Detecting Good Abandonment in Mobile Search

Kyle Williams[†], Julia Kiseleva[‡], Aidan C. Crook[Γ], Imed Zitouni[Γ],
Ahmed Hassan Awadallah[Γ], Madian Khabsa[Γ]
[†]The Pennsylvania State University, University Park, PA 16802, USA
[‡]Eindhoven University of Technology, Eindhoven, NL
[Γ]Microsoft, One Microsoft Way, Redmond, WA 98052, USA
kwilliams@psu.edu, j.kiseleva@tue.nl,
{aidan.crook, izitouni, hassanam, madian.khabsa}@microsoft.com

ABSTRACT

Web search queries for which there are no clicks are referred to as *abandoned queries* and are usually considered as leading to user dissatisfaction. However, there are many cases where a user may not click on any search result page (SERP) but still be satisfied. This scenario is referred to as *good abandonment* and presents a challenge for most approaches measuring search satisfaction, which are usually based on clicks and dwell time. The problem is exacerbated further on mobile devices where search providers try to increase the likelihood of users being satisfied directly by the SERP. This paper proposes a solution to this problem using gesture interactions, such as reading times and touch actions, as signals for differentiating between good and bad abandonment. These signals go beyond clicks and characterize user behavior in cases where clicks are not needed to achieve satisfaction. We study different good abandonment scenarios and investigate the different elements on a SERP that may lead to good abandonment. We also present an analysis of the correlation between user gesture features and satisfaction. Finally, we use this analysis to build models to automatically identify good abandonment in mobile search achieving an accuracy of 75%, which is significantly better than considering query and session signals alone. Our findings have implications for the study and application of user satisfaction in search systems.

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval — *Information filtering, Selection process*

Keywords: Mobile search behavior, good abandonment, touch interaction models, implicit relevance feedback

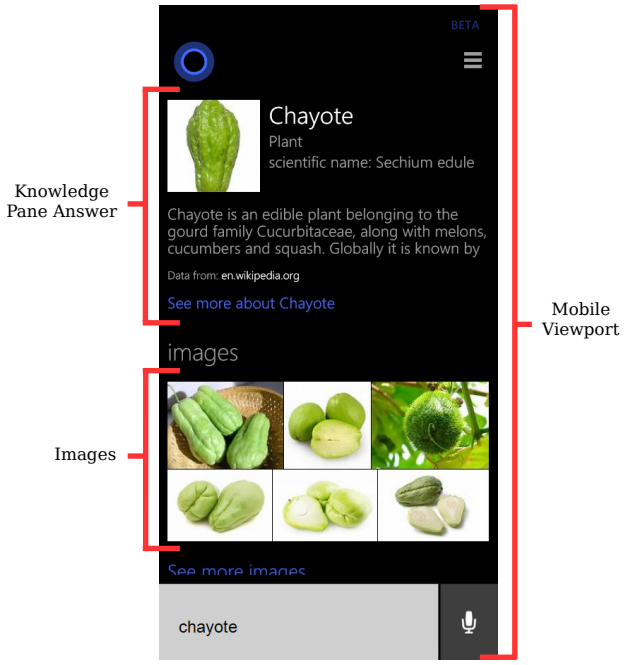

Figure 1: An example of a mobile SERP, showing the viewport, an answer and images.

1. INTRODUCTION

In recent years, there has been a large increase in people using their mobile phones to access the Internet, with it being reported that, in 2013, 63% of Americans used their mobile phones to go online compared to 31% in 2009 [7]. Having immediate access to mobile devices capable of searching the Web has led to important changes in the way that people use search engines. For instance, previous research has shown that search on mobile devices is often much more focused and that the query length and intents differ from traditional search [23]. It has also been found that mobile users might formulate queries in such a way so as to increase the likelihood of them being directly satisfied by the SERP [30]. In addition to these differences, the mobile screen sizes are typically much smaller than that of non-mobile devices. As a result of these differences, search engines have had to adapt in order to be able to better satisfy mobile users.

One way this has been done is by search engines presenting *answers* on the SERP in response to user queries. These answers typically come in the form of boxes containing a fact and, when present, they have the ability to satisfy the user need immediately. On mobile devices, there are many times

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883074.

when this may occur. For instance, a user may be out with friends and need to find the answers to questions that come up in conversation, such as *what will the weather be like tomorrow? What time does the movie start tonight?* Or *what year was a celebrity born?* Many of these types of questions can be answered by search engines without users needing to click on search results. Figure 1 shows an example of an answer that appears in the mobile search on Microsoft's digital assistant Cortana. The answer, which shows information about a plant, has the potential to directly satisfy the user's information need on the mobile SERP and thus may negate the need for the user to click on any hyperlinks. Furthermore, while it is clear that answers on a mobile SERP may satisfy a user, it is also possible for other elements on the SERP to do this. For instance, users can be satisfied by good snippets and images in SERPs.

Good abandonment refers to the case where a user is directly satisfied by the SERP without the need to click on any hyperlinks and the user is said to *abandon* the query [33]. This is in contrast to bad abandonment where a user abandons their query due to them being dissatisfied by the search results. It has been shown that good abandonment is more likely in mobile search. For instance, a study in 2009 estimated that 36% of abandoned mobile queries in the U.S. were likely good compared to 14.3% in desktop search [30].

Traditionally, abandoned queries have been considered a bad signal when measuring the effectiveness of search engines; however, recently there has been increasing awareness that abandonment can also be a good thing [1, 4, 30, 33]. However, most approaches for measuring search satisfaction and success have been based on implicit feedback signals such as clicks and dwell time [10, 18, 19, 26, 27]. However, these approaches to measuring satisfaction are not appropriate when good abandonment is taking place, especially in cases where mobile SERPs are being designed with the explicit goal of satisfying users without them needing to click. It thus becomes necessary to measure user satisfaction in the absence of clicks and recent studies have investigated various click-less approaches for doing this, such as those based on properties of the query [19] and the session [6, 33] and those based on gaze and viewport tracking [28].

We take a different approach and hypothesize that a user's gestures provide signals for detecting user satisfaction. Specifically, we focus on mobile search where gestures are prevalent and seek to answer the following main research question:

> *In the absence of clicks, what is the relationship between a user's gestures and satisfaction and can we use gestures to detect satisfaction and good abandonment?*

In this study, we use the term *gestures* to refer to users' click-less interactions with their mobile devices, such as touch gestures, swipe gestures and reading actions. In addressing this main research question, we focus on three sub-questions:

> **RQ1:** *Do a user's gestures provide signals that can be used to detect satisfaction and good abandonment in mobile search?*
>
> **RQ2:** *Which user gestures provide the strongest signals for satisfaction and good abandonment?*
>
> **RQ3:** *What SERP elements are the sources of good abandonment in mobile search?*

To our knowledge, this is the first work to consider the use of gestures to predict user satisfaction in mobile search and to use it to differentiate between good and bad abandonment. Furthermore, to our knowledge, this is also the first work to measure the relationship between user gestures and good abandonment in mobile search.

In summary, we make the following contributions:

- We construct gesture features for measuring user satisfaction in mobile search.
- We build a classifier that can automatically differentiate between good and bad abandonment and that performs significantly better than several baselines.
- We measure the correlation between gestures and satisfaction.
- We identify the SERP elements that lead to good abandonment in mobile search.

As this paper will show, gesture features are useful for detecting good abandonment, especially those that focus on user engagement with SERP elements. Furthermore, there are multiple causes for good abandonment on a mobile SERP, such as answers, snippets and images.

2. RELATED WORK

Related work falls into three categories: satisfaction in search; detecting good abandonment; and user gestures.

2.1 User Satisfaction in Search

Satisfaction is a subjective measure of a user's search experience and has been referred to as the extent to which a user's goal or desire is fulfilled [24]. For instance, satisfaction may be influenced by the relevance of results, time taken to find results, effort spent, or even by the query itself [25]. Thus, satisfaction is different from traditional relevance measures in information retrieval, such as Precision, MAP and NDCG, which are based on the relevance of results and not the overall user experience. However, similar to the case for relevance metrics, such as NDCG, satisfaction can also be fine-grained [22] and personalized [17] and it has been shown that search success does not always lead to satisfaction [13].

Several methods for measuring and predicting user satisfaction have been proposed. For instance, it has previously been shown that clicks followed by long dwell times are correlated with satisfaction [10]. Hassan et al. [19] propose to use query reformulation as an indicator of search success and thus satisfaction and show how an approach based on query features outperforms an approach based on click features, with the best performance being achieved by a combination of the two. Like our proposed work, this work does not consider clicks; however, it differs from ours since we consider gestures rather than query reformulation. Furthermore, we focus on good abandonment rather than general satisfaction.

In [18], the search process is modeled as a sequence of actions including clicks and queries and two Markov models are built to characterize successful and unsuccessful search sequences. In [16], a sequence of actions is also considered, but a semi-supervised approach is shown to be useful for improving performance when classifying Web search success.

Kim et al. [26] consider three measures of dwell time and evaluate their use in detecting search satisfaction. In [27] it

is shown that the SAT and DSAT dwell times for a page depend on the complexity and topic of a page. To address this issue, the authors propose query-click complexities in modeling dwell times on landing pages. Since we only consider abandoned queries in our study, landing page dwell times do not exist; however, we do consider a similar feature based on visibility and reading times for various elements in a SERP.

2.2 Good Abandonment

Diriye et al. [6] investigate the rationale for abandonment in search. In a survey involving 186 participants, it was found that satisfaction was responsible for 32% of abandonment. They also studied 39,606 queries submitted to a search engine of which about 22% were abandoned and, for half of the abandoned queries, rationale for abandonment were collected via a popup window. For the cases where feedback was provided, it was found that satisfaction was responsible for 38% of abandonment.

In [34] it was found that 27% of searches were performed with the pre-determined goal of having the search satisfied by the SERP and that 75% of searchers were satisfied this way. In [30] it was found that, for queries that could potentially lead to good abandonment, 56% were clearly or possibly satisfied by the SERP on the desktop, and 70% on mobile. The authors hypothesized that one of the reasons for the higher potential abandonment rates on mobile is because users may formulate queries in such a way so as to increase the likelihood of them being answered on the SERP due to a clumsy experience in retrieving webpages for display on mobile. In [3], the effect that answers have on users' interactions with a SERP is studied and it is observed that the presence of answers *cannabilize* clicks by reducing interaction with the SERP. A similar finding was presented in [5] where it was found that high quality SERPs decrease click-through rates and increase abandonment. For this reason, we consider features that incorporate non-click interactions with answers, such as element visibility duration and attributed reading time (see Section 5.1).

In [33], context is considered in predicting good abandonment. Query-level features, such as query length and reformulation, SERP features that consider clicks in neighboring queries and the presence of answers on a SERP, and session features are used to identify good abandonment. In [4], topical, linguistic features are used to detect potential good abandonment and achieved F-scores of 0.38, 0.55 and 0.71 for maybe, good and bad abandonment, respectively. Our work differs from these approaches in that we use non-click gesture features for detecting good abandonment.

2.3 Gestures for Relevance & Satisfaction

User gestures have been used in various ways to detect success and satisfaction in search. One of the common approaches is to use scroll and mouse movement behaviors in satisfaction prediction [2, 11, 12, 32]. In [12] post-click behavior, such as scrolls and cursor movement, is used to estimate document relevance for landing pages. In [14] similar features are used to predict session success. Our work differs from this work in that we do not attempt to detect post-click satisfaction, but instead predict satisfaction in the absence of a click. Furthermore, scrolls and cursor movements do not exist in mobile search; however, the swipe interaction performs a similar function and we use swipe interactions as signals for detecting good abandonment.

The two studies most similar to ours evaluate the use of user interaction on mobile phones for detecting search result relevance [15] and use eye- and viewport-tracking to measure user attention and satisfaction [28]. User interactions on mobile phones, such as swipes, dwell times on landing pages and zooms are used in [15] to predict Web search result relevance. While our study uses similar gesture features to [15], our study differs from this since, instead of predicting relevance of landing pages, we differentiate between good and bad abandonment. Furthermore, landing page interactions are used in [15], whereas we use gestures on the SERP itself and do not take visited pages into consideration. Similar features were combined with server-side features such as click-through rate in [13] to predict search success. Once again, our approach differs from this work in that we attempt to predict good abandonment. In [28] viewport- and eye-tracking were used to measure user attention and satisfaction. The authors establish the correlation between gaze time and viewport time and also studied the effect of having relevant/irrelevant answers on the user behavior and the correlation between individual signals and relevance. The authors focus on SERPs containing answer-like results since clicks on these answers do not occur frequently. Through a user study, it was shown that users are more satisfied when answers or knowledge graph information is present in the SERP. Our work differs in that, instead of only focusing on answers, we consider multiple sources of satisfaction and good abandonment in mobile search; we also consider a large number of gesture-based features beyond gaze and viewport times. Lastly, the authors in [28] suggest building a model to predict satisfaction and good abandonment as a future application; such an application is presented here, through a model for automatically identifying satisfaction and good abandonment using gesture-based features in mobile search.

3. PROBLEM DESCRIPTION

In this paper we seek to understand and differentiate between good and bad abandonment in mobile search. We seek to identify the sources of good abandonment, to understand the relationship between user behavior and good abandonment and to identify click-less features that can be used for differentiating between good and bad abandonment.

Our main hypotheses in conducting this study are that: 1) gestures provide useful features for detecting good abandonment; and 2) there are many reasons for good abandonment.

To address these problems and investigate our hypotheses we require a set of queries and satisfaction labels, which we collect through a user study and crowdsourcing. We also require a set of gestures that can be used as signals for measuring satisfaction, which we develop as part of this study. In the following sections, we present the datasets we created as well as the signals we identified.

4. DATA SETS

To collect data to understand good abandonment in mobile search, we conducted a focused user study whereby users completed a set of search tasks and provided satisfaction ratings. This led to a dataset of high quality user supplied data that we use for our analysis. However, this dataset is relatively small; thus, we also collected a second dataset via crowdsourcing that we use to validate our findings. This section describes our data collection.

Table 1: SAT Rating Distribution.

SAT Rating	Number of Tasks
1	14
2	19
3	47
4	82
5	112

4.1 User Study

We recruited 60 participants from the United States where 75% of the them were male and the remaining 25% female. The majority (82%) of participants were from a computer science background and the remaining 18% specified their background as either mathematics, electrical engineering or other. English was the first language for 55% of the participants and the mean age was 25.5 ($\pm$5.4) years.

In the user study, 5 information-seeking tasks, which represent atomic information needs [31], were designed in such a way that they may lead to good abandonment. The tasks were not designed to encourage exploration, but rather to allow the user to answer a question. They were:

1. A conversion between the imperial and metric systems.
2. Determining if it was a good time to phone a friend in another part of the world.
3. Finding the score from a recent game of the user's favorite sports team.
4. Finding the user's favorite celebrity's hair color.
5. Finding the CEO of a company that lost most of its value in the last 10 years.

After each task users provided a satisfaction rating on a 5-point scale, specified if they were able to complete the task and the amount of effort required, and provided feedback on the SERP element that provided the information they were looking for and the query that led to them being satisfied.

4.1.1 Data Description

In the user study, the total number of potential abandonment tasks was 274. A total of 607 queries were submitted for these tasks, with the minimum, maximum, mean and median number of queries per task being 1, 9, 2.2 and 2, respectively. Of the 607 queries, 576 were classified as abandoned queries since they received no clicks.

The SAT distribution (on a scale of 1-5) is shown in Table 1. As can be seen from the table, SAT ratings of 4 and 5 make up the majority of the task satisfaction labels. In this study, we follow the approach in previous studies [13, 21] and binarize these values and consider ratings of 4 and 5 as SAT and the remainder of the ratings as DSAT. With this binarization, there are 194 SAT tasks and 80 DSAT tasks.

4.1.2 Label Attribution

Labels in the user study were collected at the task level. However, good abandonment takes place at a query level. Thus, a way is needed to attribute labels to individual queries. Since users were asked to stop when they found the information they were looking for, the method for doing this is based on the observation that, if a user continues querying then they are likely not satisfied; however, when a user stops querying then they are either a) giving up the task or, b) satisfied. Based on this observation, individual impressions were labeled as follows: If the task was assigned a DSAT label, then every query for that task was assigned DSAT. If the task was assigned a SAT label, then the final query for the task was assigned the SAT label and every query before it was assigned DSAT. The assumption here is that the queries lead to DSAT until the user meets their information need at which point the query leads to SAT. After filtering queries for which not all features were available, we retained a total of 563 queries of which 461 were abandoned queries.

4.2 Crowdsourcing

The data collected in the user study is of high quality since users could directly provide information on their satisfaction; however, with only 563 queries, this dataset is relatively small. We thus collected a second set of labeled data via crowdsourcing, which is a common approach to collecting labeled data [35] and that we use to validate our findings. This section describes the collection of that data.

4.2.1 Approach

Since our focus is on good abandonment, we randomly sampled abandoned queries from the search logs of a personal digital assistant during one week in June 2015. We filtered the data such that: no adult queries were sampled; all queries originated from within the United States; all queries were input via speech or text (as opposed to, say, suggested queries); and all queries generated a SERP containing organic Web results and possibly answers.

We made use of a commercial crowdsourcing platform. Judges were shown a video explaining the task and how to judge queries with good or bad abandonment, for instance, by considering the query and the SERP and by taking the query context into consideration. Judges needed to pass qualification tasks in order to participate in labeling real data and the crowdsouring engine had built in spam detection. For each query randomly sampled from the logs, judges were shown: the query, a screenshot of the mobile SERP returned for that query, the previous query in the session and the next query in the session. Judges were asked to provide two judgments: 1) their perception of user-satisfaction on a 5-point scale and 2) if they believed the user was satisfied, which we defined as the user finding the information they were looking for, which type of element on the SERP satisfied the user. Though we asked judges to provide feedback on a 5-point scale, we binarized the labels in the same way as the user study data. We had up to 3 judges provide labels for each query and took the majority vote.

4.2.2 Data Description

We gathered a total of 3,895 labeled queries. Among the first two judgments collected for each query, the judges agreed on the label 73% of the time. We measured inter-rater agreement using Fleiss' Kappa [9], which allows for any number of raters and for different raters rating different items. This makes it an appropriate measure of inter-rater agreement in our study since different judges provided labels for different items. A kappa value of 0 implies that any rater agreement is due to chance, whereas a kappa value of 1 implies perfect agreement. In our data, $\kappa = 0.46$, which, according to Landis and Locke [29], represents moderate agree-

ment. This relatively low κ is indicative of a difficult task. After filtering queries for which not all features were available, we retained 1,565 queries for which the judgment was SAT and 1,924 queries for which the judgment was DSAT.

5. GESTURES AS SATISFACTION SIGNALS

Click signals are not available for measuring satisfaction in abandoned queries. This section describes a set of click-less features that we developed to measure good abandonment.

5.1 Gesture Features

One of the main contributions of this study is in the use of gesture features for detecting good abandonment and satisfaction on mobile devices. Specifically, we focus on gesture features related to the way in which the user interacts with the screen and features based on the elements visible to the user. As noted in [20], capturing touch events is difficult in practice; however, it is possible to infer touch-based interactions based on the mobile viewport, which is the visible region on the device. For instance, if an element is visible in the viewport at some point in time and then no longer visible, one can infer that a gesture must have taken place.

Table 2 lists the features used in this study. As previously specified, we use the term *gestures* to refer to touch- and reading-based actions. We also group element visibility features with gesture features since the visibility of an element may imply reading. We separate our features into 6 categories: viewport features (VP); first visible answer features (FA); aggregate answer features (A); aggregate organic search result features (O); focus features (F); and query-session features (QS). We describe these features now.

5.1.1 Viewport Features

Viewport features, which are represented by features VP1-VP9 in Table 2, capture the user's overall touch gestures with their mobile device. Swipes refer to the gesture whereby the user *swipes* on their device screen to move the content that is visible on the screen. We count the total number of swipes (VP1), the number of up swipes (VP2) and the number of down swipes (VP3). We also count the number of times the user changed swipe direction (VP4), i.e., a down swipe followed by an up swipe or vice versa. We also measure the total distance in pixels swiped on the screen (VP5) and the average distance per swipe (VP6). These features capture the number of SERP features seen by the user. We capture the total time spent on the SERP (VP7) and also the average amount of time between swipes (VP8), which captures how long the user spent looking at the screen after it changed. Lastly, we capture the swipe speed (VP9) as it is has been shown that a slow swipes are associated with reading and fast swipes are associated with skimming [15].

5.1.2 First Answer Features

One of our hypotheses in conducting this study was that the highest ranked visible answer on a SERP, by nature of being highly ranked, has the highest likelihood of satisfying the user. Thus, we capture a set of features that relate to the first visible answer on a SERP. We estimate the attributed reading time for the first visible answer on the SERP (FA1) based on the hypothesis that a higher attributed reading time suggests more engagement with the answer and thus may potentially result in higher satisfaction. We calculate attributed reading time for answer e, ART_e as:

$$ART_e = \sum_{v \in V} t_v \times \frac{AA_{e,v}}{VA_v}, \quad (1)$$

where V is the set of viewport instances, t_v is the duration of time for which viewport v was visible and $AA_{e,v}$ and VA_v are the visible areas of answer e and viewport, respectively, in the viewport v. We also attribute a reading time to each pixel belonging to the first answer (FA2). We calculate the attributed reading time per pixel for an answer e, RTP_e as:

$$RTP_e = \frac{1}{AA_{e,O}} ART_e, \quad (2)$$

where $AA_{e,O}$ is the pixel area of the answer e that was ever observable by the user across all viewports corresponding to the impression.

We calculate the total duration for which the first answer is (even partially) shown (FA3), which differs from attributed reading time since it is not scaled according to the visible area of the answer. Lastly, we calculate the fraction of visible pixels belonging to the first answer (FA4) as $\frac{AA_{e,O}}{AA_e}$, where AA_e is the physical pixel area of the underlying answer, observed or not.

5.1.3 Aggregate Answer Features

Features FA1-FA4 related specifically to the first visible answer on a mobile SERP. Features A1-A16 are similar in this regard, except that they aggregate and provide descriptive statistics based on the set of answers visible on a SERP. Specifically, we calculate the min, max, mean and standard deviation of the following features for the set of answers: attributed reading time (A1-A4); attributed reading time per pixel (A5-A8); total duration shown (A9-A12); and fraction of visible pixels (A13-A16).

5.1.4 Aggregate Organic Result Features

We also aggregate the same set of features for organic search results by calculating the min, max, mean and standard deviation of the following for visible organic search results: attributed reading time (O1-O4); attributed reading time per pixel (O5-O8); total duration shown (O9-O12); and fraction of visible pixels (O13-O16).

5.1.5 Time to Focus Features

We define two *time to focus* features. These features capture how long it takes a user to focus on a page element where we define focus as occurring when an element is visible for 5 seconds. The intuition behind this feature is that if a user takes a long time to focus on an element, then it may suggest decreased satisfaction due to scrolling. When an element has been visible for 5 seconds, we set the time to focus as the timestamp at which the element first became visible. We calculate the time to focus on an answer (F1) and an organic search result (F2).

5.2 Query & Session Features

While the main contribution of this work is in the gesture features, it has previously been shown that other user behavior also provides strong signals for satisfaction [18, 19]. Thus, we also use a set of features based on the query and the user behavior within the session. these features are shown by features QS1-QS10 in Table 2 and are self-explanatory.

Table 2: Description of features used in this study. The last two columns show Pearson's correlation with satisfaction (SAT) for both the data gathered in the user study and the data gathered via crowdsourcing. Missing values (-) indicate that the correlation was not statistically significant ($p > 0.05$).

	Feature Description	User SAT Correlation	Crowd SAT Correlation
VP1	Total number of swipe actions	-0.08	-0.14
VP2	Number of up swipe actions	-	-0.04
VP3	Number of down swipe actions	-0.08	-0.15
VP4	Number of swipe direction changes	-	-0.09
VP5	The total distance swiped in pixels	-0.10	-0.14
VP6	The average swipe distance	-0.10	-
VP7	The dwell time on the SERP	-	-
VP8	The mean dwell time on SERP before or after each swipe	-	-
VP9	Total swipe distance divided by time spent on the SERP	-0.11	-0.11
FA1	Attributed reading time (RT) for the first visible answer]	-	0.04
FA2	Attributed reading time per pixel (RTP) of the first answer	0.10	0.08
FA3	The duration for which the first answer was shown	-	0.06
FA4	The fraction of visible pixels belonging to the first answer	-	0.15
A1-A4	Max, min, mean and SD attributed RT for answers	-/-/-/-	0.04/-/-/0.04
A5-A8	Max, min, mean and SD attributed RTP for answers	0.11/0.11/0.11/-	0.08/0.06/0.07/0.04
A9-A12	Max, min, mean and SD shown duration for answers	-/-/-/-	0.04/0.05/0.05/-
A13-A16	Max, min, mean and SD shown fraction for answers	-/-/-/-	0.15/0.11/0.14/0.10
O1-O4	Max, min, mean and SD RT for organic results	-/-/-/-	-0.15/-/-0.09/-0.12
O5-O8	Max, min, mean and SD RTP for organic results	-/0.10/-/-	-0.13/-/-0.06/-0.12
O9-O12	Max, min, mean and SD shown duration for organic results	-/-/-/-	-/-/-/-
O13-O16	Max, min, mean and SD shown fraction for organic results	-0.20/-0.19/-0.29/0.10	-0.20/-0.07/-0.22/-0.05
F1	Time to focus on an answer	-	-0.05
F2	Time to focus on an organic search result	-	-
QS1	Session duration	-	-
QS2	Number of queries in session	-0.16	-
QS3	Index of query within session	-0.24	-
QS4	Query length (number of words)	-0.17	-0.26
QS5	Is this query a reformulation?	-0.11	-0.10
QS6	Was this query reformulated?	-0.35	-0.15
QS7	Time to next query	0.16	-0.04
QS8	Click count	-	-
QS9	Number of clicks with dwell time > 30 seconds	-	-
QS10	Number of clicks followed by a back-click within 30 seconds	-	-

5.3 Endogenous & Exogenous Features

The features used in this study were designed to be *exogenous*, meaning that the system does not have direct control over them but that instead the features are based on user input, such as swipe actions and dwell times. This is in contrast to *endogenous* features that the system can directly influence. For instance, the presence of a certain answer type, e.g., weather, is an example of a likely endogenous feature. While endogenous features are useful for measuring satisfaction, they present a challenge for search engine evaluation since a system can be unintentionally optimized for these features. As an example, if the presence of a weather answer is an indicator of satisfaction, then an answer ranker may learn to always rank weather answers highly thereby *gaming* the metric. Though we found endogenous features to be very useful for detecting good abandonment, for the reasons described above we choose to only use features that are mostly exogenous in this study. It is important to note, though, that the classification of endogenous and exogenous features is not absolute, but rather falls along a spectrum depending on the search engine and metric, and that the classification will differ depending on the circumstances.

6. GOOD ABANDONMENT, INTERACTION & SATISFACTION ON MOBILE DEVICES

In this section, we present the reasons for good abandonment and show which user gestures are correlated with good abandonment. We also investigate the relationship between satisfaction and other feedback collected from users.

6.1 Causes of Good Abandonment

The main contribution of this research is an investigation into the use of gesture features to detect good abandonment. One of the first stages in doing this is understanding the causes of good abandonment. This allows us to consider the contents of a SERP when trying to determine if a query was abandoned because the user is satisfied without the need to click. Thus, in the user study, we asked users to provide feedback on the source of satisfaction. The users were asked to select from among the following:

- **Answer.** An answer on the SERP.
- **Search Result Snippet.** The text appearing below a search result.
- **Image.** An image displayed on the SERP.

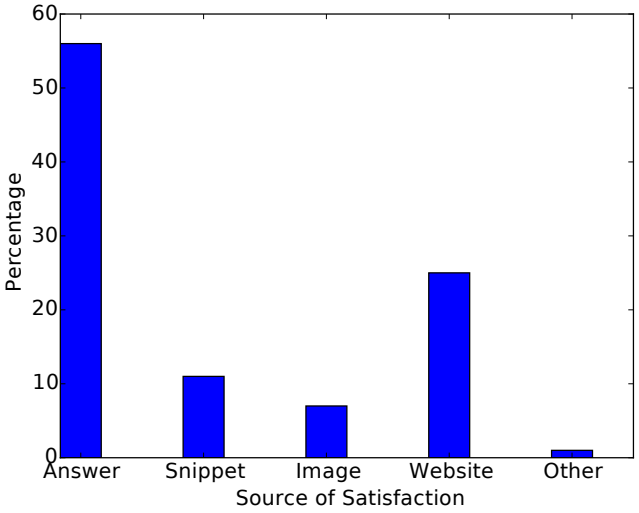

Figure 2: A comparison of the counts of the sources of satisfaction from the user study.

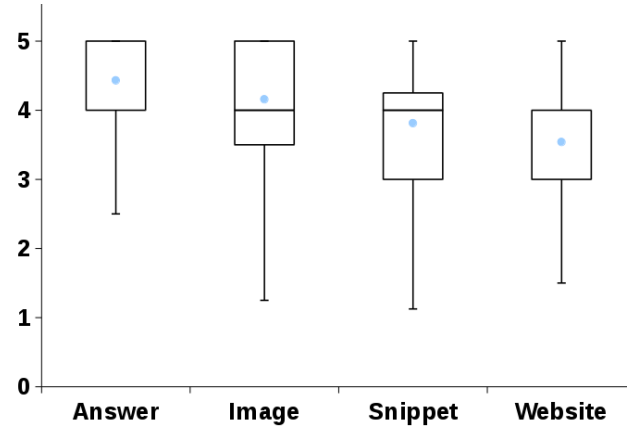

Figure 3: Satisfaction associated with each source of information.

- **Website.** If the user visited a Website to satisfy their information need.
- **Other.** An element on the SERP that does not belong to one of the above categories.

As can be seen from Figure 2, the majority of user satisfaction (56%) was due to answers on the SERP. However, an important observation is that good abandonment can be due to other sources on the SERP. For instance, images made up for 7% of satisfaction and snippets made up 11% of satisfaction. Since users were allowed to click on search results, websites were responsible for 25% of satisfaction, which is less than half of the number of times users were satisfied by answers. This analysis provides an answer to **RQ3:** *What SERP elements are the sources of good abandonment in mobile search?* It confirms our hypothesis that there are many sources of satisfaction on a SERP.

Figure 3 shows the user satisfaction associated with each of the sources of satisfaction. The mean is represented by the dot and the median by the horizontal line. As can be seen from the figure, the satisfaction ratings are highest for the answers on the SERP, and the means for images and snippets are relatively close to that for answers. The mean for websites is the lowest since users have to visit websites without knowing if it will satisfy them.

6.2 Gesture Features & Satisfaction

To better understand the relationship between gestures and good abandonment and satisfaction, we calculate the Pearson correlation between the satisfaction label and each feature. The statistically significant correlations ($p < 0.05$) for the user study data and crowdsourced data are shown in the two last columns of Table 2 where a missing value (-) indicates that the correlation was not significant ($p > 0.05$).

As can be seen from Table 2 there are several features that are significantly correlated with SAT. For instance, features from the crowdsourced data related to swipes such as the total number of swipes, the number of down swipes and the distance swiped are all negatively correlated with satisfaction. We note that one limitation with this observation is that judges were only presented with screenshots of the mobile SERP and thus were unable to swipe to see if there was additional information on the SERP not shown in the screenshot that may have satisfied the user. That being said, a similar trend is observed for the user study data where users were able to swipe. For instance, for the user study both the total number of swipes and the number of down swipes are negatively correlated with satisfaction. Furthermore, a similar finding was presented in [28] where it was shown that scrolling is negatively correlated with user satisfaction. The fact that the swipe action is negatively correlated with satisfaction suggests that the more time that users spend physically touching and moving the viewport on a mobile device, the less likely they are to be satisfied. One reason that this may be the case is that, as shown in Figure 2, a lot of good abandonment is due to answers and, when an answer is present on the viewport there may be less reason for the user to physically interact with the SERP.

Features related to the reading and visibility of answers (features FA1-F4; A1-A16), when statistically significant, are all positively correlated with satisfaction. This implies that the longer users spend viewing answers, the more likely they are to be satisfied. This is interesting when contrasted with feature VP7, which is the total time spent on the SERP, and which is not statistically significant. The data suggests that the time spent on a SERP is not a strong signal for satisfaction but that the time spent viewing answers is.

The opposite effect is observed when considering the correlation between satisfaction and the time spent reading and viewing organic search results (features O1-O16). When significant, increased interaction with organic search results is negatively correlated with satisfaction. Increased interaction with organic search results may imply that users are spending more time on the SERP unsuccessfully looking for information to satisfy their information needs.

The analysis above provides an answer to **RQ2:** *Which user gestures provide the strongest signals for satisfaction and good abandonment?* Features related to swipe actions and interaction with organic search results provide indications of bad abandonment. On the other hand, extended

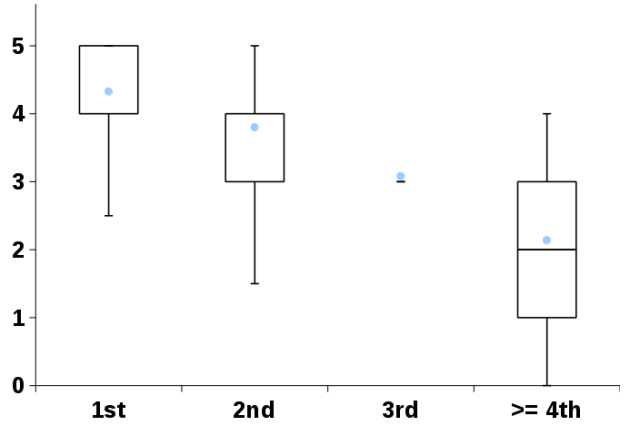

Figure 4: The relationship between query number and satisfaction.

reading-based interactions with answers on a SERP are signals that suggest good abandonment.

Table 2 also shows that the correlation between satisfaction and features based on the query and session (QS1-QS10). Our finding confirms existing findings in the literature, such as the fact that query length and reformulation are negatively correlated with satisfaction [19, 33] and, as in [33], we find in our user study data that the time to next query is positively correlated with satisfaction though we observe the opposite effect in our crowdsourced data.

6.3 User Feedback & Good Abandonment

In addition to asking users how satisfied they were and where they found the information they were looking for, we also asked them: (a) if they were able to complete the task, (b) how much effort they put into the task and (c) which query led to them finding the answers, with them being able to specify first, second, third or fourth or later. We find strong significant negative correlation of -0.65 between satisfaction and effort, and a negative correlation of -0.08 between completion and effort, indicating that less effort leads to more satisfaction and higher completion rates.

Figure 4 shows the relationship between satisfaction and the number of queries submitted by the user. As can be seen from the figure, there is a negative relationship between the number of queries required to satisfy the user's information need and their level of satisfaction. This finding makes sense for information seeking tasks, such as those used in this user study; however, we suspect that for exploratory tasks this finding may not always hold; we leave this to future work.

7. CLASSIFYING ABANDONED QUERIES

The previous section presented an analysis of the reasons for good abandonment and which behaviors are correlated with satisfaction. In this section, we present our approach to differentiating between good and bad abandonment.

7.1 Approach

We formulate a supervised classification problem where, given an abandoned query, the goal is to classify the query as being due to good abandonment or not. We use a random forest classifier[1], which is an ensemble classifier made up of a set of decision trees. Each tree is built with a bootstrap sample from the dataset and splitting in the decision tree is based on a random subset of the features rather than the full feature set [8]. In this study, the number of trees in the ensemble is set to 300 since this was empirically found to perform well and the number of features randomly selected is equal to $\sqrt{n\_features}$. At each level in the decision trees, variables are selected for splitting with the Gini index.

We use 10-fold cross validation and use grid search within each training fold to optimize for the number of leaves, tree depth and number of leaves required to split. During training, we downsample the majority class so that our class representation is even; however, we leave the class distribution unchanged in the testing data. Since we do random downsampling of training data, we repeat each experiment 100 times and report the average. For our experiments, we make use of 3 baselines and propose 2 new models.

7.2 Baselines

7.2.1 *Click and Dwell with no Reformulation*

This baseline is based on the common approach in the literature as labeling satisfaction as occurring if a user clicks on a search result and then spends a minimum of t seconds on a page and does not follow the query up with a reformulation. Spending a minimum amount of time on a webpage is known as a long dwell click and has been shown to be correlated with satisfaction [10]. In this study, we set $t = 30$ seconds. Naturally, this baseline does not make much sense for the detection of good abandonment since, by definition, abandoned queries do not have any clicks. Nonetheless, it is useful to use this baseline for comparison so as to show why click-based metrics are not appropriate.

7.2.2 *Optimistic Abandonment*

Baseline 2 is an optimistic one whereby, if there is no click and no reformulation, then it is assumed that the abandonment is good. We refer to this baseline as optimistic since it optimistically assumes that all abandonment without reformulation is good. For queries that receive clicks, the same approach as in Baseline 1 is used to measure satisfaction.

7.2.3 *Query-Session Model*

Baseline 3 makes use of features from the literature for detecting satisfaction and good abandonment. Specifically, it is a supervised classifier based on features QS1-QS10 in Table 2 that represent the query and the session.

7.3 Proposed Models

7.3.1 *Gesture Model*

This is a supervised classifier based only on the interaction features in Table 2, which is all except features QS1-QS10. The purpose of this model is to only consider the users physical behavior and gestures with the screen and investigate their usefulness in detecting good abandonment.

7.3.2 *Gesture + Query-Session Model*

This is a supervised classifier that combines the interaction-features model and the query-session model.

[1]We use the scikit-learn implementation of random forests. http://scikit-learn.org/stable/index.html

7.4 Results

We present three sets of results. First, we present results using only abandoned queries from the user study. Secondly, since the user study dataset is relatively small, to validate our approach we repeat the experiment using the crowdsourced data. Lastly, even though the focus of this study is on good abandonment, it is also useful to investigate the use of click-less interaction features for detecting satisfaction in general. Thus, we also present satisfaction detection results on all data from the user study, which includes both abandoned and non-abandoned queries. For each experiment we report the overall accuracy as well as the precision (P), recall (R) and F_1 score for SAT and DSAT separately. Bold values in the columns of Tables 3-5 indicate the best performance for that metric. When measuring result significance, we make use of the Wilcoxon signed-rank test.

7.4.1 Abandoned User Study Queries

Table 3 shows the performance on abandoned queries from the user study. As can be seen from the table, the highest accuracy of 75% is achieved by the model that combines gesture features with query-session features and is significantly better ($p < 0.01$) than the accuracy achieved by all other models. The approach based on query and session features from the literature achieves an accuracy of 73% and the gesture features alone achieve an accuracy of 70%. While the accuracy achieved by the gesture features is not as high as that achieved by the query-session features, it is still very interesting to note that, using only gesture features, it is possible to differentiate between good and bad abandonment with 70% accuracy and that this approach is significantly better ($p < 0.01$) than the other two baselines.

Table 3 also shows precision, recall and F1 scores for SAT and DSAT. As would be expected, the first baseline based on click and dwell performs very badly on SAT since there are no clicks. Thus, while it results in the highest F1 score for DSAT, the F1 score for SAT is 0. The optimistic baseline overestimates SAT and thus has low SAT precision but high SAT recall. However, this comes at the expense of having the lowest DSAT recall and lowest accuracy overall.

The model that combines query-session and gesture features achieves the second highest F_1 score for DSAT and the highest F_1 score for SAT. In fact, the model performs either best or second best for every metric and the best overall if one considers the accuracy or the F_1 scores.

7.4.2 Crowdsourced Data

To validate our model, we also consider differentiating between good and bad abandonment in the data gathered via crowdsourcing. Table 4 shows the performance. As can be seen from the table, as was the case with the user study data, the best accuracy of 68% is achieved by combining gesture and query-session features and is significantly better than all other methods ($p < 0.01$). Interestingly, for this data, the gesture features perform as well as the query-session features, with both methods achieving accuracies of 64% and both outperforming the other baselines. Overall, the query-session model and the gesture models achieve similar performance across all metrics.

As was the case with the user study data, the click & dwell baseline is unable to detect SAT since all of the queries are abandoned and have no clicks. Similarly, the optimistic baseline performs relatively poorly when it comes to its precision in detecting SAT since it overestimates good abandonment in the data; however, for this reason it achieves the highest SAT recall but the lowest DSAT recall.

The combination of query-session and gesture features achieves the highest precision for both SAT and DSAT as well as the best recall and F_1 score if one averages the values for SAT and DSAT.

7.4.3 All User Study Queries

To show the appropriateness of interaction features for detecting other types of satisfaction in addition to good abandonment, we also run a classification experiment on all data from the user study, which includes some queries that had clicks. Table 5 shows the performance on this data. As can be seen from the table, the highest accuracy when not including gesture-interaction features is 69% and is achieved by making use of the third baseline, which uses query-session features. The other baselines achieve accuracies of 66% and 61%, respectively. When only interaction features are considered, the accuracy is 66%, which is equal to the accuracy achieved by the click and dwell baseline, but less than the query-session features. However, when gesture features are combined with query-session features, the accuracy increase to 72%, which is statistically significantly better ($p < 0.01$) than all the other approaches. This combined model also achieves the highest SAT precision and F1 score, and performs second best for all other metrics.

While this paper has focused on detecting good abandonment, this experiment has shown that the gesture features are useful for detecting satisfaction in general. We expect this to be an interesting area for future research.

7.5 Providing an Upper Bound

As discussed in Section 5.3, in this study we focused on exogenous features, which are more difficult for the ranker to optimize for. This is in contrast to endogenous features, such as the presence of a certain answer type. However, to estimate an upper bound on an accuracy that may be feasible to achieve with the collected data, we also conducted an experiment where we additionally considered a set of endogenous features. Specifically, we include the following endogenous features: the number of answers and organic results on the SERP; the number of answers and organic results that came into view; the fraction of the number of answers and organic results that were visible; binary features indicating the presence of different answer types on the SERP, such as weather, currency, etc. Using these endogenous features, we achieve an accuracy of 78% on the user study data and an accuracy of 70% on the crowdsourced data. Both of these models demonstrate improvements over models where only exogenous features are used; however, as previously discussed, it is often undesirable to use exogenous features.

7.6 Discussion and Implications

We have presented various experiments for differentiating between good and bad abandonment. Our main finding is that gesture features are useful for accomplishing this goal, often achieving the same or very similar performance to an approach based on query and session features. Overall though, the best performance comes from combining these gesture features with query-session features. The reason for this is that gesture features provide us with signals that we may not be able to get from the query or session. For in-

Table 3: Performance of classifiers on only abandoned user study data. 148 SAT queries; 313 DSAT queries.

Classifier	Accuracy	SAT P	DSAT P	SAT R	DSAT R	SAT F1	DSAT F1
Click & Dwell	0.68	0.00	0.68	0.00	**1.00**	0.00	**0.88**
Optimistic	0.61	0.45	**0.93**	**0.93**	0.46	0.61	0.62
Query-Session (QS)	0.73	0.56	0.87	0.77	0.71	0.65	0.78
Gesture	0.70	0.53	0.84	0.70	0.70	0.60	0.76
Gesture + QS	**0.75**	**0.59**	0.88	0.78	0.74	**0.67**	0.80

Table 4: Performance of various classifiers on crowdsourced data. 1565 SAT queries; 1924 DSAT queries.

Classifier	Accuracy	SAT P	DSAT P	SAT R	DSAT R	SAT F1	DSAT F1
Click & Dwell	0.55	0.00	0.55	0.00	**1.00**	0.00	**0.71**
Optimistic	0.53	0.49	0.71	**0.88**	0.25	0.63	0.37
Query-Session (QS)	0.64	0.59	0.69	0.66	0.63	0.62	0.66
Gesture	0.64	0.59	0.69	0.65	0.62	0.62	0.65
Gesture + QS	**0.68**	**0.63**	**0.73**	0.69	0.67	**0.66**	0.70

Table 5: Performance of various classifiers on all user study data. 179 SAT queries; 384 DSAT queries.

Classifier	Accuracy	SAT P	DSAT P	SAT R	DSAT R	SAT F1	DSAT F1
Click & Dwell	0.66	0.27	0.68	0.67	**0.94**	0.10	**0.79**
Optimistic	0.61	0.44	**0.87**	**0.84**	0.50	0.58	0.63
Query-Session (QS)	0.69	0.52	0.84	0.72	0.68	0.60	0.75
Gesture	0.66	0.48	0.80	0.64	0.67	0.55	0.73
Gesture + QS	**0.72**	**0.55**	0.85	0.73	0.71	**0.62**	**0**.77

stance, reformulation is usually considered a strong signal for DSAT; however, the absence of reformulation does not necessarily imply SAT as was the assumption in our second baseline, which was an optimistic classifier. Instead, our findings suggest that combining signals, such as the fact that the user did not reformulate, with information on how the user interacted with the screen is more powerful.

While this study has focused on detecting good abandonment, our experiment considering all of the user study data showed that interaction features were also useful for detecting satisfaction when clicks existed and outperformed the baseline based on a click followed by a long dwell. We believe that it will be useful to consider gesture features for general satisfaction prediction and leave this for future work.

The implications of our experiments is two-fold. Firstly, it is important to develop click-less models that are able to capture satisfaction due to good abandonment. Secondly, we have shown that, while session and query features are useful for differentiating between good and bad abandonment, the inclusion of gesture features can successfully be used to improve good-abandonment detection.

8. CONCLUSIONS

This paper proposed the use of gesture features for differentiating between good and bad abandonment in mobile search. We sought to answer three research questions, the findings of which we summarize below.

RQ1: *Do a user's gestures provide signals that can be used to detect satisfaction and good abandonment in mobile search?*

By formulating a supervised classification experiment, we showed how user gesture features perform significantly better than query and session features as well as other click-based and optimistic baselines. We show this on a high quality dataset collected through a user study and verify the results on a crowdsourced dataset.

RQ2: *Which user gestures provide the strongest signals for satisfaction and good abandonment?*

Through a correlation analysis, we showed how time spent interacting with answers on a SERP are positively correlated with satisfaction and good abandonment. By contrast, swipe interactions and time spent interacting with organic search results are negatively correlated with satisfaction.

RQ3: *What SERP elements are the sources of good abandonment in mobile search?*

By analyzing data collected through our user study, we showed how good abandonment can be driven by many elements on a SERP, such as answers, snippets and images and conclude that good abandonment is due to many factors.

An interesting problem for future work would be to attribute the good abandonment to a specific entity on the screen. For instance, one might consider the attributed reading time for each element and use this information to infer which element led to good abandonment. Furthermore, it will be interesting to analyze how users' behavior differs in the presence of different entity types on the screen. This work has been performed exclusively on mobile devices, but many of the conclusions are likely transferable to tablet or desktop search; we leave this for future investigations.

Acknowledgements

We would like to thank Georg Buscher for his contribution on the difference between exogenous and endogenous signals, Sarvesh Nagpal and Toby Walker for their efforts in capturing the gesture-based interaction data, and Widad Machmouchi and Jin Kim for their insights into methods for measuring satisfaction in search.

References

[1] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz. Direct answers for search queries in the long tail. In *Proceedings of the 2012 ACM conference on Human Factors in Computing Systems*, pages 237–246, 2012.

[2] Y. Chen, Y. Liu, K. Zhou, M. Wang, M. Zhang, and S. Ma. Does Vertical Bring more Satisfaction ? Predicting Search Satisfaction in a Heterogeneous Environment. In *International Conference on Information and knowledge management (to appear)*, 2015.

[3] L. B. Chilton and J. Teevan. Addressing people's information needs directly in aha web search result page. In *Proceedings of the International Conference on World Wide Web*, pages 27–36, 2011.

[4] A. Chuklin and P. Serdyukov. Potential good abandonment prediction. In *Proceedings of the International Conference Companion on World Wide Web*, pages 485–486, 2012.

[5] A. Chuklin and P. Serdyukov. Good abandonments in factoid queries. In *International Conference Companion on World Wide Web*, pages 483–484, 2012.

[6] A. Diriye, R. White, G. Buscher, and S. Dumais. Leaving so soon? understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1025–1034, 2012. ISBN 9781450311564.

[7] M. Duggan and A. Smith. Cell Internet Use 2013, 2013. URL `http://www.pewinternet.org/2013/09/16/cell-internet-use-2013/`.

[8] W. Fan. On the optimality of probability estimation by random decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 336–341, 2004.

[9] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[10] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2): 147–168, 2005.

[11] Q. Guo and E. Agichtein. Ready to buy or just browsing? detecting web searcher goals from interaction data. In *Proceeding of the ACM International Conference on Research and Development in Information Retrieval*, pages 130–137, 2010.

[12] Q. Guo and E. Agichtein. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the International conference on World Wide Web*, pages 569–578, 2012.

[13] Q. Guo, S. Yuan, and E. Agichtein. Detecting success in mobile search from interaction. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*, pages 1229–1230, 2011.

[14] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 2050–2054, 2012.

[15] Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*, pages 153–162, 2013.

[16] A. Hassan. A semi-supervised approach to modeling web search satisfaction. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval.*, pages 275–284, 2012.

[17] A. Hassan and R. W. White. Personalized models of search satisfaction. In *Proceedings of the ACM international conference on Conference on information & knowledge management*, pages 2009–2018, 2013.

[18] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: User behavior as a predictor of successful search. *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 221–230, 2010.

[19] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2019–2028, 2013.

[20] A. D. Jeff Huang. Web user interaction mining from touch-enabled mobile devices. In *HCIR Workshop*, 2012.

[21] J. Jiang, A. H. Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. G. Kulkarni, and O. Z. Khan. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the International Conference on World Wide Web*, pages 506–516, 2015.

[22] J. Jiang, A. H. Awadallah, X. Shi, and R. W. White. Understanding and Predicting Graded Search Satisfaction. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 57–66, 2015.

[23] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my! In *Proceedings of the International Conference on World Wide Web*, pages 801–810, 2009.

[24] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundation and Trents in Information Retrieval*, 3(1-2):1–224, 2009.

[25] Y. Kim, A. Hassan, R. W. White, and Y.-M. Wang. Playing by the rules: mining query associations to predict search performance. In *Proceedings of the ACM international conference on Web search and data mining*, pages 133–142, 2013.

[26] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Comparing client and server dwell time estimates for click-level satisfaction prediction. In *Proceedings of the ACM Conference on Research & Development in Information Retrieval*, pages 895–898, 2014.

[27] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 193–202, 2014.

[28] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. *Proceedings of the ACM Conference on Research & Development in Information Retrieval*, pages 113–122, 2014.

[29] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[30] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*, pages 43–50, 2009.

[31] Z. Liao, Y. Song, L.-w. He, and Y. Huang. Evaluating the effectiveness of search task trails. In *Proceedings of the international conference on World Wide Web*, pages 489–498, 2012.

[32] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu. Different Users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*, pages 493–502, 2015.

[33] Y. Song, X. Shi, R. White, and A. H. Awadallah. Context-aware web search abandonment prediction. In *Proceedings of the ACM Conference on Research & Development in Information Retrieval*, pages 93–102, 2014.

[34] S. Stamou and E. N. Efthimiadis. Interpreting user inactivity on search results. In *European Conference on Information Retrieval*, volume 5993, pages 100–113, 2010.

[35] R. W. White, M. Richardson, and W.-t. Yih. Questions vs. Queries in Informational Search Tasks. In *Proceedings of the International Conference on World Wide Web*, pages 135–136, 2015.