Modeling and Predicting User Behavior in Sponsored Search

Josh Attenberg
Polytechnic Institute of NYU
Brooklyn, NY 11201
josh@cis.poly.edu

Sandeep Pandey
Yahoo! Research
Sunnyvale, CA 94089.
spandey@yahoo-inc.com

Torsten Suel
Polytechnic Institute of NYU
Brooklyn, NY 11201
suel@poly.edu

ABSTRACT

Implicit user feedback, including click-through and subsequent browsing behavior, is crucial for evaluating and improving the quality of results returned by search engines. Several recent studies [1, 2, 3, 13, 25] have used post-result browsing behavior including the sites visited, the number of clicks, and the dwell time on site in order to improve the ranking of search results. In this paper, we first study user behavior on sponsored search results (i.e., the advertisements displayed by search engines next to the organic results), and compare this behavior to that of organic results. Second, to exploit post-result user behavior for better ranking of sponsored results, we focus on identifying patterns in user behavior and *predict* expected on-site actions in future instances. In particular, we show how post-result behavior depends on various properties of the queries, advertisement, sites, and users, and build a classifier using properties such as these to predict certain aspects of the user behavior. Additionally, we develop a generative model to mimic trends in observed user activity using a mixture of pareto distributions. We conduct experiments based on billions of real navigation trails collected by a major search engine's browser toolbar.

Categories and Subject Descriptors: H.4.m [Information Systems Applications]: General

General Terms: Experimentation, Measurement

Keywords: implicit feedback, user behavior, sponsored search

1. INTRODUCTION

Recently, a great deal of effort in the research community has focused on improving user experience in web search through the incorporation of implicit user feedback [2, 13]. This feedback includes click-through behavior, dwell times on sites visited from query results, and other navigational behavior by search engine users. In general, implicit feedback provides valuable information about user satisfaction on web search engines.

While most prior work on implicit feedback has centered on organic search, we focus here on user navigation behavior associated with sponsored search. To our knowledge, this is the first work to focus on user behavior within sponsored search advertisements – ads which are displayed by search engines next to conventional organic search results. Revenue from these sponsored results provide much of the economic foundation of modern web search engines. However, sponsored search differs significantly from organic search in several important ways: First, while organic search is focused on satisfying users by addressing search queries, sponsored search has to optimize for ad revenue while accounting for user satisfaction and the constraints and objectives of advertisers. Second, the ranking of ads in sponsored search differs from that of organic search. Specifically, sponsored search ranking relies heavily on features such as predicted click-through rates and advertisement bid amounts, and less on hyperlink and anchortext information, which are frequently unavailable for short-lived ads. Third, click-through rates in sponsored search tend to be lower than for organic results, suggesting users might interact differently with these results than with organic results [8].

Our experiments rely on user browsing behavior collected from a navigational toolbar plug-in issued by a major search engine, comprising several million users who opted to share their browsing data. From this data, we use anonymized information about the queries submitted to search engines, the organic and sponsored results clicked on by users, and their subsequent behavior (including URLs and associated dwell times) on the visited sites. Analysis of this data leads to interesting observations. Perhaps the most compelling one is that in expectation, the CTR of ads does not have a strong correlation with what happens afterwards, in terms of click-activity or time spent. One possibly explanation for this counter-intuitive result is the proliferation of deceptive textual ad-snippets designed to entice users to visit sites, resulting in a high CTR but failing to effectively engage users once they visit. This hypothesis suggests that optimizing ad placement for high CTR does not necessarily imply the best user experience as is often assumed in the sponsored search community.

Above findings motivate the incorporation of implicit feedback, e.g., number of subsequent clicks on the site, and dwell time, into the ranking of advertisements. While many techniques have been proposed recently on how to best utilize implicit feedback in search [1, 2, 3, 5, 18, 19, 23, 25], the effectiveness of such proposals is often constrained by the limited availability of feedback data. For small search engines which do not have toolbar-like products, it is nearly impossible to gather a large amount of user behavioral information. Even major search engines can gather only a fraction of the data associated with a small subset of users. To address this issue, we demonstrate how user visit behavior on result sites depends on various properties of the queries, ads, sites, and users associated with that visit. We then go on to use such properties to predict certain aspects of the user behavior. Formally, given a query, result, and user, we are interested in modeling and predicting the post-result

KDD'09, June 28–July 1, 2009, Paris, France.

user behavior, in terms of the number of additional clicks made and the amount of time spent by users on the result. This prediction of seemingly complex user behavior is made even more difficult when one considers the proliferation of missing data that likely results in such a real-world classification task. We develop a robust method for predicting user behavior, which we evaluate thoroughly on different scenarios of information availability.

Furthermore, we present a generative model based on mixture of *pareto* distributions to describe user behavior. There are a large number of variables influencing the navigation of a user on query results. Different queries induce different probabilistic constraints on user navigation based on that query's intent. For instance, broad informational queries (e.g., camera shopping) require more browsing by users relative to more focused queries (e.g., finding a specific publication). By accounting for the varieties of behavior present in user activity, our model provides a better fit to observed data than the previous model in [14]. Overall, this paper makes the following contributions:

- A thorough analysis of sponsored search behavior. To understand the intricacies of sponsored search, we perform a high-level comparison of user visit behavior resulting from sponsored and organic query results.
- We analyze the influence of various factors on user behavior in sponsored search. Some of these observations are often counter to prior assumptions made on how users value online advertisments (e.g., high CTR implies better ads).
- We develop a novel generative model incorporating diverse trends in users' expected engagement and information need associated with issued queries.
- To deal with the sparsity inherent with user navigational information in sponsored search, we demonstrate experimentally that user behavior can be predicted with sufficient accuracy based on previously seen similar instances.

The rest of the paper is organized as follows. Section 2 discusses relevant previous research. Section 3 describes our data and experimental setup. Section 4 analyses user behavior on sites and compares behavior on organic and sponsored results. Section 5 explains our generative model for user behavior, and Section 6 provides a closer analysis of user behavior by controlling for several important factors. Our results on predicting of user behavior are presented in Section 7, and Section 8 provides some concluding remarks.

2. BACKGROUND AND PRIOR WORK

Sponsored search: There has recently been a large amount of research on sponsored search, i.e., how to best select ads to display next to search results, which is an important part of the emerging area of *Computational Advertising* [9]. While click-through behavior (and in particular CTR) is known to be an important factor in ranking sponsored search ads, we are not aware of any detailed studies of *post-click behavior*, knowledge useful for both advertiser and user satisfaction.

Using Implicit Feedback in Search: Several recent studies have focused on how implicit measures can be utilized to improve Web search [2, 13, 16, 22]. In [2] it was found that implicit feedback can improve the accuracy of a competitive search ranking algorithm by almost 31%. Various methods have been proposed for how to incorporate implicit measures into ranking. For instance, there is work on how to interpret click-through data accurately [15, 16], identify relevant websites using past user activity [1, 3, 5, 24], and rank pages based on user feedback [18, 19, 21]. Our work differs from the above in that we focus on sponsored search. Additionally, we not only analyze implicit feedback, but detail a prediction mechanism for forecasting user behavior in previously unseen scenarios. While [8] provides some notion of user behavior in the context of sponsored query results, it is with the intention of expanding textual similarity, a task sufficiently different from that tackled here.

Modeling User Behavior: Another line of work has focused on modeling user behavior [4, 10, 14]. We show that while our data conforms to a power law as in [14], the exponent of the distribution best fitting our data is substantially different from that predicted in prior work. While our study is based on search-induced behavior, [14] studied trails created from more undirected browsing. The work presented here adopts a novel pareto mixture-model based on query information need that is able to accurately fit observed user behavior.

3. EXPERIMENTAL SETUP AND DATA

In the past several years, browser enhancing plug-ins have seen wide-spread acceptance. These plug-ins are third party programs which modify the browser software to provide additional functionality when navigating the web. One particularly popular type of browser plug-in are *search toolbars*, which embed a search interface into the web browser. These toolbars typically send back to the engine various information about the user's navigational behavior, given the user's consent, and this information is used by the engines to constantly improve the quality of their search services, and in some cases also to personalize the results for the particular user.

The data used in our experiment was collected between January and July 2008, and represents a sample of users of a major search engine's search toolbar who opted into sharing their data. This sample contains roughly 4 million anonymous users, as identified by their associated browser cookies, and billions of individual page requests. Following the technique described by White and Drucker [25], we segment user navigation into post-query trails, i.e., the sequence of pages viewed as the direct consequence of following a query result. When creating these post-query trails, we introduce an additional criterion which terminates a trail upon navigation to a site other than that of the clicked query result as in [14]. Thus, we focus on the query results themselves and the implicit feedback that can be gleaned from user behavior on the corresponding sites only.

4. USER BEHAVIOR ON QUERY RESULTS

By monitoring the trails induced by post-result user navigation, we are able to compile a detailed understanding of how users interact with the sites that are offered to them. The focus of this section is to present trends in user browsing and information seeking behavior, and to offer some intuitive explanations of our findings when possible.

Many attributes could be collected to provide a quantitative summary of user behavior on query results. Following prior work in relevance feedback and user navigation, for each trail (originating from a search result page) we focus on: (a) the number of clicks the user makes on the trail (*trail length*), and (b) the total time spent in the trail (*trail duration*). These two numbers provide a useful synopsis of user navigation behavior. Additionally, to describe how individual sites are navigated in aggregate, we use the Shannon entropy describing the various trails that users take in the site (more details in Section 4.2). Using these simple features, we can obtain useful insights into the ways users interact with query results. While we will consider both organic and sponsored search results, we are particularly interested in the case of sponsored search results, which has been studied much less by previous work.

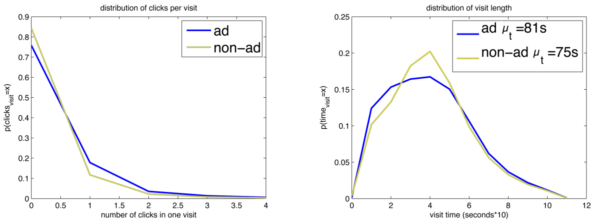
In Sections 4.1 and 4.2, we present and discuss the observed distributions for trail length, trail duration, and entropy taken from the accumulated organic and sponsored results. Section 4.3 studies the interdependence between these observed variables.

4.1 Trail Analysis

After partitioning our data into query trails, we investigate the distributions of the number of clicks made in a trail after landing on a query result (trail length), and the total time spent in the trail after landing on the result (trail duration). Figure 1(a) presents the distribution of trail lengths; not surprisingly most trails are very short, and in fact more than 70% of trails involve no additional click after the click-through on the search result. Excluding the initial click-through, the average number of clicks per trail is 0.39 for sponsored results and 0.25 for organic results. Thus, while sponsored results tend to have lower click-through rates than organic results (not shown here), once users clicks on sponsored results they are, on average, more active.

Figure 1(b) displays the distribution of trail durations. While sponsored search results tend to lead to more time spent on results than organic results – $82s$ versus $75s$ in expectation, we note that sponsored results are also much more likely to result in visits of $< 20s$ than organic results; i.e., users frequently click sponsored results and then leave almost immediately. This may be an indicator of the occasionally deceptive nature of the textual snippets designed by advertisers to be presented to users for sponsored search results. Advertisers realize that increased traffic to their site maximizes the number of potential customers and have developed expertise at engineering snippets to optimize click-through rate, possibly deceiving some users into thinking the resulting site will suit their needs. (In contrast, snippets for organic results are created not by the site owner, but automatically by the engine.)

We note that it is difficult to directly compare the expected behavior of users on organic and sponsored search results. Any given cause for initiating a web search may result in clicks on organic results, while only a subset of these scenarios tend to find users pursuing sponsored results. Thus, a query “car insurance” has a reasonable chance of a sponsored result being clicked, however, for a query such as, e.g., “mean of beta distribution”, it is much less likely that a sponsored result would be clicked or even offered. Considering that contrasting these data sets is difficult, we observe that the overall trend seems to be of similar overall shape; it is likely that similar processes underlie both organic and sponsored behavior, perhaps with different parameters or initial conditions. In the following, we focus primarily on the latter of these two, sponsored search, a subset of user behavior largely overlooked by prior work on user behavior.



(a) Distribution of Clicks (b) Distribution of Time Spent

Figure 1: Distributions of User Activity On Result Pages

4.2 Entropy of trails

The number of clicks alone is a rather course-grained feature for describing user behavior and engagement. Within any given site, there are likely many trails of a given click-length along which a user can navigate, with associated meanings for user and site owner. Five clicks on a trail describing how to file a complaint can be interpreted very differently than five clicks browsing and purchasing products on a retail site. The infinite possibilities of the web makes a large scale analysis of fine-grained user surfing behavior across many different sites difficult. However, we still seek to understand in a broad sense how users navigate on a query result site. One possible metric for this is the Shannon entropy H of the navigational history on a site S: $H = \sum_{\rho \in \{trails_S\}} -p_\rho \log_2 p_\rho$. Here ρ denotes a navigational trail in S, and p_ρ is the observed probability of a user taking that trail. This roughly translates to the number of bits needed to describe which trail a user has taken. Sites with a large entropy tend to see a wide variety of trails taken by users, while those sites with a very low H have most users take one of a few different trials, either by choice or due to site structure.

For both ads and non-ad results, we collect all sites receiving at least 50 visits and calculate the entropy as above. Figure 2 presents the distribution of relative site frequency of observed entropy values for organic and sponsored results. We notice that sites resulting from sponsored search results tend to have higher entropy than organic result sites. This is especially significant since there is a far greater number of organic observations, and entropy is dependent on the number of observations. Alternately, the longer expected trail length of sponsored results may contribute to the difference in entropies.

From the observed difference in entropy, one could conjecture differences in site function and associated query type (navigational, informational, transactional) between sponsored and organic results. For example, many advertisements are placed by large retail sites, where users can browse for many different products related to their interests. On the other hand, many organic sites are clicked in response to a simple, direct question, where only a single click is needed to satisfy the user.

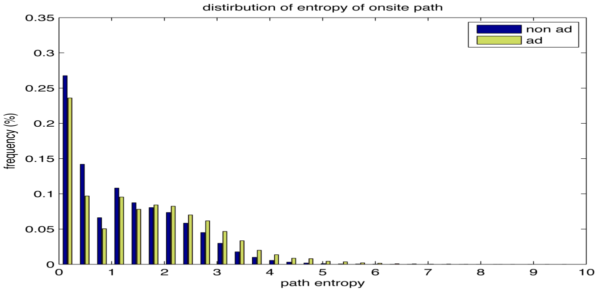


Figure 2: Distribution of Entropy for On-Site Behavior

4.3 Relation Between Time and Clicks

The number of clicks made and the time spent on a trail are highly correlated, as would be expected; it takes a certain amount of time for users to make successive clicks. More interestingly, we observe that the probability of a user making an additional click in the trail seems to be dependent on the time spent on the current page. Figure 3 shows the strong correspondence between the likelihood of making another click on the site ($\mathbb{P}(nextclick)$) and the dwell time (t) on the current page. We note that as the dwell time t increases, the probability of making the next click $\mathbb{P}(nextclick)$ also increases, for both sponsored and organic results.

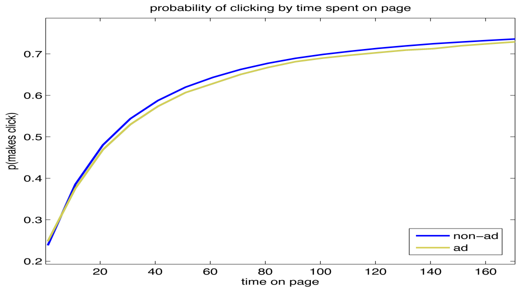


Figure 3: Probability of Another Click vs. Time Spent on Page

To study this further, we look at how $\mathbb{P}(nextclick|time \geq y)$ depends on the number of prior clicks made in the trail, i.e., $\mathbb{P}(nextclick|time \geq y, totclick = x)$. In Figure 4, we plot the

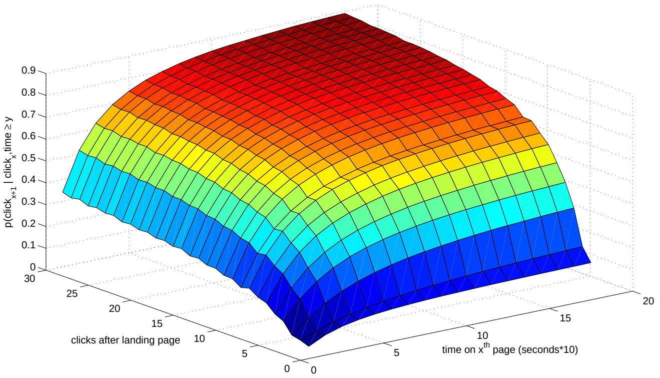


Figure 4: Probability of Subsequent Clicks Depending on Number of Previous Clicks and Time Spent on Current Page for Sponsored Results

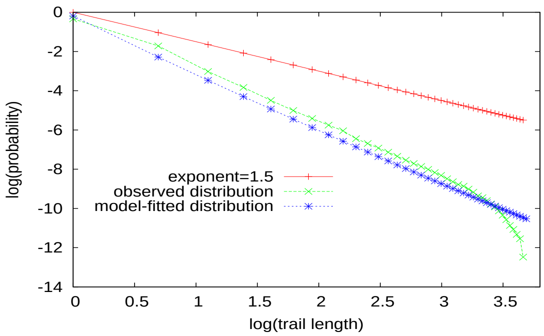


Figure 5: Distribution of Trail Lengths.

probability of a user making successive clicks conditioned on the number of prior clicks as well as the time spent on the current page. Note that the probability of making another click increases with both the number of previous clicks on the site and the time spent on the current page. One possible explanation is that of *increasing user engagement*, i.e., the probability of making additional clicks depends on some measure of the user's engagement with the site. Increased on-site activity in the form of clicks or time spent may indicate a greater chance of the user becoming engaged.

5. MODELING CLICK BEHAVIOR

Visits on query results constitute a diverse set of navigational trails. We observe that the distribution or trail lengths closely resembles a power law, i.e., the probability of observing a trail of length x, $\mathbb{P}(x)$, is proportional to $x^{-\alpha}$ where α is the *scaling exponent* of the power law. This observation is in accordance with the study by Huberman et al. in [14]. In Figure 5 we plot the observed distribution on the log-log scale. In this scale, it is evident that the distribution adheres to a straight line (whose slope is roughly equal to 3).

While both our study and [14] obtain a power law, we note that the exponent of our power law differs significantly from the exponent of 1.5 that was observed and theoretically derived in [14]. Figure 5 shows the power law with exponent 1.5 along with the observed distribution. Also shown in the figure is the distribution predicted by our model (discussed below), which provides a very good fit to the observed distribution.

Before delving into the details of our proposed model, we note the reasons for which we believe our observed power law differs from the one observed and predicted in [14]. First, the nature of the web has changed dramatically since the study in [14]. Second, our analysis focuses solely on trails originating from query result pages, while Huberman et al. focus on trails obtained from general browsing. Search-induced trails are likely to be shorter than random-surfing trails for two reasons: (i) typically, searchers seek specific information and when they find what they are looking for, they quickly end their trails, moving on to the next task in hand, while in [14], the assumption is that users continue browsing until the benefit (enjoyment) of the pages encountered becomes less than the "cost" of browsing, and (ii) in the case when users do not find the desired information following a search result, it is likely they go back to a different search result or reformulate the query to start a new trail.

5.1 Mixture of Power Laws

A cursory look at our data reveals that fitting one power law over the surfing behavior aggregated over millions of queries is inadequate, as it greatly oversimplifies complex human behavior. Intuitively, different queries have different information needs, thereby inducing very different types of click behavior. For instance, queries related to shopping entail much more browsing on behalf of users than queries on more focused tasks, e.g., finding a specific book or publication. To account for this diversity, we now propose a mixture model of user behavior based on queries [1].

Instead of one underlying power law, we assume that there is a mixture of power law distributions generating user behavior. In particular, our model consists of C clusters of queries. Each cluster, c, has its own discrete power law distribution with unknown parameter α_c. A particular c models a set of queries that have a certain information need and possess a characteristic click behavior best fit by α_c. Under the discrete power law, a user makes x clicks following a search result with probability $\mathbb{P}(x|\alpha_c) = f(x, \alpha_c)$ where $f(x, \alpha_c) = \frac{x^{-\alpha_c}}{\zeta(\alpha_c)}$ and $\zeta(\alpha_c) = \sum_{x=1}^{\infty} x^{-\alpha_c}$. Large values of α_c imply long user trails (e.g., broad queries requiring some amount of browsing), while small values of α_c imply short trails (e.g., more focused queries). The prior probability of a cluster to contain a query is π_c, where $\sum_{c=1}^{C} \pi_c = 1$. We face the problem of estimating these unknown parameters of our model, denoted by $\theta = \{\pi_c, \alpha_c\}_{c=1}^{C}$.

From the data we construct a vector for each query q where $q(x)$ denotes the number of trails of length x originated from q. Given this query vector, we can assign query q to the above mentioned clusters. We denote the probability that query q belongs cluster c by $\gamma_{q,c}$. Assuming that all visits to a query are drawn i.i.d. in accordance to that query's parameters, we can calculate $\gamma_{q,c}$ as:

$$\begin{aligned}\gamma_{q,c} &= \mathbb{P}(c|\theta) \cdot \mathbb{P}(q|\alpha_c) \\ &= \pi_c \cdot \prod_{x=1}^{\infty} f(x, \alpha_c)^{q(x)} = \pi_c \cdot \prod_{x=1}^{\infty} [\frac{x^{-\alpha_c}}{\zeta(\alpha_c)}]^{q(x)}\end{aligned}$$

Applying the law of total probability, we normalize $\gamma_{q,c}$ such that $\sum_{c=1}^{C} \gamma_{q,c} = 1$. Thus, the log-likelihood of the entire query data Q, given unknown θ, is:

$$\begin{aligned}ln\big(\mathbb{P}(Q|\theta)\big) &= \sum_{q \in Q} ln\big(\mathbb{P}(q|\theta)\big) = \sum_{q \in Q} ln\Big(\sum_{c=1}^{C}(\pi_c \cdot \mathbb{P}(q|\alpha_c))\Big) \\ &\propto \sum_{q \in Q} ln\Big(\sum_{c=1}^{C}\big(\pi_c \cdot \prod_{x=1}^{\infty} [\frac{x^{-\alpha_c}}{\zeta(\alpha_c)}]^{q(x)}\big)\Big)\end{aligned}$$

To optimize log-likelihood over the unknown parameters (π's, α's), we use the Expectation-Maximization (EM) algorithm [12]. In this paradigm, we iteratively improve our estimates on π and α, as shown in Algorithm 1.

[1] The model can be generalized to accomodate the influence of users and pages as well.

Algorithm 1 Exp. Maximization for Power Law Mixture Model

```
while convergence condition not met do
  E step:
    $\gamma_{q,c} = \frac{\pi_c \cdot \prod_{x=1}^{\infty} \left[\frac{x^{-\alpha_c}}{\zeta(\alpha_c)}\right]^{q(x)}}{\sum_{i=1}^{C} \left(\pi_i \cdot \prod_{x=1}^{\infty} \left[\frac{x^{-\alpha_i}}{\zeta(\alpha_i)}\right]^{q(x)}\right)}$
  M step:
    $\pi_c = \frac{\sum_{q \in Q} \gamma_{q,c}}{N}$
    where: $N$ is the total number of unique queries, and $\alpha_c$ can be estimated following the work of [11] as:
    $\alpha_c = 1 + \frac{\sum_{q \in Q} \left(\gamma_{q,c} \cdot q(x)\right)}{\sum_{q \in Q} \sum_{x=1}^{\infty} \left(\gamma_{q,c} \cdot q(x) \cdot \ln(2 \cdot x)\right)}$
end while
```

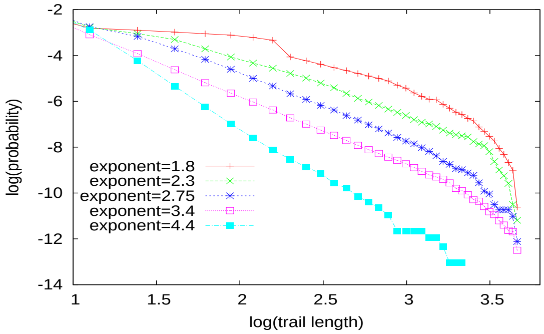


Figure 6: Trail distributions of the 5 clusters derived by EM.

We ran the EM alogorithm with $C = 5$ clusters on 2 million trails from the sponsored search data set. Figure 6 shows the trail length distribution associated with each cluster. (Each cluster contains at least 5% of the queries.) From the scaling exponents, it is evident that these clusters differ significantly from each other. This validates the hypothesis that different queries have different information intent and lead to vastly different user behavior. (The overall fit of our mixture model to the observed power law is shown in Figure 5.)

6. FACETED ANALYSIS OF BEHAVIORAL TRENDS

In the previous sections, we showed general patterns of user behavior in terms of number of clicks and time spent, and discussed possible models for this behavior. Of course, query topic, query intent, time of day, specialization, domain knowledge, the user's disposition, and countless other ingredients can contribute to how a user behaves after clicking on a query result. While considering each of these facets is impossible, in this section we perform a more detailed analysis by separately controlling for a few of the more interesting of these factors. In particular, we look at click-through rate, query topic, and several other properties of queries.

6.1 Click-Through Rate Vs. User Engagement

In the sponsored search community it is widely assumed that the ads with the highest CTR are the "best" ads – the high proportion of clicks has been interpreted as a testament to the site's quality and relevance to the user's needs [20]. It is unknown, however, if an increased click-through rate for a particular ad or site translates to more on-site activity per visit. This is an important consideration for advertisers and search engines; the results that users tend to click most frequently may not actually be the most useful results to the user, or the result leading to the most on-site activity or even purchases.

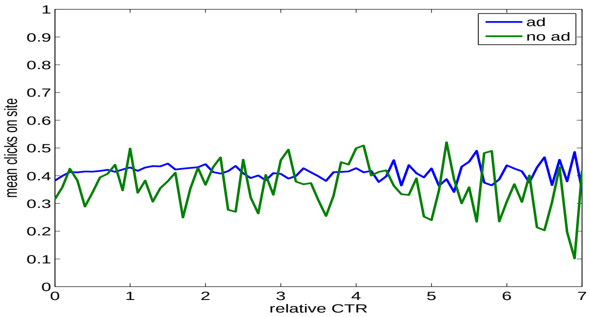


Figure 7: Expected On-Site Clicks Vs. Site Click-Through Rate

To investigate the presence of such a correlation, we compare the CTR of sites with the expected trail length and duration of visits to those sites, for sponsored and organic results. We compute the CTR of a site by dividing the number of times the site was clicked by the number of impressions for the site.[2] Figure 7 shows the mean trail lengths, depending on the site's CTR[3]. As shown in Figure 7, there is no obvious relation between the level of click-based activity and the CTR; web surfers do not seem to browse more on sites with a higher CTR. This appears to be true for both sponsored and organic results, with sponsored results resulting in slightly more activity across the range of CTRs. Measurements of query duration as influence by CTR appear very similar to Figure 7, showing very little if any noticeable correlation between CTR and trail duration. This is an interesting and somewhat unexpected result: While optimizing result placement based on CTR may optimize the payments to the search engine, from our data it seems that a higher CTR does not always lead to more activity on the site per visit. This phenomenon could possibly be explained by the proliferation of deceptive textual ad-snippets designed to entice users to visit sites, resulting in a high CTR but ineffective engagement of users once they visit.

6.2 Topical Influence On User Behavior

Previously, we have speculated that query topic likely influences the behavior of users on the site visited from the result page. In order to show the impact of query topic on site activity, we took the entire set of query trails culminating from sponsored search results and identified a query topic according to a proprietary, ad-centric topical taxonomy, using a specialized classifier. Then, within each topic, we calculated the average trail length and trail duration after landing on the result site. The results are shown in Figure 8 presents this comparison. (Note that each topic had a significant number of instances, and that overall data is roughly balanced across topics.)

As conjectured, the amount of activity does indeed vary across the topics we have considered, both for trail length and duration. This has possible implications for approaches that exploit user behavior to improve search results, in that one great care must be exercised when making comparisons across topic boundaries.

Second, and perhaps more interestingly, increased click activity in a topic is not necessarily associated with increased time spent. Note that queries in the *Travel* category tend to lead to significant on-site activity in terms of both number of clicks and time spent, while for *Finance* we have fairly small trail lengths but very long trail durations.

[2] While the clickability and thus CTR of an ad/site depends on many factors, including position on the result page, here we look at this simple definition of CTR, leaving more detailed analysis for future work.

[3] Here we present relative CTRs to preserve proprietary information.

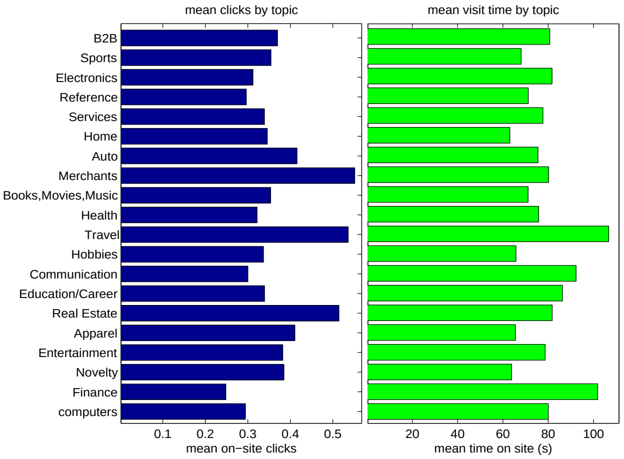


Figure 8: Distribution of Trail Lengths and Trail Durations According to Query Topic

6.3 Influence of Other Query Factors

So far, we have considered the influence of CTR and query topic on on-site activity. Next, we look at the impact of other factors, that, in the past, have often been associated with user browsing activity or page quality: (i) The ordering of the results clicked for the query. (ii) The number of query terms. (iii) The navigational vs. informational nature of the query. (iv) The PageRank of the clicked search results [6]. (v) The overall frequency of the query in our data set. These features were selected due to their intuitive influence on a user's browsing behavior or their use in prior work.

To demonstrate the impact that these facets may have on user browsing, we segregate data into sets according to the value of that facet. In order to maximize the illustrative ability permissible by such an arrangement, we partition data according to the boundaries made by a decision tree attempting to predict onsite activity based solely on the feature in question. After partitioning the data, we calculate the trail length for each bin. Due to lack of space, we do not prevent similar results for trail duration. The result of these experiments can be seen in Figure 9. The y-axis is the average trail length per bin (μ). Below we explain these features and associated observations in detail.

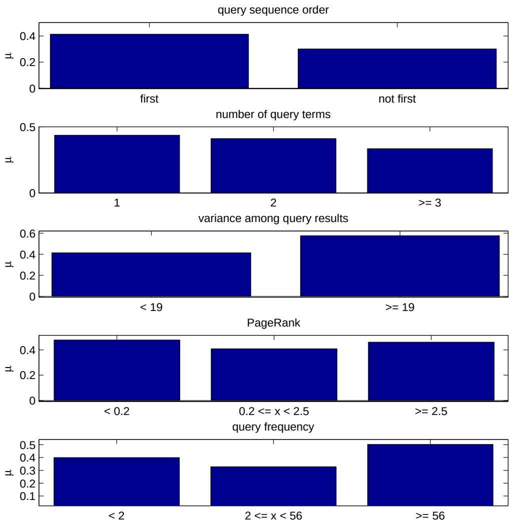


Figure 9: Influence of External Factors on Trail Length

Ordering of Results Clicked for a Query: Examining Figure 9, we notice that users tend to be more active on the first search result visited during a multi-result query session. In other words, trails originating from the first result that is clicked tend to be longer. This might be due to a loss of patience as the session carries on, or due to the fact that users first click on higher ranked and thus possibly better results. Additionally, sessions with many results clicked may simply not have any good results, therefore the need for many results to be visited. We note in this context that there is a known bias for users to click on the first result on a search result page, even if that result is worse than a lower-ranked result on the page; it is possible that this bias carries over to subsequent clicks on the site.

Number of Query Terms: Contrary to our initial assumptions, the amount of on-site surfing decreases as the number of terms in the query grows. Our expectation was that longer queries tend to be more specific than their shorter counterparts thereby leading to greater interest once a result had been chosen. There are several possible explanations. First, longer queries may lead more directly to the page that the user really wants, making additional clicks unnecessary. Second, long queries are often difficult to answer for search engines and may thus give worse results.

Navigational vs. Informational Nature of the Query: Broder [7] proposed a taxonomy of query goals into three main categories: navigational, informational, and transactional queries. We expect the query goal would influence the manner in which users navigate the results presented to them. Rather than follow the detailed techniques described in prior work (e.g., [17]) to automatically identify the category of a query, we use the simpler idea (also described in [17]) that clicks on navigational queries typically focus on one or only a few results (e.g., most clicks for the query "myspace" go to myspace.com), while the other two categories of queries may see a variety of results clicked by different users. Thus, we use the variance in the selected results for a query as a proxy for how navigational a query may be; queries with low variance see one of a few results chosen in most cases, and are therefore more likely to be navigational.

Looking at the result in Figure 9, we see that non-navigational queries tend to lead to more click activity than navigational queries. Non-navigational queries are often more exploratory in nature; a user may not know exactly what is wanted or may require some orientation when searching for the correct information [23]. Non-navigational results that are product-oriented may also lead to browsing to facilitate comparison shopping: a user searching for "hard drive" is likely to examine similar products before making a final decision.

Pagerank of Clicked Search Results: PageRank is an enormously influential static ranking of page importance based on the hyperlink structure of the web, and has occasionally been considered a proxy for page quality [6]. To study the impact of Pagerank on site activity, we computed Pagerank values for a site-based web graph. As we see, Pagerank seems to have fairly little impact on the surfing behavior of users when visiting a site via query results. Figure 9 shows the weak correlation between PageRank and click-activity. In fact, search results with the lowest values of PageRank seem to have slightly more activity than results with moderate PageRank. Results with low PageRank may be sufficiently specialized to prevent the wide-spread attention necessary for a high PageRank, however, this specialization may attract a lot of interest from users visiting that site.

Query Frequency: On the other hand, query frequency seems to have a substantial impact on the amount of on-site activity, with the most frequent queries resulting in many more clicks than infrequent queries. We conjecture that more common queries lead to better results, which in turn result in more activity [1].

7. PREDICTING USER ENGAGEMENT

In the previous sections, we examined toolbar logs to study user behavior, characterizing user behavior in terms of post-result click activity, time spent, and entropy. Metrics such as these could be leveraged to improve ranking of sponsored and organic results, as illustrated in a rich set of prior work in IR [2, 5, 18, 19, 23]. Furthermore, our study revealed that these metrics may be even more important given the fact that neither CTR nor PageRank provide a strong signal for how users behave after a result is clicked.

However, a major hurdle in leveraging user behavior for ranking is its limited availability. Given the vast size of the internet, the data available to any one party is likely to cover only a small fraction of the trail activity that takes place on the Web. To address this sparsity, next we look at the problem of predicting user behavior: given a user, query, and result, we want to predict the activity on the trail originating from this result, conditioned on the fact that the result is clicked [4].

Predictions about on-site user activity are made using standard statistical tools, incorporating features extracted from past user browsing behavior, the various query features from Section 6, and data about how users navigate on individual landing pages and their associated sites. We show that some measures of behavior can be predicted reliably when all three entities involved in the given triplet (i.e., user, query, result) have been previously seen in the training data. Also, we show that the classification can be made robust to the case where some of these entities are new, i.e., unseen in the training data. These results have two implications: While prediction on known entities shows that user behavior tends to be consistent (i.e., past behavior predicts future behavior), prediction on new entities demonstrates generalizability of the prediction task.

Since our emphasis is on understanding user engagement in terms of the number of clicks and amount of time spent on-site, we now demonstrate that we can predict with sufficient reliability whether or not a user will spend more than a certain amount (in our case, 60 seconds) on a result and whether that user will make one or two additional clicks on that site. We believe the ability to predict these simple metrics implies that more general forecasts can be made, a task which we leave to future work. In the remainder of this section we present the set of features used when making these predictions, and discuss issues associated with their collection. We then present our results for predicting user activity. Finally, in order to simulate a more realistic setting and to evaluate the influence of certain features, we perform predictions with intentionally excluded feature sets.

7.1 Feature Extraction

The features in our experiments are divided into four entities:

User Features: Observations compiled on a specific user's query result and navigational behavior. Strict anonymity is ensured throughout the course of our experiments.

Query Features: How clicked results are distributed, and how results tend to be browsed for particular queries[5].

Site Features: Navigational features aggregated across all landing pages associated with a particular site. Site-wide features are much less sparse, and therefore more likely to be available, in quantities needed to make valid statistical inferences, than single-page features.

Page Features: Features pertaining to individual landing pages. While this set of features is potentially more revealing than site-wide features, we are less likely to have enough information in the logs for a particular landing page. The set of specific features used in this set is identical to that for the Site Feature set, differing only in the granularity of aggregation.

The specific nature of each feature is described below. Some of these features are accumulated for all four of the above entities, while others are specific to one or two. This list of features is by no means exhaustive, and additional features may lead to further gains in prediction.

Click Probabilities: The probability of making $i \in [1, 5]$ clicks, and the mean number of clicks made per visit. Used in all four entities.

Distribution of Times: The amount of time spent on the page resulting on the i^{th} click, $i \in [1, 5]$. Additionally, the mean time spent per visit. Used in all four entities.

Navigational Shannon Entropy: The number of bits needed to encode the path chosen on a particular site or page. In addition to computing path entropy values for all paths visited on a site, we compute entropy for limited-depth user navigations, giving some notion of how a site is shaped from the perspective of a site's visitors. These features are used in the Page and Site Feature entities.

Query Intent: Rather than perform more advanced computation in order to determine the desired type of action a user wishes to perform, we compute the variance among clicked results for a particular query, and the information entropy for describing query results. The intuition behind this approach, as described before, is that navigational queries will tend to focus on a single result, while non-navigational queries will have more diffuse results, and therefore greater variance and entropy. This is a Query feature, of course.

User Activity: The number of queries issued by a user, the average number of clicks per query by this user, the probability a clicked result will be an ad, and the expected position of a user's clicked results. These features are applicable only to the User entity.

Activity on Queries: The frequency, click rate, ad click rate, and mean position of the clicked result. Also the diversity of results clicked, that is $\frac{|urls_{unique}|}{|freq_{query}|}$.

Query Topic: The topic of a particular query, as determined by our proprietary query taxonomy. Additionally, we look at the number of terms present in a particular query.

Click-Through Rate A simple estimation of click-through rate for a particular url: The number of times a url is clicked divided by the number of times a url is returned as a query result. This is a page-specific feature.

7.2 Experimental Setup

Our data consists of more than 2 million instances, where each instance consists of a triple < user, query, result URL > and the resulting click trail on the site. Since many features listed above require aggregations over many instances (e.g., click probabilities), we need to be careful during the feature extraction process to ensure that the information from the test set is not "leaked" to the classifier. More specifically, we perform the aggregation and agglomeration of features as follows: our instances are partitioned into two equal sized sets. One of these two sets is used to compile features which require

[4] Note that the probability of a result to get clicked is its CTR value. CTR prediction has been studied before, and we focus on the orthogonal problem of predicting activity after the result is clicked.

[5] Future work could follow previous work in [5] and use individual query terms (rather than complete queries) to alleviate sparseness in the data.

Feature Set Used	Click AUC	Time AUC
Top $100k$ Ad	0.708	0.594
Random $100k$ Ad	0.672	0.585
Top $100k$ Non-Ad	0.704	0.598
Random $100k$ Non-Ad	0.613	0.560

Table 1: Predictive Performance on the Various Data Sets Used for Ad and Non-Ad Data.

cross-instance aggregation. We call this set the *training set* and the other set the *test set*.

From the training set, a classifier is trained. While evaluating on the test set, given a test instance we probe into our training set to check whichever features are available for the entities involved in the test instance. For instance, if the user involved in the test instance is entirely new, then we may not get any user-centric features for him/her from our training set. The same is not true for queries though, since features like query topic or query intent can be extracted even if the query is new. The classifer then takes these features to predict user behavior on the test instance. This evaluation process ensures that no information from the test set is disclosed to the classifier.

While it is impossible to make meaningful predictions in a complete absence of features, we show that even a few features are sufficient to provide reasonable classifier performance. In order to evaluate the predictive response to missing features, two data sets are compiled for both organic and sponsored search results. (Both data sets consist of training and test data, in order to ensure fair evaluations, as described above).

- *RandomSet*: This dataset consists of $100,000$ instances sampled at random, restricted to ensure that at least *some* features are present. This random sampling leads to a feature distribution which should reflect performance of our classifier in the wild.
- *TopSet*: This is a filtered dataset to ensure that feature density is high and that all sets of features are represented. For each of the four entities (user, query, site and page) listed before, a subset of the data is extracted, consisting only of instances where the associated feature is present; instances where an entity's associated feature is missing are filtered out. For each of these four sets, the $25,000$ instances containing the most number of features are selected; these are best-case samples in which the presence of certain features is ensured. Finally, these four sub-sets are combined to make $100,000$ optimistic instances. This data is used to give something of a best-case evaluation of our classifier. Since the presence of all features is ensured, we can use this data set to estimate the discriminative ability of each feature entity.

All classification tasks in this section were performed using cost-sensitive two-class logistic regression with ridge regularization. The vector, w^T is optimized using Newton's Method. Parameters and weights used for each classification sub-problem were hand tuned by compiling a large parameter set, performing model training and test evaluation on each parameter. The configuration offering the best test performance is retained. To account for imbalances in the size of positive to negative instances in each of our experiments, we use the area under the receiver operating characteristic curve (AUC) as our metric for evaluating predictive ability.

7.3 Results

Initial Results: Classifier performance on each of these data sets is presented in Table 1. The first experiment we evaluate is the binary classification task deciding whether or not a user will make one additional click (column titled "Click AUC"), or if a user will spend at least one minute ("Time AUC" column), on a particular result site given that the result was clicked. Both data sets are used to show how well we can do with the Top $100k$ data set as well as a more natural setting of the Random $100k$ data. We conduct these experiments on both organic and sponsored search data.

From Table 1, we see that in both the ad and non-ad cases, the Top $100k$ data set offers an improvement in classification performance over the Random $100k$ data set. This result is unsurprising since the former data set consists of instances where the most information is available, while the latter follows the natural, often sparse distribution of feature availability. We note that while Random $100k$ offers degraded classification performance, we believe the results are still acceptable. With more optimized feature extraction user behavior can be conjectured with sufficient reliability in the wild.

Of interest is the improved predictive ability on the ad data set in comparison to the non-ad data. This is surprising, since our data set contained more non-ad data, by an order of magnitude, implying denser feature availability. We believe this can be attributed to different user expectations when clicking on ads as opposed to non-ads: There are many reasons why a user may click on an organic result presented by a search engine, but only a small subset of these tend to lead to clicks on sponsored search results.

Predicting Two Clicks: While predicting a single click on a query result is a challenging task, we seek to discover whether further activity can be forecasted. As a simple test, we take query results from the ad and non-ad Top $100k$ data sets, and predict if a user will go to make two clicks on a particular result. Table 2 presents the results of this experiment. We see that performance is comparable to that achieved in predicting a single click. This may indicate that most of the uncertainty involved in predicting user behavior occurs at the first step. For future work, we would like to perform a much more detailed prediction of user activity.

Data Set Used	AUC
Ad Top $100k$	0.69
Non-Ad Top $100k$	0.697

Table 2: Accuracy in Predicting Two Clicks

Ablative Feature Experiments: Missing some features is common in our data. In order to understand the reduction in classifier performance from missing a particular feature set, we evaluate the performance of our classifier in the presence of all *but* a given entity of features. This ablative feature classification is performed by removing certain features from the training and test data, and using logistic regression as before. Top $100k$ data sets are used to ensure that the feature removed was present in sufficient numbers in the initial data set, and while only results from our sponsored search data set are used due to space constraints, the results are similar for non-ad data. The results of this study are presented in Table 3.

Feature Set Removed	Click AUC	Time AUC
None	0.708	0.594
User Features	0.708	0.594
Query Features	0.708	0.59
Site Features	0.671	0.572
URL Features	0.667	0.586
User & Query	0.708	0.59
Site & URL	0.594	0.546

Table 3: Accuracy Results In an Ablative Feature Comparison for Ads on the Top $100k$ Set

Single Feature Experiments: Table 3 reveals the dependency of our classifier on site and URL specific information; missing either of these feature sets severely restricts the ability of our classifier, while user and query can be removed with little consequence on the output. At this point, it is unclear if users or queries are really useless in determining on-site behavior, or if this information is largely subsumed once site and URL information is known. We conduct a single feature classification experiment on our ad Top $100k$ data in order to understand exactly how much information related to each feature set helps in prediction. This is done by filtering all *but* a particular feature entity, then classifying as above; the results are shown in Table 4:

Feature Set Used	Click AUC	Time AUC
User Features	0.503	0.505
Query Features	0.594	0.546
Site Features	0.659	0.573
URL Features	0.644	0.564
All Features	0.708	0.594

Table 4: Accuracy Results Using Individual Feature Sets for Ads on the Top $100k$ Set

From Table 4, we see that user features provide very little information as to user navigation behavior in our model. This has the upside that data can be collected in a way that maximizes privacy. Queries prove to be a better discriminator then users by a wide margin; however, the performance is still significantly below that achieved when all features are present. This could be because result quality often varies significantly for a particular query, or that query synonymy or ambiguity confounds predictions based on query features alone. Site and URL features seem to offer the most classification information – some sites or pages are predictably more prone to elicit clicks or browsing time from users. It is interesting to note that site features tend to outperform URL features, even though a URL provides information on a finer resolution then a site alone as one site can contain many URLs. One possible explanation is that there exists much more site information in our data set, enabling better feature estimates. Regardless of the cause, this observation is promising since site level information is the most frequently available data in our logs.

8. CONCLUSION

In this paper, we have performed the first detailed study of post-click through user behavior on sponsored results, and compared it to the case of organic search. We also presented a generative model based on a mixture of power laws, and showed how to predict user behavior on result sites using a set of user, query, site, and page features.

Overall, our results show that user behavior in sponsored search has a number of similarities, but also some differences from that in organic search. This observation is interesting since sponsored search results originate from a distinct and specialized mechanism, ruled more by bid prices than relevance to users, who only indirectly affected ad selection via click-through rate. However, it is becoming increasingly clear that to be successful, a sponsored search platform has to balance the interests of advertisers, searchers, and search engines, and this requires use of a richer set of features including those gleaned through implicit feedback.

Our work here takes a first step towards using such features in sponsored search, but leaves open a number of questions. First, prediction results could be improved greatly by using more sophisticated methods and by incorporating additional features, some of which such as the position of the clicked ad among sponsored results were not available at the time of this study. It would also be interesting to relate post click-through behavior to actual convergence (e.g., a purchase, or as defined by the advertiser), and to explore the long-term changes in user behavior due to engagement with ad sites (e.g., do people return to a site after engaging in more clicks on a previous visit).

9. REFERENCES

[1] E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *Proc. of SIGCHI 2008*.

[2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proc. of SIGIR 2006*.

[3] E. Agichtein and Z. Zheng. Identifying "best bet" web search results by mining past user behavior. In *Proc. of KDD 2006*.

[4] R. Baeza-Yates and C. Castillo. Crawling the infinite web: five levels are enough. In *3rd Workshop on Algorithms and Models for the Web-Graph*, 2004.

[5] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In *Proc. of WWW 2008*.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual search engine. In *Proc. of WWW 1998*.

[7] A. Broder. A taxonomy of web search. *SIGIR Forum, 36, 2002*.

[8] A. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using web relevance feedback. In *Proc. of CIKM 2008*.

[9] A. Z. Broder. Computational advertising and recommender systems. In *RecSys '08: Proc. of the 2008 ACM Conf. on Recommender systems*, 2008.

[10] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *Proc. of SIGCHI 2001*.

[11] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *ArXiv Technical Report*, 2007.

[12] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal statistical Society B*, 39:1–38, 1977.

[13] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, 2005.

[14] B. H. Hager, M. A. Richards, P. T. R, B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose. Strong regularities in world wide web surfing. *Science*, 280:95–97, 1998.

[15] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. of SIGIR 2005*.

[16] T. Joachims, 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. *ACM Trans. Inf. Syst.*, 25(2):7, 2007.

[17] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. of WWW 2005*.

[18] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. Browserank: letting web users vote for page importance. In *Proc of SIGIR 2008*.

[19] M. R. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Ranking web sites with real user traffic. In *Proc. of WSDM 2008*.

[20] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search: a query substitution approach. In *Proc. of SIGIR 2008*.

[21] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proc. of SIGKDD 2005*.

[22] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proc. of CIKM 2008*.

[23] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proc. of SIGCHI 2004*.

[24] R. W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proc. of SIGIR 2007*.

[25] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc of WWW 2007*.