Explicit In Situ User Feedback for Web Search Results

Jin Young Kim
Microsoft
Redmond, WA USA
jink@microsoft.com

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

Nick Craswell
Microsoft
Redmond, WA USA
nickcr@microsoft.com

ABSTRACT

Gathering evidence about whether a search result is relevant is a core concern in the evaluation and improvement of information retrieval systems. Two common sources of evidence for establishing relevance are judgements from trained assessors and logs of online user behavior. However, both are limited; it is hard for a trained assessor to know exactly what users want to find, and user behavior only provides an implicit and ambiguous signal. In this paper, we aim to address these limitations by collecting explicit feedback on web search results from users in situ as they search. When users return to the search result page via the browser back button after having clicked on a result, we ask them to provide a binary thumbs up or thumbs down judgment and text feedback. We collect in situ feedback from a large commercial search engine, and compare this feedback with the judgments provided by trained assessors. We find that in situ feedback differs significantly from traditional relevance judgments, and that it suggests a different interpretation of behavior signals, with the dwell time threshold between negative and positive in situ feedback being 87 seconds, longer than the more common heuristic of 30 seconds. Using text feedback from users, we discuss why user feedback may differ from editorial judgments.

CCS Concepts

• **Information systems→Evaluation of retrieval results**

Keywords

IR evaluation, explicit user feedback, web search

1. INTRODUCTION & RELATED WORK

Web search engines are able to collect large scale behavioral log data as users search. This has enabled significant advances in search engine's ability to identify relevant content and evaluate various approaches. Because logs are collected in situ, they provide a rich picture of real people performing self-motivated searches. However, logs are not a direct source of relevance judgments. Instead they provide some implicit evidence of relevance, which must be interpreted with care.

One common way that search engines use log data is to try to determine if a search result is relevant. The most obvious approach is to assume that any search result that a user clicks after issuing a query is relevant, but it quickly becomes clear that users sometimes click on irrelevant results. To refine this measure, researchers began using the amount of time that a user spends engaging with the clicked result before performing their next action, referred to as their *dwell time* on a result, as a signal of relevance. A popular heuristic for this is to consider search results with a dwell time equaling to or exceeding 30 seconds as relevant [3, 6]. While experimental results suggest that dwell time can serve as a proxy for relevance or interest [2, 3, 5], researchers have shown that it can be impacted by characteristics of the query [8], task [1, 5], or search result page [6, 8].

To better understand a user's intent behind observed web search log data, it is useful to correlate the implicit data with some sort of explicit ground truth. A number of early studies [2, 3, 5] did this by instrumenting people's browsers to actively request feedback from real users in situ as they performed real self-motivated tasks. For example, Claypool et al. [2] developed a custom browser that was used by 75 students during undirected web browsing in a lab setting. Over the course of several days, these students provided feedback by answering the question, "How interesting is this page?" each time they left a page. Other researchers [3, 5] collected in situ feedback in the context of web search. Kelly & Belkin [5] provided seven graduate students with laptops installed with custom logging software to collect feedback related to their search task and the usefulness of the content they found over the course of 14 weeks. Fox et al. [3] provided a web browser plug-in to 146 Microsoft employees that popped up a dialog box requesting explicit feedback whenever a participant searched. A general finding across this research is that users dwell longer on web content they find relevant or interesting. We take a similar approach to collect explicit feedback from search engine users, updating these findings with insight from a more recent search system, as web search has changed a lot in the past 10 years. Additionally, because we collect feedback within the context of an existing search system we are able to capture a much broader array of users, versus just students or employees at a single institution.

Of course, there are drawbacks to collecting explicit feedback from users in situ. The feedback request may be missed or ignored by many users. Giving feedback may be inconvenient and disruptive to the search process, possibly impacting the logged user behavior. One way to address the challenges with collecting user feedback or interpreting log data is to instead collect explicit relevance judgments from expert raters. For example, Kim et al. [6] develop a rich model of dwell time by collecting relevance judgments from trained assessors for thousands of queries. However, we know that personal context (task [5], individual preferences [9], etc.) matters a lot, making it hard to provide external judgments. For this reason, we look at the difference between external relevance judgments from trained assessors and in-situ feedback from the actual user.

This paper presents an illustration of the different ways to understand large scale behavioral data via in situ relevance feedback and judge ratings, using dwell time as an example. We begin by describing our approach to collecting data – including implicit log data, explicit in-situ user feedback, and relevance judgments from trained assessors – and show that there are significant differences between a user's explicit feedback and that of trained assessors. We look at both data in terms of dwell time,

SIGIR'16, July 17–21, 2016, Pisa, Italy .
 ISBN 978-1-4503-4069-4/16/07…$15.00.
DOI: http://dx.doi.org/10.1145/2911451.2914754

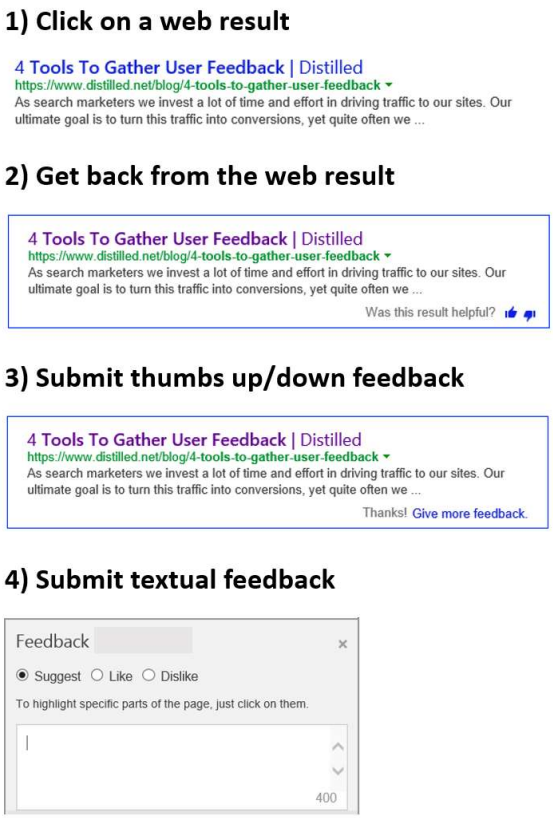

Figure 1. The user experience for collecting in situ feedback.

and suggest based on the data we collect that the current dwell time cut-off for satisfaction in search should be extended to 87 seconds.

2. METHODOLOGY

Here we describe the data collection method for our analysis. We focus on the in-situ feedback collection, which a novel aspect of our work. To explore dwell time from many different perspectives, we collected data for the same set of query-URL pairs from three different sources: search engine logs, in situ user feedback, and relevance judgments from external trained assessors.

2.1 Search Engine Log Data

We analyze the behavioral logs collected by a major internet search engine. We sampled five days of log data from September 2015 for users in the United States English language locale, collecting data for 4,738,204 total search result page views (see Table 1). The sample was filtered to remove bots, spam, and outliers. The logs contain anonymous user ID, time stamp, query, and results clicked. To determine how long a user dwelt on a search result page from the logs, we use the time stamp of the result click and the timestamp of any subsequent behavior the user has with the search within 30 minutes. Note that this means we are only able to identify the dwell time for results where the user has a subsequent interaction with the search engine.

2.2 In Situ User Feedback

In addition to collecting implicit behavioral information from users, the search engine was also instrumented to solicit in situ explicit feedback from users.

Table 1. Statistics of the Feedback Data

Statistics	Count	Ratio
# Overall	4,738,204	N/A
# Request	370,126	7.81%
# Response	1,044	0.28%
# Comments	18	1.7%

2.2.1 Feedback User Experience

Figure 1 shows how this was feedback was collected. As is the case with dwell time, it is only possible for a search engine to collect feedback from a user after having visited a search result is the user returns to the search engine to interact with it. For this reason, we created a feedback experience for users following a back button click. A user first clicks on a web search result (step 1) and visits that result. When the user returns to the search engine results page (step 2), the result they clicked is augmented with the question, "Was this result helpful?" along with a thumbs up and thumbs down button. To draw the user's attention to the feedback request, the result is outlined with a blue bounding box. Users are free to ignore the request, but if they click thumbs up or down (step 3), we request additional comments with the prompt, "Give more feedback." Again, users can ignore this additional request, but if they click it (step 4), we show a comment box in the bottom right corner of the screen.

The feedback experience is designed to be lightweight and unobtrusive. For example, while previous studies have offered between 3 [3] and 7 [5] options for feedback, we limit the interface to a simple binary to make it familiar and easy for users. It is only visible when users return after clicking, and only on the clicked result. We also stop showing feedback requests for the rest of the session after the user responds.

2.2.2 Feedback Data Collection

We collected feedback from a fraction of real users of a large commercial search engine. As summarized in Table 1, we displayed 370,126 feedback requests out of a sample of more than 4.7 million, touching 7.8% of the traffic in this sample. Among the feedback requests we made, we received 1,044 responses, which translates into 0.28% response rate. All of the results for which a user provides a thumbs up feedback we consider "satisfactory," and all results that receive a thumbs down are considered "unsatisfactory." Among the thumbs up/down feedback responses, we further received textual comments in 18 cases (1.7%). While the response rate is low, it is consistent with the fact that collecting in situ feedback from people is challenging. For example, the search engine displays a generic "Send Feedback" button at the bottom of each search result page, and receives feedback on less than 0.00001% of all pages displayed.

2.2.3 Limitations to the Feedback Collected

While the collected feedback represents real people's experiences on real search tasks, there are some limitations of data collected in this manner. Of particular note are potential query and user biases.

Potential query and result biases: Because the search engine can only instrument the results page, feedback is limited to cases where the user returns to the search engine after clicking. While over half of all queries involve a return [7], many queries (including most navigational queries) do not. Previous work (e.g., [3]) has addressed this issue by requiring people to install software. This

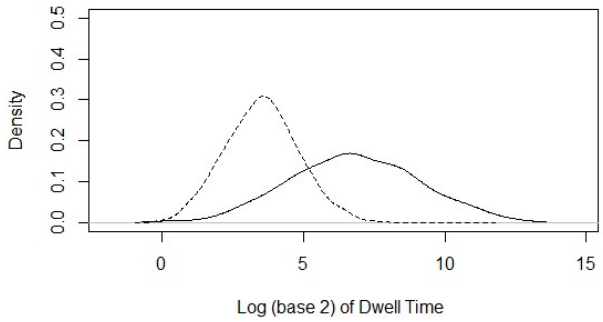

Figure 2. Density estimation for dwell time (in log scale) on web results for satisfactory results (solid line) vs. unsatisfactory results (dotted line)

enables feedback to be captured for all queries, but limits the set of users from whom feedback can be captured.

Potential user biases: Because users are able to decide whether or not to provide feedback, there is a selection bias among users. As shown in Table 2, We see that the subset of users who provided feedback tended to interact more with the results, with much higher chances to click on SERP. However, their dwell time behavior is similar (median dwell time of users who did not provided the feedback: 59 second, who provided any feedback: 62 second), and that is the area of focus for this paper.

Table 2. Comparing the aggregate behavior of user groups

Response to Feedback	% of pages with any click on SERP	Median dwell time on web results
Yes	88.4%	62 seconds
No	55.2%	59 seconds

2.3 External Assessor Relevance Judgments

To compare in situ user feedback with relevance judgments provided by trained assessors, we also collected external relevance judgments for the 1,044 query-URL pairs for which we received in situ user feedback. The assessors were trained for this kind of labeling, receiving instructions to rate the relevance of the URL for given query in 5-point scale representing: Perfect, Excellent, Good, Fair, and Bad. In doing so, judges are asked to imagine a population of possible users who issued the query, considering multiple possible intents behind the query and their relative likelihood.

3. DATA ANALYSIS

We correlate the in situ feedback we collected with user behavior observed in the logs, focusing on dwell time, and then look at the dwell time threshold suggested using external assessor relevance judgments. We see that the ideal dwell time to separate satisfied and unsatisfied clicks is different depending on the type of judgment used, and show that the reason for this is that the in situ feedback looks very different from external relevance judgments.

3.1 Dwell Time Using In Situ Feedback

Online user behavior may be used to understand various aspects of the search service. Here we focus on dwell time as an indicator of the user's satisfaction with the page. We know that users can spend a long time and still be dissatisfied, or vice versa. Our goal is to determine how accurate dwell time is as the measure of user satisfaction using our in situ user feedback.

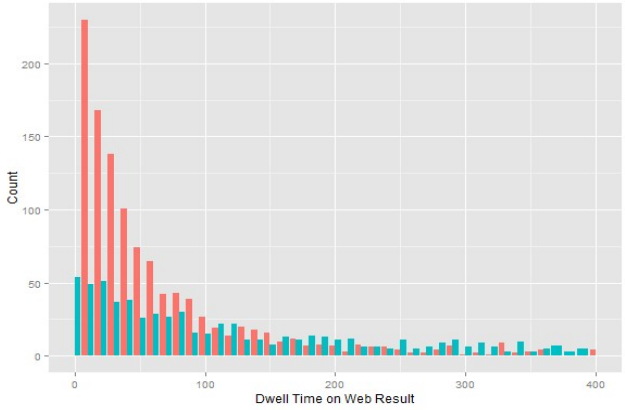

Figure 3. Histogram of dwell time distribution for satisfactory (green) vs. unsatisfactory results (red)

We refer to the thumbs up case as a satisfied (SAT) user and the thumbs down case as a dissatisfied (DSAT) user. Figure 2 shows dwell time density estimates (in log scale) for the two cases, with SAT as a solid line and DSAT as a dotted line. It is clear that the dwell time distribution is skewed toward right for SAT results. The median dwell time for SAT results is 110 seconds, 34 seconds for DSAT results.

But can the dwell time alone clearly separate between satisfactory vs. unsatisfactory results? In Figure 3 we plotted the histogram of dwell time distribution for SAT results (green) vs. DSAT results (red). Here, while there is clear trend of satisfactory results having relatively longer dwell time, there are some fraction of results where users are satisfied with less than 30 seconds of visit, and dissatisfied after spending quite a long time. This confirms the

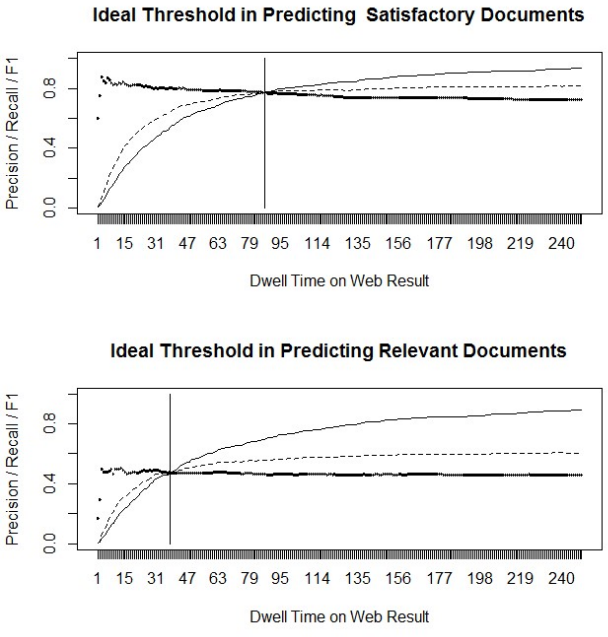

Figure 4. The precision-recall curve when using dwell time threshold to predict in situ user feedback (top) or external judgments by trained assessors (bottom).

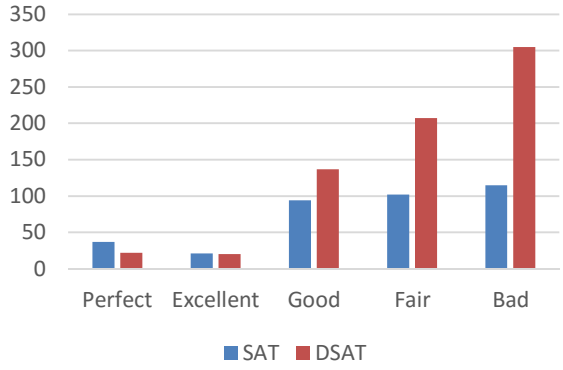

Figure 5. Comparison of in situ user feedback with external judgments from trained assessors.

conjecture that time is an important, but not a sufficient indicator of user satisfaction.

One way to find the ideal threshold is to consider the dwell time threshold as the classification boundary between SAT vs. DSAT cases, where we consider any impression with dwell time above the threshold as SAT, and vice versa. When we try to predict user satisfaction, we found that the threshold which gives the best F1 score is 87 seconds as shown in a dashed lines in Figure 4 (top).

3.2 Dwell Time Using External Judgments

Using the same analysis we used to extract the ideal dwell time based on the in situe feedback we collected, we also look at the ideal dwell time that can be calculated using the external assessor relevance judgments. To map 5-scale label to SAT and DSAT, we binarized the judgments to consider Perfert, Excellent, Good and Fair labels as SAT, and Bad as DSAT. Here we find that the threshold was 38 seconds, which roughly agrees with the threshold known in literature [3] yet is significantly shorter than the 87 seconds based on user feedback as shown in Figure 4 (bottom).

3.3 In Situ Feedback vs. External Judgments

To understand why the dwell time differs so significantly as a function of the data we use, we look more closely at the data we collected via in situ feedback and judgments from trained assessors. Figure 5 shows how SAT / DSAT labels from in situ user feedback compares with 5-scale labels from trained assessors. Overall, we can see the increase in DSAT labels as we move into lower scale in relevance judgment. However, there are considerable fraction of DSAT feedback in Perfect judgment, and vice versa for Bad judgment.

As was the case for user behavior, the results indicate a reasonable correlation with user satisfaction, yet the degree of disagreement is nontrivial. Among many possible explanations, the most plausible one is that the nature of human judgment where judges are asked to target the satisfaction of user population. For instance, we found several instances where user expressed dissatisfaction when we returned the right result for queries with clear navigational intent. (e.g., www.facebook.com for query 'facebook'). This illustrates that users in situ have idiosyncratic search intents which are not easy to be predicted by external judges.

Some of text comments we received following feedback reveal insights for why user feedback may not match the external judgments. There was a user who left a negative feedback on a page about a nurse practitioner after typing the name of the nurse. The comment said, "No phone number listed," implying that the user was looking for phone number. (Judge rating: Fair.) Also, there was a query {what is a connection string} which resulted in a click on a page with reasonable explanation of the concept. But the textual user feedback said, "No example to go with your instruction." (Judge rating: Good.) Examples like this show that user feedback can reveal varied nuances on users' intent which would not have been clear from the query.

4. CONCLUSIONS

We discussed the limitations to the use of human judgement or online user behavior for improving information retrieval systems, and addressed these limitations by collecting explicit feedback for web search results from searchers in situ as they search. When users return to the search result page after having clicked on a result, they are given an opportunity to provide a binary thumbs up or thumbs down judgment and text feedback. Our experiment with large commercial search engine reveals that this high-quality in situ feedback differs significantly from relevance judgments inferred using human judgment or online behavior. Using search result dwell time as an example, we show that the best possible cutoff to identify relevant documents is 87 seconds, while the 30 second limit that is commonly used in the literature risks accidentally identifying irrelevant documents as relevant. Based on users' comments, we also find that individual user's feedback can be different from labels from trained judges who aim to predict the satisfaction of user population as a whole.

5. REFERENCES

[1] Chilton, L. and Teevan, J. (2011). Addressing information needs directly in the search result page. In *Proceedings of WWW 2011*.

[2] Claypool, M., Le, P., Wased, M. and Brown, D. (2001). Implicit interest indicators. In *Proceedings of IUI 2001*.

[3] Fox, S., K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve Web search. TOIS, 23(2), 2005, 147-168.

[4] Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. TOIS 25(2), 2007.

[5] Kelly, D. and Belkin, N. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of SIGIR 2004*.

[6] Kim, Y., Hassan, A., White, R.W. and Zitouni, I. (2014). Modeling dwell time to predict click-level satisfaction. In *Proceedings of WSDM 2014.*

[7] Lee, C.J., Teevan, J. and de la Chica, S. (2014). Characterizing multi-click behavior and the risks and opportunities of changing results during use. In *Proceedings of SIGIR 2014*.

[8] Liu, C., White, R.W. and Dumais, S.T. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of SIGIR 2010*.

[9] Teevan, J., Dumais, S.T. and Horvitz, E. (2010). Potential for Personalization. TOCHI, 17(1).