping between the presence or absence of answers and these labels.

6.5 Learning Procedure

Each of the models described in the previous section is used to generate a rationale prediction for each of the 1,799 observed abandonment instances in our set. Predictions are made at the end of the search session containing the abandoned SERP. This aligns with our application setting where predictions would be made retrospectively. This also lets us capitalize on session features. Ten-fold cross validation was performed to increase the reliability of the results over 10 randomized experimental runs. We now present findings on the prediction effectiveness using the different feature classes. We report averages over all runs and folds and present the results of statistical testing as appropriate.

6.6 Predicting Abandonment Explanations

We begin by analyzing the performance of the binary predictions for each class of SERP abandonment. The top row of Table 1 presents the $F_{0.5}$ values representing how effectively our binary classifiers can predict *SAT*, *DSAT*, *Unintentional*, and *Other* using all features. The next to bottom row of the table contains the Marginal baseline score for each class. The bottom row contains results for the Answer Presence (AP) baseline in the operational search engine. Significance values are also indicated using paired t-tests in comparison with the marginal and AP baseline. As can be seen from the table, AP outperforms the marginal for both *SAT* and *DSAT* predictions. Since it is stronger than the marginal, we use AP as the preferred baseline in the remainder of our analysis.

Turning attention to the performance of the classifiers that use *All* features, we see that our classifiers significantly outperform the baselines in predicting *SAT*, *DSAT*, and *Other*. However, there are no significant gains over the baseline *Unintentional*. One explanation is that unintentional abandonments are affected by external factors, such as distractions, loss of interest, or task shifts, which may not be predictable using the features of the SERP or searchers' interactions with it. The table also shows that it was easier to predict *DSAT* abandonment than *SAT* abandonment, perhaps because it may not be obvious that the user found the sought information directly on the SERP, especially if they do not interact in any way with it. Another reason is that *DSAT* abandonment is related to re-querying (as shown in Figure 7), providing a clear signal for prediction. We now study the impact of the feature classes.

6.6.1 Impact of Feature Classes

Table 1 also shows the breakdown of performance by the different feature classes and the two class combinations defined earlier. There is interesting variance in the feature classes that matter depending on the prediction task. For example, using only *Query* features yields a classifier that does not significantly outperform the AP baseline (Answer Presence: 0.5443 vs. Query: 0.5532, ns). Since the presence of a direct answer on the SERP is dependent on the query, much of the value from query features may already be encoded in AP. Cursor movements capture aspects of how people examine the SERP [17]. Given the range of possible cursor behaviors it is encouraging that *Cursor* features alone yield reasonable prediction performance, primarily for *DSAT* predictions, and for *SAT* and *DSAT*, adding *Cursor* to *All* helps.

Importantly, Table 1 shows that both *SAT* and *DSAT* can be accurately predicted using *All.NoInteraction*, which is accessible to search engines without needing further instrumentation. Focusing briefly on the *Other* classification task, we see in the last column of the table that it is features of hyperlink clicks and dwells that were most useful in predicting the *Other* class. Recall that the *Other* classification contained the *better query* subclass and was often used by participants to capture query dissatisfaction (e.g., typographical errors). Closer inspection of the key features showed that they were associated with an immediate query reformulation, rapid clickthrough on the SERP following the current (abandoned) one, or engagement with spelling corrections on the current SERP, both reflecting problems with the abandoned query.

6.6.2 Individual Feature Contributions

In addition to studying performance at the feature class level, it can also be informative to examine the individual features that contributed the most evidential weight to the classifications. We focus on the *SAT* and *DSAT* prediction tasks in the remainder of the paper both to simplify our analysis and given that these are important motivations for search providers. In Table 2 we present the top five most important features for the *SAT* and *DSAT* prediction. Descriptive names are used for each feature and the suffixes "prevSERP" or "nextSERP" are used to represent the features of the previous or next SERP respectively. For each feature we also present the normalized weight with respect to the most predictive feature, and the Pearson correlation coefficient (*r*) between each feature's values and the ground truth labels. The feature weights are returned as part of the MART classifier output and reflect the relative importance of the features in the classifications models that were generated. This helps determine the tendency of the feature value, e.g., the similarity to the next query positively correlates with *DSAT* but negatively correlates with *SAT*.

Table 2 shows that the features that were most important in predicting *SAT* abandonment were associated with similarity between successive queries, the presence of inline answers on the page (especially an direct answer at the top of the list of search results), the quality of the search results retrieved, and the degree of examination of the SERP, measured in terms of total time spent dwelling on AOIs. The features associated with *DSAT* were also associated with query reformulation, but this time the reformulation was positively related to *DSAT*. Interestingly, the presence of non-hyperlink clicks was strongly associated with *DSAT* abandonment. *NonHyperlinkClickCount* is the second most important feature for *DSAT*. Although these clicks can occur in any non-link SERP location many of the clicks occur in the search box as shown by the importance of the *ClickCountInTopSearchBox* feature. However, the fact that *NonHyperlinkClickCount* is more important than *ClickCountInTopSearchBox* suggests that non-hyperlink clicks in regions of the SERP other than the search box are also important in predicting *DSAT*. Result quality and the speed with which users clicked the search result on the follow-on SERP (if there was a click) were also strong predictors of *DSAT*, the latter perhaps suggesting that the search task is difficult with users needing to spend more time examining the search results.

So far we have shown that we can predict *SAT* and *DSAT* abandonment with good accuracy. However, since there was some similarity in the top-five features for *SAT* and *DSAT* in Table 1 (just different directionality), we wanted to better understand what features distinguished between these two key abandonment types.

6.7 Distinguishing SAT from DSAT

We used the labeled data and focused on the subset of abandonment instances that represented a *SAT* or *DSAT*, ignoring the others. In these experiments, *SAT* is the positive class. Using all features, we can effectively differentiate between the two abandonment types. Specifically, we achieved an $F_{0.5}$ score of 0.7847, significantly more than the AP baseline performance of 0.6133 ($t(99)=1.41$, $p=.016$). The features that best distinguish *SAT* from *DSAT* abandonment are in Table 3. The positive label was assigned to *SAT*, so the tendencies of the correlations are positive if they correlate with satisfaction and negative if they relate more to dissatisfaction. The features that matter most in this task are associated with the queries (similarity to next query and next query length) and results (result quality, presence of answers, and match between query and result snippets). Only two of the top-ranked features are associated with interaction / cursor movements. One of these are the number of non-hyperlink clicks on the SERP which seem to carry a signal from both search box clicks, i.e., reformulations, as well as further types of interaction on the SERP (e.g., clicks to restore focus to the page, clicks prior to scrolling). Interestingly, although answers to user requests may be found in captions for many queries, the match between the query and the caption is actually more closely associated with *DSAT*.

Table 3. Top ten features by evidential weight for distinguishing *SAT* from *DSAT* abandonment.

Class	*Features*	*Weight*	*r*
Query	PercentOverlapWithNextQuery	1.000	–0.436
Query	CosineSimToNextQuery	0.775	–0.421
Result	MinRankerScore	0.677	+0.259
Result	TotalNumAnswersShown	0.554	+0.077
Result	SERPHasAnswer@Pos1	0.470	+0.288
Result	AvgRankerScore	0.408	–0.052
Cursor	NonHyperlinkClickCount	0.403	–0.266
Result	MinCosSimNextQueryCaption	0.341	–0.366
Query	NextQueryLengthChars	0.335	–0.348
Cursor	StDevCursorMoveDistance	0.312	+0.102

7. DISCUSSION AND IMPLICATIONS

Although abandonment occurs often and is critical for measuring satisfaction, there has been very limited study of why it occurs. Our study is the first to address key issues in understanding abandonment, namely gathering rationales in-situ, characterizing the reasons that searchers abandon and their abandonment behavior at scale, and predicting abandonment rationales from search behavior. It is only through extensive studies and detailed analysis of this nature that we can truly understand search abandonment.

Not only did we observe the same satisfaction-oriented rationales identified in previous studies, but we also found that a significant fraction of abandonment (around 20%) does not fit under SAT or DSAT (e.g., formulating and issuing a better query before inspecting the results). We need to further analyze these explanations, especially the *better query* class, which we leave for future work. Participants in both of our studies were employees at Microsoft and are may therefore be more familiar with technology than the average user. Studies with more general user cohorts are needed.

Our classifiers significantly outperformed the marginal and AP baseline for the categories *SAT*, *DSAT*, and *Other*, but not *Unintentional*. One explanation is that unintentional abandonment is either independent of SERP content or search behavior, making it difficult to learn with those features. Focusing on particular feature classes, *Result* and *All.NoInteraction* classes were found to perform reasonably well, and only slightly worse than *All*, which include cursor movements, clicks, and dwells. Since the features in *Result* and *All.NoInteraction* are available in the search engine

logs, extra instrumentation may not be required to achieve strong prediction accuracy. Interestingly, *Cursor* features also provide reasonable performance on their own, and significantly impact prediction accuracy when combined with all other features for this task. Cursor features were also prominent in the one versus all predictions (Table 1). More exploration of that is required.

Our findings of positive and negative associations of single features with abandonment rationales are of particular importance to search engine providers since these facilitate the definition of robust metrics for capturing aspects of perceived search quality.

Being able to accurately predict the reasons for abandonment has many implications for search engine design and evaluation. The frequency of SERP abandonment and its predicted rationale can be used as a metric to evaluate engine performance. Such a metric could be used to supplement existing click-based metrics by not only assigning an abandonment rate to the query, but also estimating the type of SERP abandonment that occurred. This also has implications for the layout of the SERP: certain queries could be altered to ensure good abandonment is encouraged, for example by inserting more inline answers, or bad abandonment designed out. This can also be used to reduce instances of bad abandonment. For example, Sarma et al. [28] propose algorithms for learning to rank with the goal of minimizing query abandonment, and such algorithms could be trained using the output of our classifiers to minimize bad abandonment rather than all abandonment. In addition, search engines that can predict abandonment given only a single SERP can adapt the search experience and/or the ranking algorithm for follow-on queries if DSAT is estimated

8. CONCLUSIONS AND FUTURE WORK

There is limited knowledge of why people abandon SERPs and only restricted offline analysis can be performed on abandonment in search logs. Better knowledge of the causes of search engine abandonment and methods to accurately predict these reasons are critical for understanding and modeling user satisfaction with search engines. We used data gathered by retrospective and in-situ surveys to characterize abandonment rationales, showing that one-in-five abandonment instances does not relate to satisfaction. We developed classifiers capable of accurately predicting why searchers abandon and studied the features that distinguish SAT and DSAT abandonment. Future work will involve further study of the reasons behind abandonment, the application of our models to search logs to develop abandonment-based metrics for assessing search engine performance, and training ranking algorithms using log data labeled with our classifiers to minimize instances of bad abandonment and ultimately improve searcher satisfaction.

REFERENCES

[1] E. Agichtein et al. (2006). Learning user interaction models for predicting web search result preferences. *SIGIR*, 3–10.

[2] E. Agichtein, E. Brill, and S.T. Dumais. (2006). Improving web search ranking by incorporating user behavior information. *SIGIR*, 19–26.

[3] G. Buscher, S. Dumais, and E. Cutrell (2010). The good, the bad and the random: An eye-tracking study of ad quality in Web search. *SIGIR*, 42–49.

[4] G. Buscher et al. (2012). Large-scale analysis of individual and task differences in search result page examination strategies. *WSDM*, 292–307.

[5] B. Carterette and R. Jones. (2007). Evaluating search engines by modeling the relationship between relevance and clicks. *NIPS*, 217–224.

[6] C. Castillo et al. (2010). When no clicks are good news. *SIGIR 2010 Industry Track*.

[7] L. Chilton and J. Teevan. (2011). Addressing people's information needs directly in a Web search result page. *WWW*, 27–36.

[8] E. Cutrell and Z. Guan (2007). What are you looking for? An eye-tracking study of information usage in Web search. *CHI*, 407–416.

[9] D. Downey et al. (2008). Understanding the relationship between searchers' queries and information goals. *CIKM,* 449–458.

[10] H. Feild, J. Allan, and R. Jones. (2010). Predicting searcher frustration. *SIGIR*, 34–41.

[11] S. Fox et al. (2005). Evaluating implicit measures to improve the search experience. *ACM TOIS*, 23(2): 147–168.

[12] J. Friedman, T. Hastie, and R. Tibshirani. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2): 337–407.

[13] Q. Guo, S. Yuan, and E. Agichtein. (2011) Detecting success in mobile search from interaction. *Proc. SIGIR*, 1229–1230.

[14] Q. Guo and E. Agichtein. (2010). Ready to buy or just browsing? Detecting web searcher goals from interaction data. *SIGIR*, 130–137.

[15] Q. Guo et al. (2011). Why searchers switch: understanding and predicting engine switching rationales. *SIGIR*, 335–441.

[16] A. Hassan, R. Jones, and K. Klinkner. (2010). Beyond DCG: user behavior as a predictor of a successful search. *WSDM*, 221–230.

[17] J. Huang, R.W. White, and S.T. Dumais. (2011). No clicks, no problem: using cursor movements to understand and improve search. *CHI*, 1225–1234.

[18] T. Joachims et al. (2005). Accurately interpreting clickthrough data and implicit feedback. *SIGIR*, 154–161.

[19] T. Joachims and F. Radlinski. (2007). Search engines that learn from implicit feedback. *Computers,* 40, 34–40.

[20] M. Kellar, C. Watters, and M. Shepherd. (2006). A goal-based classification of Web information tasks. *JASIST*, 43(1)

[21] E.A. Kensinger. (2007). Negative emotion enhances memory accuracy: Behavioural and neuroimaging evidence. *Curr. Dir. Psychol. Sci.* 16(4): 213–218.

[22] J. Li, S.B. Huffman, and A. Tokuda (2009). Good abandonment in mobile and PC internet search. *SIGIR*, 43–50.

[23] L. Lorigo et al. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *JASIST,* 59(7): 1041–1052.

[24] A.Y. Ng. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. *ICML*, 78–85.

[25] F. Radlinski, R. Kleinberg, and T. Joachims. (2008). Learning diverse rankings with multi-armed bandits. *ICML*, 784–791.

[26] K. Rodden et al. (2008). Eye-mouse coordination patterns on Web search results pages. *CHI*, 2997–3002.

[27] D. Rose and D. Levinson. (2004). Understanding user goals in Web search. *WWW*, 13–19.

[28] A. Sarma, S. Gollapudi, and S. Ieong. (2008). Bypass rates: Reducing query abandonment using negative inferences. *KDD*, 177–185.

[29] S. Stamou and E.N. Efthimiadis. (2009). Queries without clicks: Successful or failed searches? *SIGIR 2009 Workshop on the Future of IR Evaluation*, 13–14.

[30] S. Stamou and E.N. Efthimiadis. (2010). Interpreting user inactivity on search results. *ECIR*, 100–113.

[31] R.W. White and S.M. Drucker. (2007). Investigating behavioral variability in Web search. *WWW*, 21–30.

[32] R.W. White and S.T. Dumais. (2009). Charactering and Predicting search engine switching behavior. *CIKM*, 87–96.