Predicting Web Search Success with Fine-grained Interaction Data

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

Dmitry Lagun
Mathematics & Computer Science Department
Emory University
dlagun@emory.edu

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

ABSTRACT

Detecting and predicting searcher success is essential for automatically evaluating and improving Web search engine performance. In the past, Web searcher behavior data, such as result clickthrough, dwell time, and query reformulation sequences, have been successfully used for a variety of tasks, including prediction of success in a search session. However, the effectiveness of the previous approaches has been limited, as they tend to ignore how searchers actually view and interact with the visited pages. We show that fine-grained interactions, such as mouse cursor movements and scrolling, provide additional clues for better predicting success of a search session as a whole. To this end, we identify patterns of examination and interaction behavior that correspond to search success, and design a new Fine-grained Session Behavior (FSB) model to capture these patterns. Our experimental results show that FSB is significantly more effective than the state-of-the-art approaches that do not use these additional interaction data.

Categories and Subject Descriptors

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

General Terms

Design, Experimentation, Human Factors

Keywords

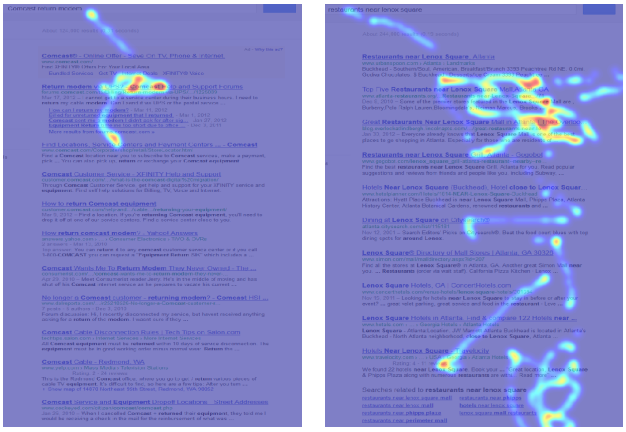
success prediction, search session, mouse cursor analysis

1. INTRODUCTION

Detecting and predicting searcher success is essential for automatically evaluating and improving Web search engine performance at scale [7]. Furthermore, real-time intervention and assistance could be provided if lack of success could be predicted earlier in the search session [2, 6].

While previous research has made great use of search behavior data such as result clickthrough, dwell time and sequences of searches [7, 2, 6, 1] for predicting searcher success, the effectiveness of the existing models is largely limited as they are agnostic about how users actually view and interact with the visited pages. For example, seeing a search session with a high result clickthrough rate, existing methods would typically consider that the searcher is very likely to be successful in her search. However, the clickthrough information should be coupled with time information to have a better understanding of the search success – for example, a shorter time before click may be indicative of higher confidence in the perceived relevance about the search result while shorter time spent on the landing page might suggest that the searcher found out that the document was actually not relevant.

Yet, the dwell time does not provide a full picture – spending a long time on a landing page might suggest that the searcher was struggling and could not find the relevant information if she was actually scanning instead of reading during the stay [5]. Sometimes, spending shorter time on a landing page might suggest that the searcher was actually successful if she quickly found the needed information. Similarly, spending some time carefully reading a result snippet before clicking on it is different from quickly scanning the whole search result page within the similar amount of time – in the latter case, the user appears to be less satisfied with the returned results and is less likely to be successful. Based on the time information alone, such search sessions might be considered successful and thus the search engine would not be able to improve on them; or, in an on-line setting, the search engine would not provide additional help when the searcher is struggling and becoming frustrated.



(a) successful *(b) unsuccessful*

Figure 1: Example mouse cursor heat maps: (a) - search result page in a successful search session; (b) - search result page in a unsuccessful search session.

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

Recently, *fine-grained interactions* such as mouse cursor movements and scrolling behavior on a Web page have shown to be valuable signals in inferring viewing behavior (e.g., [12, 4, 10]) and searcher preferences (e.g., [9, 5, 8, 13]). In this paper, we propose to model these richer observations to improve the estimation of searcher success. Figure 1 shows the mouse cursor heat maps overlaid on the visited search engine result pages from a successful and an unsuccessful search sessions respectively. As we can see, the mouse cursor positions in the unsuccessful session (Figure 1(b)) are more spread-out than in the successful session (Figure 1(a)) and spread to the lower part of the result page, suggesting that when the searcher examines multiple search results (especially the lower-ranked ones) before click, she is more likely to be unsuccessful in finding the needed information. Most closely related to our work, Guo and Agichtein [5] found that the fine-grained post-click behavior correspond to different viewing patterns and are indicative of document relevance. Complementary to previous efforts, this paper is the first to study the association between the Web *search success* and the *fine-grained behavior* such as cursor movements, and the first to develop a predictive model, FSB, that jointly models the session-level behavior patterns on both search engine result pages and pages on the search trails [14]. As the rest of the paper demonstrates, FSB achieves significant improvements for predicting search success over the state-of-the-art methods.

2. FINE-GRAINED SESSION BEHAVIOR (FSB) FEATURES

In this section, we describe our proposed *Fine-grained Session Behavior (FSB)* features to capture the page examination patterns that could be indicative of search success. In addition, we also include features from the queries, clicks and dwell time. The brief descriptions about some of the FSB features along their correlation with the success labels (Section 4.1) are reported in Tables 1, 2, and 3 and expanded below. Note that the features in the fine-grained interaction groups, namely, cursor and scroll, are first computed for each page view and then aggregated over the entire search session. These two groups can be further divided into pre-click and post-click sub feature groups, corresponding to behavior on the search engine result pages and the behavior on the pages in the search trail. Additional details about the full list of features in the cursor and scroll groups will be available online due to the space constraints.

Group	*Feature*	ρ
Query	*avg_qwords:* average number of words of the queries in the session	-0.117
	num_queries: total number of queries in the session	-0.522**
Click	*num_clicks:* total number of clicks in the session	-0.363**
	ctr_q: total number of clicks over number of queries in the session	0.150*
	ctr_s: total number of clicks over number of SERP views in the session	0.358**
Time	*tasktime:* total time duration in the session	-0.473**
	avg_time_s: average deliberation time on SERP pages in the session	-0.107
	avg_time_c: average dwell time on clicked landing pages in the session	-0.119
	satr_c: the ratio of satisfactory clicks (clicks with dwell time at least 30 seconds)	0.059
	dsatr_c: the ratio of dis-satisfactory clicks (with dwell time at most 10 seconds)	-0.083

Table 1: Coarse-grained behavior feature descriptions and Pearson's correlations with success ratings (indicates statistical significance at p < .01 level; * indicates statistical significance at p < .05 level).**

Query Features Query features derived from the query string itself, include the query length in words and characters, average number of characters of query terms, the number of submitted queries, SERP views, and unique queries. Intuitively, the longer the query, the more likely the task is difficult and the searcher ends up unsuccessful. On a session-level, the larger the number of queries users have to submit the more likely that the user is struggling. Notice the subtle difference between submitting a query and viewing a SERP – one query might correspond to multiple SERP views. The former can be captured from server-side logs while the latter can only be obtained with a client-side instrumentation.

Click Features Click features include the number of clicks and clickthrough rate (over queries and SERPs). Clickthrough generally is an indicator of success as searchers click on a document when they think that the document can satisfy their information needs. However, a large number of clicks, especially when paired with an even larger number of submitted queries, might indicate that the clicked documents are actually not relevant and the search goal is unsuccessful. Here, the clickthrough rate may provide additional evidence about search success.

Time-related Features We consider both the time users spend on the SERP and the pages on the search trail. In the literature [3, 7, 5], the former is referred as "deliberation time" while the latter is referred as "dwell time". Usually, these measurements of time are defined as the intervals, in seconds, between the time the page is loaded and the time the searcher leaves the page. To aggregate the time information across multiple page views in the session, we follow previous research [7, 6] and compute the total time span during the session, averaging time spent on different types of pages, and the ratio of clicks that result in SAT (dwell time $\geq$ 30 seconds) and DSAT (dwell time $\leq$ 10 seconds) [3].

Group	*Feature*	ρ
Cursor (SERP)	*avg_ymax_s:* average maximum y coordinate on SERP pages	-0.384**
	min_ymax_s: minimum maximum y coordinate on SERP pages	0.163*
	max_ymax_s: maximum of maximum y coordinate on SERP pages	-0.330**
	med_ymax_s: median of maximum y coordinate on SERP pages	0.073
	num_low_ymax_s: number of maximum y coordinate on SERP pages that are below 400 pixels	-0.138*
	num_high_ymax_s: number of maximum y coordinate on SERP pages that are above 800 pixels	-0.031
Cursor (Trail)	*avg_ymax_t:* average maximum y coordinate on search trail pages	0.179**
	min_ymax_t: minimum maximum y coordinate on search trail pages	0.253**
	max_ymax_t: maximum of maximum y coordinate on search trail pages	0.200**
	med_ymax_t: median of maximum y coordinate on search trail pages	0.289**
	num_low_ymax_t: ratio of maximum y coordinate on search trail pages that are below 400 pixels	-0.132
	num_high_ymax_t: ratio of maximum y coordinate on search trail pages that are above 800 pixels	0.361**

Table 2: Sample fine-grained cursor feature descriptions and Pearson's correlations with success ratings (indicates statistical significance at p < .01 level; * indicates statistical significance at p < .05 level).**

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

Group	*Feature*	ρ
Scroll (SERP)	*avg_ymax_s:* average speed of vertical scrolls on SERP pages	-0.318**
	min_ymax_s: minimum speed of vertical scrolls on SERP pages	0.069
	max_ymax_s: maximum speed of vertical scrolls on SERP pages	-0.331**
	med_ymax_s: median speed of vertical scrolls on SERP pages	-0.074
Scroll (Trail)	*avg_ymax_t:* average speed of vertical scrolls on search trail pages	-0.087
	min_ymax_t: minimum speed of vertical scrolls on search trail pages	-0.071
	max_ymax_t: maximum speed of vertical scrolls on search trail pages	-0.068
	med_ymax_t: median speed of vertical scrolls on search trail pages	-0.131

Table 3: Sample fine-grained scroll feature descriptions and Pearson's correlations with success ratings (indicates statistical significance at p < .01 level; * indicates statistical significance at p < .05 level).**

Vertical Scrolling Features In addition to modeling the overall amount of scrolling, we also model the frequency and speed of scrolling behavior, as well as the overall scroll distance and range in pixels, following [5]. The intuition behind is to capture the searcher's examination patterns. For example, high frequency and speed of scrolling may indicate that the searcher was "scanning" or skipping parts of the document, while a moderate range of scrolling with low speeds may indicate that the searcher was "reading".

Aggregation of Interaction Features We explore different strategies in aggregating the page-level features, including computing the mean, median, minimum, and maximum of all the page views and counting the number of page views that meet some specific requirements. The four statistics are more generic treatments of aggregation, with the mean or average more frequently used, median more robust to outliers, and minimum and maximum capture the extreme behavior. The threshold-based counting aggregation is more customized towards individual features, which may result in more effective predictors when appropriately applied. This approach may require a deeper understanding of each individual feature to define meaningful thresholds. We will discuss the different strategies in more depth in Section 4.1 and compare their effectiveness.

Note that aggregation over an entire section might be problematic when a search session consists of multiple sub-goals, in which case the searcher behavior may exhibit larger variations. To address this issue, search goal boundary detection algorithms (e.g., [11, 7]) can be applied to ensure aggregation is over single search goal. Alternatively, one may also consider aggregation over each search trail [14], which may also reduce the variance that comes from different types of pages. In this paper, our aggregation is on a single search goal as each session in our dataset consists of one single search goal (Section 3). While further improvement may be possible (e.g., through trail-level aggregation), as we show later, this goal-based aggregation formalism already results in effective models (Section 4).

3. DATA

The data set we used for our experiments, which has hundreds of search tasks and explicit relevance judgements of visited Web pages, is from a user study conducted by researchers at the University of Massachusetts [2]. The usage data of the participants was tracked, containing the URLs the searchers visited, the fine-grained interactions with the browsed pages, such as clicks, cursor movements, and scrolling, the time-stamp of each page view and interaction was also recorded. The search tasks in the user study were designed to be representative of Web search and difficult to solve with a search engine (i.e., the answer was not easily found on a single page). This is particularly valuable as these more difficult and long-tailed search tasks are the main challenge for the state-of-the-art search engines, and an accurate success prediction algorithm would enable search engines to evaluate and improve performance in these search tasks at a large scale.

The original dataset is publicly available online [1]. Similarly, the processed data and source code for this paper is available at `http://ir.mathcs.emory.edu/data/CIKM2012/`. Next, we describe the details of the user study and the collected data (additional information can be found along with the original dataset).

Explicit Judgements: Each time the participants completed a search task, they were asked the degree to which their information need of the task was satisfied during the entire search session on a five point scale ("1" indicates the search session "did not satisfy the information need in any way" and "5" indicates that the search session "completely satisfied the information need").

We used this self-reported explicit judgement as our ground truth for search success. A total of 211 search tasks were completed and provided feedbacks from 30 participants, with 463 queries submitted and 711 pages visited.

4. EXPERIMENTS

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

4.1 Feature Association with Search Success

We now discuss the association between each individual session feature and the explicit success judgements. Specifically, we computed Pearson's Correlation for each feature and conducted statistical significant testing. The results are summarized in Tables 1, 2 and 3. We organize the discussions by feature groups and compare alternative session-level aggregation strategies and sources of evidence (e.g., pre-click vs. post-click) when appropriate.

Query, Click, Time As we can see from Table 1, the average query length is negatively correlated with the session length, though not significant, while the number of submitted queries exhibits much stronger negative correlation of -0.522, confirming our intuition that the longer the queries the user had to submit and the larger

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

the number of queries, the more likely that the user was struggling and more likely to fail.

The number of clicks turns out to be negatively correlated with search success, which may seem counter-intuitive as clickthrough is typically considered as a signal of finding relevant information. One explanation is that the large number of clicks may come from the large number of queries. Indeed, divided over the number of queries, the clickthrough rate measures results in positive correlations, with the ratio computed over SERP views much more significant. This suggests benefits of client-side instrumentation.

As for the time measures, it turns out that the overall time span of a session exhibits the most significant correlation of -0.473, which makes sense as it characterizes the session length as the number of queries and clicks do. Somewhat surprisingly, the average dwell time on landing pages is negatively correlated with search success. One explanation is that as the task difficulty increases, users need to spend on average longer time to find the information on a page. The SAT and DSAT clickthrough rates, in contrast, match our intuitions and exhibit positive and negative correlations with success. However, the correlations are not significant, which may also be explained by the fact that the search tasks in our dataset are relatively more challenging.

Cursor Movements The analysis of this feature group is given in Table 2. For the simplicity of discussion, we only focus on analyzing the most discriminative feature of this group – maximum y coordinates and compare the different aggregation functions on this feature as well as two different sources of evidence, namely the pre-click behavior on search engine result pages (SERP) and the post-click behavior on the search trail pages.

For the SERPs, the averaging function for cursor (*avg_ymax_s*) appears to be most effective, which is substantially stronger than the deliberation time counterpart of cursor *avg_time_s* (Table 1). Interestingly, the minimum aggregation function (*min_ymax_s*) results in a significant positive correlation of 0.163. Note that very small maximum y coordinate suggests abandonment of search results and the minimum aggregation function of maximum y coordinate to some extent quantifies the likelihood of abandonment.

For the search trail pages, the averaging function (*avg_ymax_t*) only results in a moderate significant positive correlation of 0.179, which is stronger than its dwell time counter-part (*avg_time_s*) with a negative insignificant correlation of only -0.107 as shown in Table 1. Interestingly, the averaging function does not seem to be the most effective for the search trail pages. Instead, median function appears to be the most effective statistic, likely due to its robustness to outliers and larger variance. The best aggregation function for this feature turns out to be counting with meaningful thresholds. For example, the number of "SAT" page views, whose maximum y coordinate is on or above 800 pixels (*num_high_ymax_t*), exhibits a substantially stronger correlation of 0.361. The gain is significant compared to its dwell time based counterpart (*satr_t*), whose correlation is only an insignificant 0.09. Similarly, the number DSAT page views, whose maximum y coordinate below 400 pixels(*num_high_ymax_t*) exhibits stronger correlation than its dwell time counterpart (*dsatr_t*). These observations match the finding in [5] that post-click cursor features such as maximum y coordinates have stronger association with relevance as compared to dwell time. However, as we haven seen, careful selection of the session-level aggregation functions could have significant impacts on the predictive power of such a page-level feature.

Vertical Scrolling The analysis of this feature group is given in Table 3. Similar to the cursor feature group, we focus only on analyzing the most discriminative feature of this group – the scroll speed, to illustrate the differences in the various aggregation options as well as the patterns in pre-click and post-click pages. Overall, the correlations of different aggregations of sources of evidence all result in negative correlations for scroll speed, while the correlations are stronger for SERPs than trail pages. This may be explained by the larger variance in the types of trail pages – some of which may be shorter and do not require scrolling. As a result, scrolling may be more sparse and less reliable than the cursor features such as maximum y coordinate, as suggested by the overall weaker associations for search trail pages. In contrast, the correlations for scroll speeds on SERPs are much stronger with mean and maximum aggregation functions, resulting in significant correlations of -0.318 and -0.331, which is likely attributable to the relative fixed layout of search engine result pages, where user behavior tend to have smaller variance.

4.2 Predicting Search Success

Now we report our results and findings in predicting search success explicitly judged by the users using the different groups of features (Section 2). We formulate the success prediction problem as classification, and consider a search session with explicit success judgements (Section 3) equal to or larger than 4 as *successful* and *unsuccessful* otherwise. This definition of *search success* corresponds to the $Q^+R_*^+A^+V^?$ type of success according to the Query-Result-Answer-Verification (QRAV) model proposed by Ageev et al. [1], where a participant was satisfied with her search session and believed that she found an correct answer, without a verification whether the submitted answer was actually accurate.

The underlying success prediction algorithm used was logistic regression, which is a widely used generalized linear model for classification [2, 6], where the predictors can take different forms, such as continuous, discrete, dichotomous, or a mix of these. The response variable, in our case, whether the search goal was *successful* or *unsuccessful*, is not a linear function of the predictors but a logit transformation of their linear combination. Logistic regression has the advantages of simple implementation, good interpretability, and time-efficiency in training at scale.

For training and testing, we used 10-fold cross-validation with 100 randomized experimental runs. We evaluated our full model *FSB* [2], its four single feature group components: *query*, *click*, *time*, and *cursor*. We also evaluated the *cursor_serp* and *cursor_trail* sub-groups, which are based on the *cursor* feature group computed for the SERPs and trail pages respectively. Two baselines considered are a naïve Majority Baseline (*MB*) that always guesses the majority class *successful* and a state-of-the-art baseline *QCT* model trained on the Query, Click and Time feature groups. Feature group ablation analysis was also conducted by removing the single feature groups one at a time. Finally, we compare the three selected aggregation functions, namely, average (*avg*), median (*med*), and the threshold based counting function (*thres*). We report accuracy and weighted averages of Precision, Recall, and F1-measure over the two search success classes.

Single Feature Groups: The results are summarized in Table 4. The differences between different methods are statistically significant at .05 level under paired t-test except the difference between *cursor_serp* and *QCT* and the difference between *cursor_trail* and *click*. As we can see, our full model *FSB* significantly outperforms the two baselines as well as all the single feature groups. The *cursor* group performs the best among all the single feature groups and is the only single feature group that outperforms both of the two

[2]Some features such as the scroll group are excluded in feature selection. Details are available online.

baselines. The remaining single feature groups outperform the *MB* baseline but underperform the *QCT* baseline. Among the two cursor subgroups, the *cursor_serp* group is significantly more predictive than the *cursor_trail* group, which may be due to the more severe data sparsity and larger variance lies in the different search trail pages compared to the SERPs. Nevertheless, the combined feature group *cursor* significantly outperforms each of the two sources of evidence individually, suggesting the two are complementary.

Methods	*Acc (%)*	*P*	*R*	*F1 (% Imp.)*
FSB	**77.1**	**77.9**	**77.1**	**77.5 (+7.6%)**
cursor	75.3	76.0	75.3	75.6 (+5.1%)
cursor_serp	71.2	72.0	71.2	71.6 (-0.6%)
cursor_trail	65.8	65.8	65.8	65.8 (-8.6%)
query	68.9	68.7	68.9	68.8 (-4.4%)
click	66.3	66.2	66.3	66.3 (-8.0%)
time	70.1	70.5	70.1	70.3 (-2.4%)
QCT	71.8	72.3	71.8	72.0 (n/a)
MB	61.7	38.1	61.7	47.1 (-33.4%)

Table 4: Accuracy, Precision, Recall and F1-measure for the full FSB model, the single feature groups, and the QCT, MB baselines. The percentage of improvement over the QCT baseline is reported for the F1-measure.

Feature Group Ablation: The results are summarized in Table 5. The differences between the full model *FSB* and all the feature ablation methods are significant at .05 level under paired t-test except for *FSB-query*, suggesting all the feature groups except the *query* group contribute significantly to the full model even when other feature groups are presented. The largest decrease comes from removing the *cursor* feature group. Interestingly, even though *cursor_serp* seems to contribute more than the *cursor_trail* subgroup, the contributions from both of the two subgroups are statistically significant as supported by the fact that *FSB-cursor* significantly underperforms *FSB-cursor_serp* and *FSB-cursor_trail*.

Methods	*Acc (%)*	*P*	*R*	*F1 (% Diff.)*
FSB	77.1	77.9	77.1	77.5 (n/a)
FSB-cursor	71.8	72.3	71.8	72.0 (-7.3%)
FSB-cursor_serp	72.2	72.6	72.2	72.4 (-6.6%)
FSB-cursor_trail	73.7	74.5	73.7	74.1 (-4.4%)
FSB-query	77.3	78.0	77.3	77.6 (+0.2%)
FSB-click	74.8	75.7	74.8	75.2 (-2.9%)
FSB-time	73.3	73.9	73.3	73.6 (-5.0%)

Table 5: Accuracy, Precision, Recall and F1-measure for feature group ablation. The difference compared to the FSB full model is reported for the F1-measure.

Aggregation Functions: The results are summarized in Table 6. The differences between different methods are statistically significant at .05 level under paired t-test. As we can see, all the single aggregation functions underperform the full *FSB* model that utilizes all the three functions. Among the individual functions, *FSB (thres)* performs the best, followed by *FSB (med)* and *FSB (avg)*, showing the importance in selecting the aggregation functions.

5. CONCLUSIONS

In this paper we introduce a new model for representing the searchers' fine-grained session behavior (FSB) that captures not only information about the queries, clicks and time, but also fine-grained interactions both *before* and *after* clicking on a search result, such as cursor movements. To our knowledge, FSB is the first successful attempt to exploit and aggregate such "low-level" behavioral signals on the session-level, aiming to predict search success.

Methods	*Acc (%)*	*P*	*R*	*F1 (% Diff.)*
FSB	77.1	77.9	77.1	77.5 (n/a)
FSB (avg)	71.9	72.4	71.9	72.1 (-7.2%)
FSB (med)	72.8	73.4	72.8	73.1 (-6.0%)
FSB (thres)	73.1	73.8	73.1	73.4 (-5.5%)

Table 6: Accuracy, Precision, Recall and F1-measure for individual aggregation functions. The difference compared to the FSB full model is reported for the F1-measure.

Our experimental results show that these behavioral signals indeed correlate with searchers' explicit judgements of search success, and provide additional valuable information beyond queries, clicks and the amount of time users spend on the pages in a search session. We found that the different sources of evidence (i.e. behavior *before* and *after* a click) carry valuable complementary information about search success and that the feature aggregation choice was crucial. In combination, these signals enable FSB to exhibit significant improvement of predicting search success over the state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant IIS-1018321. The authors also thank Henry Feild for sharing the user study data.

6. REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proc. of SIGIR*, 2011.

[2] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proc. of SIGIR*, pages 34–41, New York, NY, USA, 2010. ACM.

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

[4] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. In *Proc. of CHI*, pages 3601–3606, New York, NY, USA, 2010. ACM.

[5] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proc. of WWW*, 2012.

[6] Q. Guo, R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais. Why searchers switch: understanding and predicting engine switching rationales. In *Proc. of SIGIR*, pages 335–344, New York, NY, USA, 2011. ACM.

[7] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *Proc. of WSDM*, 2010.

[8] J. Huang, R. W. White, G. Buscher, and K. Wang. Improving searcher models using mouse cursor activity. In *Proc. of SIGIR*, 2012.

[9] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proc. of CHI*, pages 1225–1234, New York, NY, USA, 2011. ACM.

[10] D. Lagun and E. Agichtein. Viewser: enabling large-scale remote user studies of web search examination and interaction. In *Proc. of SIGIR*, pages 365–374, New York, NY, USA, 2011. ACM.

[11] B. Piwowarski, G. Dupret, and R. Jones. Mining user web search activity with layered bayesian networks or how to capture a click in its context. In *Proc. of WSDM*, pages 162–171, 2009.

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

[13] R. W. White and G. Buscher. Text selections as implicit relevance feedback. In *Proc. of SIGIR*, 2012.

[14] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. In *Proc. of SIGIR*, 2010.