

Characterizing and Predicting Search Engine Switching Behavior

Ryen W. White
Microsoft Research
Redmond, WA 98052
ryenw@microsoft.com

Susan T. Dumais
Microsoft Research
Redmond, WA 98052
sdumais@microsoft.com

ABSTRACT

Search engine switching describes the voluntarily transition from one Web search engine to another. In this paper we present a study of search engine switching behavior that combines large-scale log-based analysis and survey data. We characterize aspects of switching behavior, and develop and evaluate predictive models of switching behavior using features of the active query, the current session, and user search history. Our findings provide insight into the decision-making processes of search engine users and demonstrate the relationship between switching and factors such as dissatisfaction with the quality of the results, the desire for broader topic coverage or verification of encountered information, and user preferences. The findings also reveal sufficient consistency in users' search behavior prior to engine switching to afford accurate prediction of switching events. Predictive models may be useful for search engines who may want to modify the search experience if they can accurately anticipate a switch.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: *search process*.

General Terms

Measurement, Experimentation, Human Factors.

Keywords

Search engine switching.

1. INTRODUCTION

Search engines such as Google, Yahoo!, and Live Search facilitate access to the vast quantities of information present on the World Wide Web. A user's decision to select one search engine over another can be based on factors including reputation, familiarity, effectiveness, and interface usability [19]. Searchers may not use the same engine for all queries; they often switch between different engines within and between sessions [14,18,21]. Previous work on switching has promoted multiple search engine use [22], predicted when users are going to switch [11,15], studied switching to develop metrics for competitive analysis of engines in terms of estimated user preference and user engagement [14], or built conceptual and economic models of search engine choice [18,21]. However, despite the economic significance of engine switching to search providers, and its prevalence among engine users, little is known about the rationale behind switching, the behavior itself, or the features most useful in predicting switching events.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'09, November 2–6, 2009, Hong Kong, China.

Copyright 2009 ACM 978-1-60558-512-3/09/11...\$10.00.

In this paper, we present research on the characterization and prediction of search engine switching behavior. We focus on switches within a session rather than between-session switches (that may be task-oriented) or long-term switches (that may represent significant shifts in user preferences or settings). Within-session switching is most common and allows us to study the antecedents of switching in more detail. We use two complementary methods – large-scale log analysis and user survey data – to provide a rich picture of switching behavior. Log data enables us to examine patterns of behavior for large numbers of individuals, and the survey data enables us to understand some of the rationale behind the observed patterns. The reasons behind the switches, such as user frustration, a desire for topic coverage or fact verification, prior experience, and interface usability, are challenging to reliably study in logs but can be identified in survey responses.

In addition to characterizing switching behavior we also investigate the effect of different features on the accuracy of switch prediction models. We build models with rich sets of features derived from the active query, recent interaction behavior from within the current search session, and/or the user's long-term search history. Earlier work on switch prediction applied data mining techniques to user actions encoded as character sequences [11,15]. However, such sequences are only one way to represent interaction behavior and may not always be available to search engines. It is therefore important to understand what other features can yield accurate switch predictions. We extend previous switch prediction research using a broad set of features derived from our log and survey analysis. Through our methodology we characterize properties of queries, sessions, and user histories that are potentially useful in prediction. A better understanding of which features contribute most to improving prediction accuracy can yield powerful models that do not depend on complex representations of user interaction history, making them more attractive for large-scale deployment.

The remainder of this paper is structured as follows. Section 2 outlines previous work on predicting query difficulty and characterizing search engine switching behavior. Section 3 provides an overview of the log-based analysis and survey methodologies. In Section 4 we characterize switching behavior, including aspects of the pre- and post-switch interaction. In Section 5 we investigate the predictive value of query, session, and user features in isolation and combination. We discuss our findings and their implications in Section 6 and conclude in Section 7.

2. RELATED WORK

Two lines of work are relevant to our research: predicting query difficulty and characterizing search engine switching behavior.

There is an established record of research in information retrieval that addresses the challenge of predicting query performance, and the influence of different query representations or document re-

presentations on such performance. A high-level goal of that work is to understand differences in performance across queries to devote additional resources or use alternative methods, as appropriate, to improve the overall search experience. For example, if a system knows which queries are difficult, it could devote additional resources to enhancing search results for those queries, or if a system knows which algorithms work best for a particular query, it could improve performance by selecting the most appropriate algorithm for each query. While it is easy to show that using different query representations [2] or retrieval models [1] can improve search performance, it is more challenging to accurately predict in advance which methods are most appropriate.

Measures such as query clarity [6], Jensen-Shannon divergence [4], and weighted information gain [23] have been developed to predict performance on a query (as measured by average precision, for example). Leskovec et al. [16] used graphical properties of the link structure of the result set to predict the quality of the result set and the likelihood of query reformulation. Teevan et al. [20] developed methods to predict which queries could most benefit from personalization. In research more closely related to search engine switching, White et al. [22] developed methods for predicting which search engine would produce the best results for a query. For each query they represented features of the query, the title, snippets and URLs of top-ranked documents, and the results set, for results from multiple search engines, and learned a model that predicted which engine produced the best results for each query. The model was learned using a large number of queries for which explicit relevance judgments were available. One way in which such results could be leveraged is to promote the use of multiple search engines on a query-by-query basis, using the predictions of the quality of results from multiple engines.

A user's decision to use one search engine over another is dependent on many factors including reputation, familiarity, retrieval effectiveness, and interface usability [19]. Similar factors can influence a user's decision to switch from one search engine to another, either for a particular query, a particular task if another engine specializes in such tasks, or more permanently, as a result of unsatisfactory experiences or relevance changes, for example.

Some research has examined engine switching behavior. Some of the earliest research in this area was by Mukhopadhyay et al. [18] and Telang et al. [21]. They used economic models of choice to understand whether people developed brand loyalty to a particular search engine, and how search engine performance (as measured by within-session switching) affected user choice. They found that dissatisfaction with search engine results had both short-term and long-term effects on search engine choice. The data set is small by modern log analysis standards (6,321 search engine switches from 102 users), somewhat dated (data from June 1998 – July 1999 including six search engines but not Google), and only summary-level regression results are reported. Juan and Cheng [14] described some more recent research in which they summarize user share, user engagement and user preferences using click data from an Internet service provider. They identify three user classes (loyalists to each of the two search engines studied and switchers), and look at the consistency of engine usage patterns over time. Neither of these studies addressed the challenge of predicting switch behavior. Accurately predicting if a user is about to switch allows the search provider to offer additional search support.

Heath and White [11] and Laxman et al. [15] developed models for predicting switching behavior within search sessions using sequences of user actions (e.g., query, result click, non-result

click, switch) and characteristics of the pages visited (type of page and dwell time) as the input features. Heath and White [11] used a simple threshold-based approach to predict a switch action if the ratio of positive to negative examples exceeded a threshold. Using this approach they achieved high precision for low recall levels, but precision dropped off quickly at higher levels of recall. Working with the same data, Laxman et al. [15] developed a generative model based on mixtures of episode-generating Hidden Markov Models and achieved much higher predicative accuracy. The research reported in this paper is similar to this line of work, but extends it in several ways. We use a richer set of features to characterize properties of the query, the search session, and the user. We compliment a large-scale log study with a survey to develop insights about people's motivations for switching and characteristic behaviors, which we use to develop more abstract features such as "several related queries in quick succession without clicks". We also observe user behavior over a longer period of time (six months), and study both pre- and post-switch behaviors.

We now describe the log analysis and user survey used as the basis for our characterization of switching behavior.

3. LOG-BASED ANALYSIS AND SURVEY

We collected data from two complimentary methods – large-scale log analyses and a user survey or questionnaire. The log analyses provide insight into a range of user activities in situ. The survey provides insight into the reasons for the observed behaviors. [10, 13] have more on combining logs and other data capture methods.

We analyzed six months of interaction logs from September 2008 through February 2009 inclusive, obtained from hundreds of thousands of consenting users through a widely-distributed browser toolbar. These log entries include a unique identifier for the user, a timestamp for each page view, a unique browser window identifier (to resolve ambiguities in determining which browser a page was viewed), and the URL of the Web page visited. Intranet and secure (https) URL visits were excluded at the source. In order to remove variability caused by geographic and linguistic variation in search behavior, we only include entries generated in the English speaking United States locale. Any personally identifiable information was removed from the logs prior to analysis. From these logs we extracted *search sessions*. Every session began with a query issued to Google, Yahoo!, or Live Search and could contain further queries or Web page visits. A session ended if the user was idle for more than 30 minutes. Similar criteria have been used in previous work to demarcate search sessions, e.g., [7].

We compliment our log analysis with a survey of users' experiences with search engine switching. We distributed the survey via email to 2,500 randomly-selected employees within Microsoft Corporation. 488 employees completed the survey, for a response rate of 19.5%. The survey contained a mixture of open and closed questions. We were particularly interested in eliciting responses concerning the rationale behind engine switching since this is something that the log data does not provide. We also asked questions regarding the frequency with which people switched engines, characteristics of their most recent switching episode, and patterns of activity that preceded switching events. Five-point scales were used where appropriate, with: *Never, Rarely, Sometimes, Often, Always* used to elicit frequency information.

4. CHARACTERIZING SWITCHING

We now analyze our logs and survey data with the objective of characterizing aspects of switching behavior. We first present an

overview of the log data and the survey data. We then focus on aspects of the search behavior prior to the switch, including common actions, temporal dynamics, and significant user action sequences. In addition, we study post-switch behavior, including post-switch activity and estimates of post-switch user satisfaction.

4.1 Overview of log data

From the logs described in the previous section we extracted 1.1 billion search sessions beginning with a query to Google, Yahoo! or Live Search in the six month duration of the study. A search engine switch occurs if consecutive queries within a session are issued to different engines (e.g., query Google then query Live). Of the 1.1 billion search sessions, 42.9 million (4.0%) contained at least one search engine switch between two of the three engines, and 10.8 million (25.1%) of those switching sessions had multiple switches. In total, we observed 58.6 million instances of search engine switching behavior comprising 1.4% of all Google, Yahoo!, and Live queries in the six-month period. Of all switches, 7.4 million (12.6%) exhibited the same query on the pre-switch and post-switch engines.

As noted above, search engine switches were observed in 4% of all search sessions. However, switches are more likely to occur for longer search sessions. Figure 1 shows the probability of switching, $P(Switch)$, for sessions of varying length, as measured by the number of queries in the session.

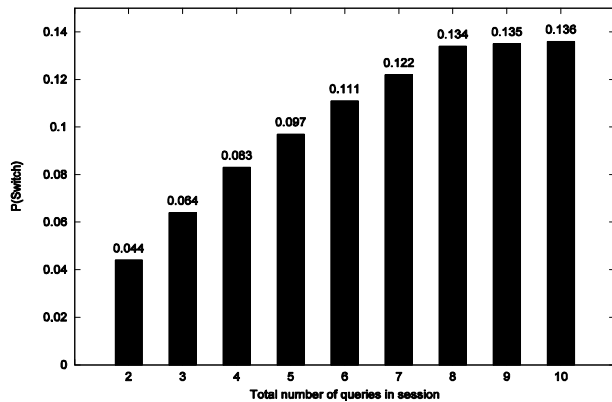


Figure 1. Probability of switching given session length.

As search session length increases, perhaps because of the nature of the user’s task or the quality of search results, the likelihood of switching also increases. For sessions that include five or more queries, switches occur approximately 10-14% of the time.

Of the 14.2 million users in our log sample, 10.3 million (72.6%) used more than one engine in the six-month duration of the logs, 7.1 million (50.0%) switched engines within a search session at least once, and 9.6 million (67.6%) used different engines for different sessions (i.e., engaged in between-session switching). In addition, 0.6 million users (4.4%) “defected”¹ from one search engine to another and never returned to the previous engine.

Although search engine switching describes the activity of voluntarily shifting from one search engine to another, the switch itself can happen in at least three ways:

- 1) *Browser*: Issue query directly from a browser search box or browser toolbar by first selecting search provider if needed.
- 2) *Navigate*: First visit search engine homepage via the browser address bar and then issue query.
- 3) *QueryToNavigate*: First query for a search engine name (e.g., search Yahoo! for [google], [google.com], etc.), visit the search engine’s homepage, and issue query.

A switching *event* is defined as any of these three switch types.

In the 58.6 million examples of switching behavior these switch types were distributed as follows: *Browser* is 69.2%, *Navigate* is 18.3%, and *QueryToNavigate* is 12.5%. It appears that browser search boxes and optional browser toolbars facilitate search engine switching behavior. *Navigate* and *QueryToNavigate* both rely on the explicit recall of the destination engine name or URL by the user before the switch can occur. This presents a possible barrier to switching in this way. In contrast, *Browser* requires only user recognition of an engine in a list of search providers or switching cues provided to users when searching on other engines.

4.2 Overview of survey data

70.5% of survey respondents reported that they had switched between different search engines either within or between sessions. This percentage is remarkably similar to the percentage (72.6%) observed in the log-based analysis reported in the previous section. This increased our confidence about the consistency of the two data sources used for this study. The respondents who did not switch did so because they were satisfied with the engine they used (57.8%), they believed that no other engine would perform better (24.0%), or felt that it was too much effort to switch engines (6.8%). Other reasons provided included loyalty derived from features such as long-term histories or privacy protection, consistency, and distrust or dislike of other search engine brands.

66.8% of those who reported that they switched engines did so within a session at least *Sometimes* and 24.4% of subjects switched within a session *Often* or *Always*. As part of our survey we asked those respondents who switched with a session to provide the rationale for their switching behavior. They did so by selecting at least one explanation from a list of possible reasons provided to them. In Figure 2 we present the breakdown of responses, grouped by the response options offered to respondents.

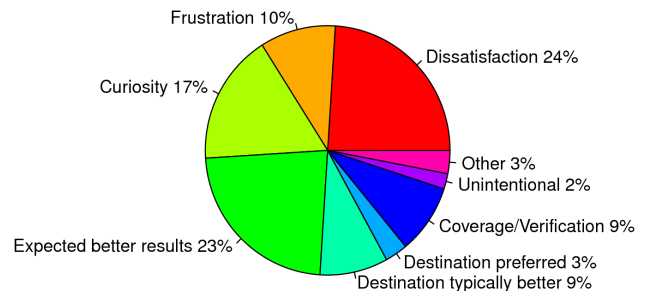


Figure 2. Reasons given for search engine switching.

There are three general types of reasons: dissatisfaction with the quality of results in the original engine (dissatisfaction, frustration, expected better results), the desire to verify or find additional information (coverage/verification, curiosity), and user preferences (destination preferred, destination typically better). These same three motivations were also seen in free-form survey feedback. Respondents who answered *Other* listed reasons such as

¹ Our definition of defection was a switch from one engine to another, issuing at least one additional query on the post-switch engine, and never returning to the origin engine. More relaxed variants of these criteria would likely yield more defections.

loyalty, hope, and search applications that let them view the results from more than one search engine simultaneously.

Although we focus on within-session switching in this paper, we also asked survey respondents to describe and rationalize any between-session switching behavior (*i.e.*, attempt one session on one engine and another session on a different engine) or long-term switching (or defection). 46.5% of those who switched did so between search sessions at least *Sometimes* and 14.2% of switching respondents did so between sessions *Often* or *Always*. The reasons that respondents gave for between session switching were that the destination engine typically performs better for the task they were attempting (55.2%), any engine would have sufficed (18.6%), or unintentional (*e.g.*, different entry point or different computer) (12.8%). Other reasons included trust and differences in engine performance for different markets.

40.4% of subjects reported having defected from one search engine to another and never or very rarely returning to the pre-switch (origin) engine. 82.7% of subjects reported that they were happy with their decision to defect. This is substantially higher than the 4.4% observed in our log analysis, and likely reflects the fact that we used only three popular engines in that analysis (but our survey respondents may try new engines for short periods of time), and our strict definition of defection. The main reasons for defection were many dissatisfactory experiences with the origin engine (43.9%), one particularly dissatisfactory experience with the origin engine (7.9%), more relevant results on other engine (20.1%), or a new entry point such as a browser search box or optional browser toolbar (28.1%). Since the effect of dissatisfaction appears cumulative, search providers should promptly address all forms of dissatisfaction in order to retain their users.

We now describe aspects of pre-switch behavior.

4.3 Pre-switch behavior

A better understanding of the antecedents of switching can help explain switching behavior and facilitate the accurate prediction of switching events. We used all 58.6 million switching events in our logs and analyzed important pre-switch interactions.

4.3.1 Actions preceding a switch

We began our analysis of pre-switch behavior by calculating the frequency of actions immediately preceding a switching event, defined earlier as one of *Browser*, *Navigate*, or *QueryToNavigate*. There are five actions that we consider: Query, Pagination (*i.e.*, requesting the next page of search results for the current query), Clicking on a search engine result page (SERP), Clicking on another (non-SERP) page, and Navigation to another page not associated with a click (*e.g.*, through browser address bar). We also identify cases in which the switch occurs immediately at the start of the session and the preceding event is Start session. Figure 3 shows the breakdown of actions immediately before a switch. The most common pre-switch actions are queries, followed by non-SERP clicks, SERP clicks, and navigation to other pages.

We also studied in the extent to which this distribution of activities held across the search process. Figure 4 (overleaf) shows the temporal dynamics across all 58.6 million switches in more detail. In particular, we show the probability of an action, $P(\text{Action})$, occurring at different time points leading up to a switch. We consider the five actions described above: Query, Pagination to the next SERP, Click SERP result, Click non-SERP link, or Navigate to page. The top panel of the figure shows the proportion of each action as a function of time in the session before the switch. The

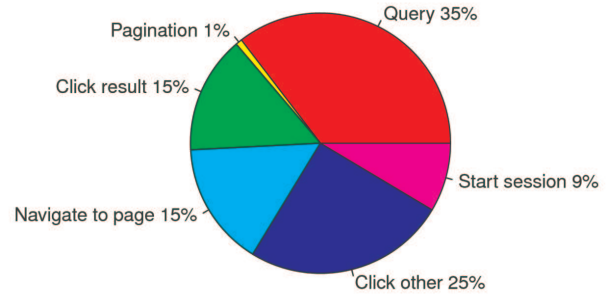


Figure 3. Observed actions immediately preceding a switch.

time scale is normalized to show proportions of the total *pre-switch* session time. Visible oscillations in $P(\text{Action})$ can be attributed to bucketing noise during normalization. Actions that occur shortly after the first query in the session are shown at the left, and those that occur just before the switch are shown at the right. A query occurs 100% of the time at the beginning of a session by definition. The proportion of total actions that the query represents decreases as other actions become important. SERP clicks are common early in the process, accounting for 50% of the actions immediately following the query, but fall off after that. This is similar to a result reported by Downey et al. [7] in which SERP clicks were more frequent than another query for the 25 seconds after a query, but another query was more common subsequently.

It is also interesting to consider which actions increase just before a switch. Looking at the far right of Figure 4 we see that clicks (on either the SERP or non-SERP) decrease, and that pagination, queries and navigation actions increase. The reason for the small drop in navigation behaviors is unclear, but may reflect users abandoning alternative resources they have navigated to in favor of trying another engine. Immediately before a switch, users are less likely to click URLs relative to other points during the session and more likely to try another query or to page to see more results. We have also investigated the types of URLs that people click on. The bottom panel of the figure shows the proportion of clicks that are to pages that the user has previously viewed in the session, represented as the probability of revisitation, $P(\text{Revisit})$. The proportion of revisits increases as the session progresses as users return to previous SERPs or other pages. $P(\text{Revisit})$ rises sharply immediately before a switch, perhaps confirming the frustration or dissatisfaction suggested in our survey responses.

4.3.2 Multi-action pre-switch sequences

To obtain further insight into what users do before a switch that may be useful for both characterizing and predicting switching we asked survey respondents the following question: “Is there anything about your search behavior immediately preceding a switch that may indicate to an observer that you are about to switch engines?” We analyzed subject responses to this question and identified the following five most common answers:

- A1: Try several small changes to the query (word order, phrases, synonyms, more specific), often in pretty quick succession.
- A2: Go to more than the first page of results, again often in quick succession and often without clicks.
- A3: Go back and forth from SERP to individual results, without spending much time on any.
- A4: Click on lots of links, then go to another engine for additional information.
- A5: Do not immediately click on something.

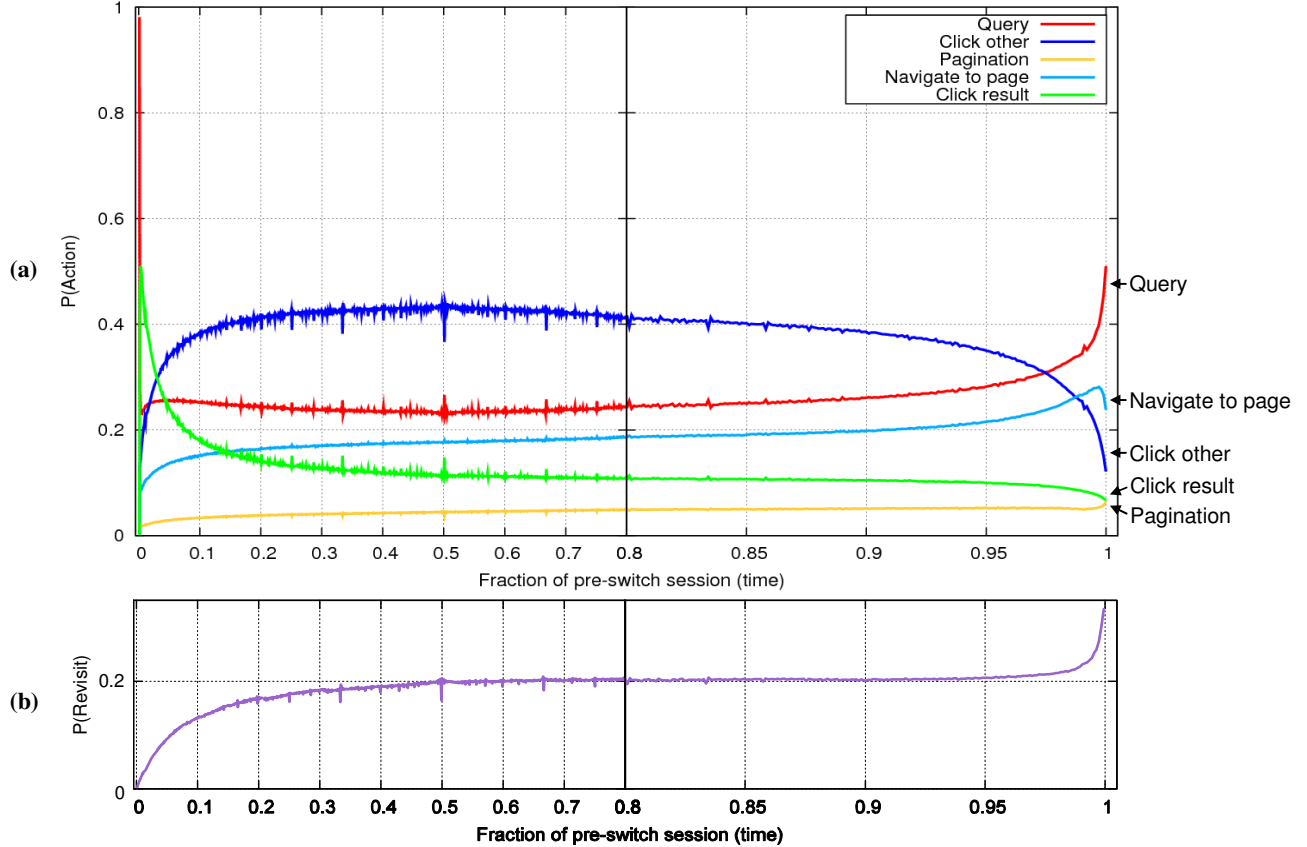


Figure 4. Temporal dynamics of pre-switch search activity: (a) probability of an action and (b) probability of revisitation.

To verify whether these five behaviors also appeared in our logs and to use them as features in a predictive model we needed to first encode them in some way. Earlier work (e.g., [7,9,11]) has already introduced formal models and languages that encode search behavior as character sequences, with a view to comparing search behavior in different scenarios. We formulated our own alphabet with the goal of maximum simplicity. We encode the pre-switch interaction behaviors as a sequence of characters, where each character corresponds to either: (i) a user action such as a query or click, or (ii) attribute(s) of the page visited such as SERP or non-SERP. We encoded page visits in two ways: *basic* and *advanced*. In the basic representation we only differentiate between SERP and non-SERP pages. However, in a similar way to [7], we felt that page dwell times could be useful and we encoded these also. Dwell times were bucketed into “short”, “medium”, and “long” based on a tripartite division of the dwell times across all users and all pages viewed. The *advanced* representation uses this more detailed characterization of page visits that includes information about dwell time. Table 1 shows the alphabet used in our study. We automatically encode all actions by stepping through the action series in chronological order, and at each point categorizing the page and the action taken to get there. For example, a user issuing a series of queries, each time viewing the resultant SERP for a short duration but not clicking on any search results and then navigating to a non-SERP page through the browser address bar, and viewing that page for a long time, would be represented in basic form as $qRqRqRqRnP$ (or in abbreviated form $qR*nP$), or in advanced form as $qAqAqAqAnH$ (or in abbreviated form $qA*nH$).

Table 1. Characters assigned to actions and pages visited.

Action	Page (basic)	Page (advanced)
q Query	R SERP	A SERP (short)
p Pagination	P Non-SERP	D SERP (medium)
s Click result		E SERP (long)
c Click other		F Non-SERP (short)
b Back one page		G Non-SERP (medium)
j Back many pages		H Non-SERP (long)
n Navigate to page		

We encoded all pre-switch interaction activity in the 58.6 million switching events (including all substrings) in this format and computed the frequency with which they appeared before a search engine switch. We also encoded all 1.1 billion search sessions in this format (creating tens of billions of action sequences) and calculated how frequently each of the pre-switch strings was observed in all sessions independent of switching. From these frequency counts we identified significant pre-switch patterns called *sequence motifs* by calculating the *point-wise mutual information* (PMI) for each sequence. PMI is a measure of association based on information theory that compares the probability of observing two items together with the probabilities of observing two items independently (c.f. [5]). We apply it in our context to estimate which sequences had a genuine association with pre-switch behavior and which were observed by chance. Table 2 presents the five basic and advanced sequence motifs with the highest PMI values.

To generate these motifs we required that each appear at least 10,000 times in all search sessions over the six months. This threshold allowed us to filter infrequent sequences that also co-occurred with switching events, giving them a high PMI value.

Table 2. Top significant pre-switch sequence motifs.

PMI rank	Basic representation	Advanced representation
1	$qR*sPbR$	$qA*sFbD$
2	$qR*nPcP*$	$qA*qDsF$
3	$qRsP*qR*$	$qAsF*$
4	$qRsP*qRjP*$	$qDpF*$
5	$qRpR*$	$qA[sFbA]*$

The sequence motifs reveal some interesting behavioral patterns. For example, it appears that repeat submission of queries followed by no SERP clicks (*i.e.*, $qR*$) commonly precede engine switches, and that revisitation and pagination also seem important. Such features were also mentioned in common survey responses. Three of the responses (*A1*, *A2*, and *A3*) suggest that pre-switching users view SERPs for short time durations, often without clicking on search results. Repeat queries with no SERP clicks is mentioned in common survey response *A1*, and pagination and revisitation are mentioned in responses *A2* and *A3* respectively. Note that the suffix of the fifth-ranked advanced sequence motif in Table 2 (*i.e.*, $[sFbA]*$) means that the same action – click search result, view page for short time, and return to SERP for short time – appeared repeatedly in sequence, and often before a switch. Such behavior was also highlighted in *A3*. We do not see any strong evidence for *A4*, clicking on lots of links before going to another engine for verification, and this may indicate that this behavior is less common than switching because of dissatisfaction or that the behavioral antecedents are difficult to encode. We also did not see any strong evidence for *A5*, perhaps because SERP views with medium-long dwell times and no clicks occurred frequently in many sequences, independent of engine switching.

We have investigated aspects of the pre-switch behavior of search engine users. In Section 5 we will evaluate the effectiveness of features derived from this analysis for predicting engine switching. We now focus on the behavior following a switching event.

4.4 Post-switch behavior

Once again we use the 58.6 million switch examples from toolbar logs and study user behavior following a switch. We also include a log-based analysis that estimates whether users were satisfied with the results they encountered following the decision to switch.

4.4.1 Actions following a switch

We begin our analysis by focusing on user actions after a switch. In Figure 5 we present a summary of the actions that immediately follow a switching event. We consider six actions in total: Click on a SERP result, Re-query the destination engine, Query on other engine (*i.e.*, switch again to a third engine), Re-query origin engine (*i.e.*, switch back to pre-switch engine), Navigate to another page without clicking on a link, or End session.

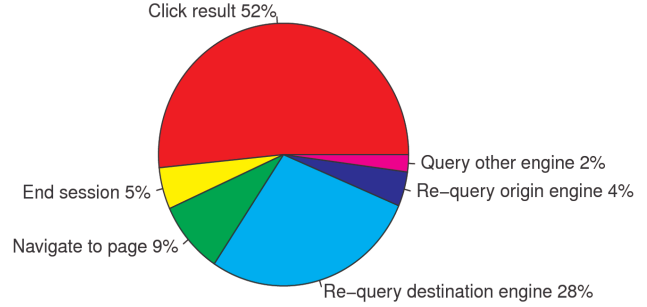


Figure 5. Actions immediately following an engine switch.

As can be seen from Figure 5, around half of switches were followed by a search engine result click. This suggests that around half of switches were successful in getting users to information that appeared relevant. For the remaining switches users engaged in a range of activities including ending the search session and navigating to another page through the browser address bar or favorites list. Around one third of all switches led to another query as the immediate follow-on action; suggesting dissatisfaction with the immediate search results. Most of those queries are on the destination engine, however around 15% of those queries involve immediately switching back to the origin engine (*e.g.*, query Live Search then Google then return to Live Search) or querying a third engine (*e.g.*, query Google then Yahoo! then Live Search).

Extending the analysis beyond actions immediately following the switch allowed us to look further at returns to the origin engine and the utilization of multiple engines. If we examine the next query, ignoring events in-between if required, we find that around 20% of all switches lead to a return to the origin engine on the next query and around 6% of all switches lead to the use of a third engine. These behaviors may be attributable to the destination engine not meeting users’ information needs or to users seeking to verify encountered information or obtain more information (as we saw in the survey responses). Further analysis of the queries for which this behavior was observed revealed that many were informational in nature (*e.g.*, computer error messages, medical diagnosis, legal advice, or term-paper questions). For such queries search engines may be ineffective or users may wish to verify encountered information or explore topics in greater detail.

4.4.2 Satisfaction

One interesting question is the extent to which switching to a new engine improves the user’s task success. It is difficult to know for sure whether an information need was satisfied using only log data, but we explore several possible measures. We report two measures of overall user effort and activity (number of queries and number of actions), and two measures that summarize the quality of the interaction. The first measure is the fraction of queries that result in no SERP clicks ($\%NoClicks$). The intuition is that no clicks are a likely indicator of poor quality results. We realize that some queries are satisfied by the search results themselves and do not require any additional actions. But others, *e.g.*, [8], have found that SERP clicks are less likely for low frequency queries and goals, so we include that measure in our analysis here. The other measure we use is based on work by Fox et al. [9] in which they showed that clicks which are followed by a dwell time of more than 30 seconds on the destination page are more likely to be rated as “satisfied” by users than those that result in a quick

return to the SERP. Thus we define a *SatAction* as the first SERP click that a user dwells on for more than 30 seconds.

In Table 3, we summarize these measures for actions that occur before a switch (origin engine) and after a switch (destination engine). We show this separately for all switches and for switches involving the same query on the origin and destination engine.

Table 3. Measures of effort / activity / quality of interaction.

Activity	# Queries		# Actions	
	Origin	Destination	Origin	Destination
All Queries	3.14	3.70	9.85	11.62
Same Queries	3.08	3.73	9.03	10.25
Success	% NoClicks		# Actions to SatAction	
	Origin	Destination	Origin	Destination
All Queries	49.7	52.7	3.81	4.71
Same Queries	54.5	59.7	3.67	4.61

The results are very similar for both types of switches. Note that given the large sample sizes all differences are significant with independent measures *t*-tests at $p < .001$. Users issue more queries and perform more actions on the destination engine than on the origin engine. They also seem less satisfied by our two measures – there are more queries with no clicks on the destination engine, and there are more actions before the first *SatAction*. Thus, on average, switches do not appear to lead to a quick resolution of the users’ information needs.

An area for future research would be to examine different classes of queries in more detail to see if we can identify consistent classes of queries for which there are advantages to switching and those for which there are no such benefits.

In this section we have focused on characterizing aspects of search engine switching behavior. As well as characterizing the behavior, it is important to understand the role that features derived from this behavior can play in a predictive model of switching. An ability to accurately predict when a user is going to switch allows the origin and destination search engines to act accordingly. The origin engine could offer users a new interface affordance (e.g., additional query suggestions, or richer support for sorting or filtering using metadata about the search results), or search paradigm (e.g., engage in an instant messaging conversation with a domain expert) to encourage them to stay. In contrast, the destination engine could pre-fetch search results in anticipation of the incoming query. In the next section we describe an investigation of the predictive value of query, session, and user features.

5. PREDICTING SWITCHING

The prediction task is to estimate whether a user’s next action will be an engine switch given the interaction observed in a session so far and possibly knowledge about the user’s long-term interaction history. For this task we developed a learning model that uses logistic regression (cf. [12]), a technique that has been shown to have good performance in many domains and can effectively handle numerical and categorical predictor variables. The aim of this experiment is not to optimize the model but rather to determine the predictive value of the query/session/user feature classes for the switch prediction challenge. The model is held constant throughout the experiment and only the features used change per the experimental design. We now describe the features we use, the evaluation of models that use them, and the experimental findings.

Table 4. Features used in switch prediction.

Query class
<i>abandonmentRate</i> : Fraction of times query has no SERP click
<i>avgClickPos</i> : Average SERP click position (starts at zero)
<i>avgNumClicks</i> : Average number of SERP clicks
<i>avgNumAds</i> : Average number of advertisements shown
<i>avgNumQuerySuggestions</i> : Average number of query suggestions
<i>avgNumResults</i> : Average number total search results
<i>avgTokenLength</i> : Average length of query tokens
<i>followOnRatio</i> : Fraction of times query leads to another query
<i>frequencyCount</i> : Total query frequency
<i>hasAlteration</i> : True if alteration applied (e.g., remove plurals)
<i>hasOperators</i> : True if query has operators (e.g., site:)
<i>hasQuotes</i> : True if query contains quotation marks
<i>hasSpellCorrection</i> : True if spell correction fires
<i>paginationRate</i> : Fraction of times request next page of results
<i>queryLength</i> : Query length in characters
<i>queryTokens</i> : Query length in tokens
Session class
<i>avgTimeBetweenQueries</i> : Average time between queries
<i>currentEngine</i> : Current search engine name
<i>currentSequenceAdvanced</i> : Advanced string rep. of session so far
<i>currentSequenceBasic</i> : Basic string representation of session so far
<i>hasMotifAdvanced</i> : True if <i>currentSequenceAdvanced</i> has seq. motif
<i>hasMotifBasic</i> : True if <i>currentSequenceBasic</i> has sequence motif
<i>numBacks</i> : Number of revisits in the session so far
<i>numPaginations</i> : Number of paginations in session so far
<i>queriesInSession</i> : Number of queries in the session so far
<i>ratioQueriesWithNoClicks</i> : Fraction of queries with no clicks
<i>ratioQueriesWithOneClick</i> : Fraction of queries with one click
<i>ratioQueriesWithMultipleClicks</i> : Fraction of queries with many clicks
<i>timeInSession</i> : Time in the session so far (in seconds)
<i>URLsInSession</i> : Number of URLs in session so far
User class
<i>avgSessionLengthQueries</i> : Average session length in queries
<i>avgSessionLengthTime</i> : Average session length in time
<i>avgSessionLengthURLs</i> : Average session length in URLs
<i>avgQueryLength</i> : Average query length in characters
<i>avgQueryTokens</i> : Average query length in tokens
<i>propPreferredEngine</i> : Fraction queries issued to preferred engine
<i>sessionCount</i> : Total number of sessions

5.1 Features

Table 4 summarizes the features that comprise the three feature classes. This list is not exhaustive, but does cover important aspects of search interaction that may have value in this context, including many that emerged from the analysis in Section 4.

5.1.1 Query Features

Query features are assigned to the most recent query in the session within which the prediction is being made. They are derived from the query itself (e.g., number of tokens) and from the search logs of one of the engines in our study. The logs were gathered over the same six-month time span as the toolbar logs used to characterize switching (i.e., September 2008 to February 2009 inclusive). Unlike toolbar logs, search logs contain records of the SERP contents shown to users at query time (e.g., the number of advertisements shown or the total number of query results).

5.1.2 Session Features

Session features are computed based on the observed interaction in the session up until the point that the switch prediction is made. Session features include information about the distance into the session (e.g., the number of queries issued or pages visited so far),

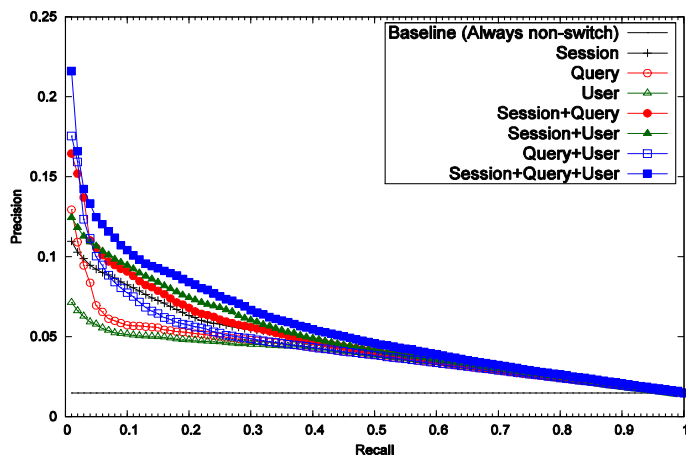


Figure 6. Precision-recall curve for all sessions.

result inspection behavior (*e.g.*, the number of revisits or paginations), search success (*e.g.*, the ratio of queries with no result clicks), and patterns of interaction (*e.g.*, basic and advanced string sequences, *currentSequenceBasic* / *currentSequenceAdvanced*). Also included were binary variables *hasMotifBasic* and *hasMotifAdvanced*. These were set to true if any of the top-100 sequence motifs emerging from our Section 4.3.2 analysis appeared in *currentSequenceBasic* or *currentSequenceAdvanced* respectively.

5.1.3 User Features

User features are computed at all points in the session based on the current user's search history gathered over the six-month period from September 2008 to February 2009. From each user's history we extracted features of their queries, their frequency of searching, their average session length (in terms of queries, time, and URLs), and the proportion of queries that they issue to their preferred engine. We would expect users who switched frequently to issue a smaller fraction of queries to their preferred engine than those who switch infrequently or never switch.

5.2 Evaluation

As stated earlier, the prediction task was to predict given features of the query, session, and user whether an engine switch was about to occur as the next user action. The goal of this experiment was to assess predictive value of each of the feature classes and highlight the individual features that performed well. We learned seven models, representing the three individual feature classes and combinations of them. We also include a baseline which always predicts the most common action, no-switch.

Switching immediately follows around 1% of all search-related interactions. This makes the switch prediction task extremely challenging. Since the task was to predict whether the next action was a switch and not whether a full session contained a switch we used session *states* rather than complete sessions in our evaluation. A session state contains the observed interaction in a session to a given point, as well as the most recent query and a unique user identifier used to locate user history if required.

We used a sample of 100,000 session states randomly chosen from the six months of logs from September 2008 to February 2009 inclusive to train a version of our learning model for each of the seven feature combinations. To mirror the distribution of real switches, the sample contained 1,000 randomly-chosen switching

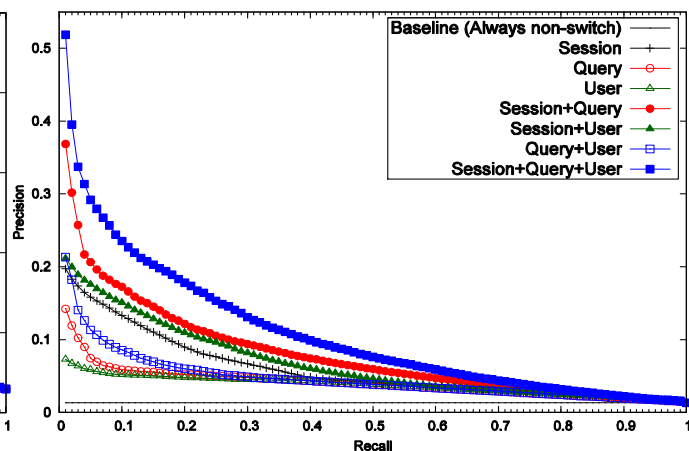


Figure 7. Precision-recall curve for sessions with three or more queries observed so far.

states and 99,000 randomly-chosen non-switching states. However, the class imbalance caused by the small number of switching events may hurt the performance of the learning model.

A common way of addressing class imbalance is to artificially rebalance the training data. To do this we down-sample the majority class (non-switches) using a technique similar to [17]. In our case, this involved holding the 1,000 positive examples constant, randomly selecting without replacement 1,000 non-switching examples, and training a logistic regression model on the 50/50 split. We repeat this process until all non-switching examples were used in training exactly one time. This yielded a total of 99 different sub-models that each make a prediction about whether a switch is about to occur. The majority vote among the predictions is then used to determine the overall prediction via a form of *bagging* [3].

To test our models we created a separate test set that succeeded the training set. We extracted approximately 300 million search sessions from toolbar logs for March 2009 and April 2009 using the method described in Section 3. From these sessions we randomly-selected a subset of 10,000 session states ensuring that the ratio of switching to non-switching in each subset was 1:99 to match the global likelihood of the switching event. To reduce sampling bias we constructed 100 subsets using this approach.

Evaluation proceeds as follows. At each of the 10,000 session states in the current subset, the model predicts whether a switch will occur as the next action given the features of the most recent query, the session so far, and/or the user search history. To do so, the model obtains a binary (switch/no-switch) prediction from each of the 99 sub-models, counts the number of switch and non-switch predictions, and makes the final prediction based on which outcome has the most votes. The performance of the model with the assigned feature classes is then determined using precision and recall averaged across all 100 subsets.

We now describe the findings of our analysis.

5.3 Findings

We evaluated the performance of each of the seven feature classes plus the no-switch baseline using precision and recall. Different levels of recall are achieved by setting different confidence thresholds for our model ranging from extremely low confidence to extremely high confidence. In this context, precision is defined as the number of true switches (*i.e.*, predicted switches that actually

were switches) divided by the total number of session states in the test set labeled with the switching event. Recall is defined as the number of true switches divided by the total number of switches in the test set. Figure 6 shows precision-recall curves for our predictions of whether a switch will occur at the next action. Separate curves are shown for models using query features, user features, and session features (for actions preceding the action we are predicting), feature combinations, and a baseline that always predicts no-switch since this is by far the most likely outcome. Error bars are too small to be visible on Figures 6 or 7 (forthcoming).

First, we consider performance using just a single class of features (query, user, or session). The best performance is obtained for the session features, followed by query features, and user features. Even user features, which perform the most poorly, still provide considerable lift over the baseline model. Users differ along many dimensions and the simple measures we have encoded (*e.g.*, the average length of queries they have issued, the proportion of queries for which they have previously switched search engines), provide some improvements prediction accuracy. Knowing characteristics of the query, such as its length and previous click patterns, can improve predictive accuracy even more. And, knowing characteristics of the session to date, such as the time in the session or previous clicks, are the most useful for improving accuracy. Second, we examined combinations of these features. Combining the different types of features results in marked improvements in accuracy, suggesting that they provide complimentary evidence about the task. At low levels of recall, adding the session features typically improves accuracy by 50% or more. For example, at recall level 0.10, precision for the query model is 0.057 (shown in the curve with open red circles), and adding the session features increases precision to 0.091 (shown in the curve with the filled red circles). The best performance is obtained when all three classes of variables are used, resulting in precision of 0.104 at recall 0.10. For this model, the most predictive features in the logistic regression include query features (*queryLength*, *avgTokenLength*), session features (*timeInSession*, *actionsInSession*), and user features (*avgSessionLengthURLs*).

Figure 7 shows the precision-recall curves for sessions with three or more observed queries, since such sessions provide additional context about the user’s progress on their task. The overall pattern of results is very similar. When considered individually session features are better than query features which are better than user features, and the best performance is obtained using all three types of features. There are also some interesting differences compared with the overall performance seen in the previous figure. First, prediction accuracy is much higher – *e.g.*, at 0.10 recall, the precision for the full model is now 0.235 compared with 0.104 in the previous figure. This is a result of longer sessions providing more context to identify sequence motifs and due to task differences. More difficult tasks result in longer sessions and more switching. Second, the session variable provides more of a lift when added to the user and query variables than it did previously. At low levels of recall, adding the session features typically improves accuracy by 200-300% or more. For example, at recall level 0.10, precision for the query model is 0.059 (shown in the curve with open red circles), and adding the session features increases precision to 0.172 (shown in the curve with the filled red circles). The best performance is obtained when all three classes of variables are used, resulting in precision of 0.235 at recall 0.10. For this model, the most predictive features in the logistic regression include session features (*timeInSession*, *actionsInSession*, *numPaginations*),

query features (*queryLength*), and user features (*avgSessionLengthURLs*). In addition, the two sequence motif features (*hasMotifAdvanced*, *hasMotifBasic*) are also strongly predictive indicating that the abstract patterns of behavior, such as *qR*sPbR* (*i.e.*, multiple queries with no clicks, then a single SERP click and a SERP revisit), can improve prediction accuracy. The *currentSequenceBasic* or *currentSequenceAdvanced* features were not strongly predictive because they required an exact match between a learned sequence appearing in the training data and the sequence generated from recent session interaction. More experimentation with sequences is required, especially with sequence suffixes that target recent session interaction over all session interaction.

Predicting which (if any) actions during the course of a session will involve a switch to another search engine is a challenging task, in part because of the low frequency of such events. Using features of the query, user and search session (prior to the switch), we can predict switches with much higher accuracy than a simple baseline model. Although the absolute level of performance is not too high, we believe that it is sufficient to support some kinds of user support (*e.g.*, additional query suggestions or other search aids), especially in the case of longer sessions.

6. DISCUSSION AND IMPLICATIONS

A primary focus of this research has been the characterization of search engine switching behavior. Through our analysis we have shown that approximately 4% of search sessions involve one or more switches between search engines. We have also shown that this percentage increases to over 10% for longer search sessions. The reasons for switching are varied and include: perceived poor quality of results on original engine, desire for verification or additional coverage, and user preferences. Approximately half of all users in our log sample and around two-thirds of survey respondents engage in within-session switching. It is clear that the utilization of multiple search engines is an important aspect of users’ Web search behavior. Since switching is mainly associated with dissatisfaction with the search results on the origin engine, that engine could tailor the search experience for queries with a high observed switching rate.

Given that search engine switching may also be attributed to a desire for additional information, a search engine may wish to discourage switching away from their engine by offering topic coverage or redundancy (for verification purposes) as optional ranking criteria in addition to relevance. Tools to proactively notify users when other engines may have different results or results that support or refute a line of argument could also help users.

Though studying the pre-switch activities of search engine users we identified important patterns through temporal analysis and sequence motifs. Our findings revealed that some actions, such as SERP clicks and non-SERP clicks, decreased before a switch, whereas queries and navigation to other pages increased. Influential interaction sequences also emerged as important from the survey data and log-based analysis. For example, repeat submissions of queries followed by no SERP clicks, was the most discriminating sequence motif. By better understanding pre-switch behavior we can personalize switch predictions to the current user and their search context. In addition, we can use global switching rates for different queries or search patterns, independent of user.

We analyzed the post-switch activities of users with a particular focus on search success. Overall, switching to another search engine does not provide a quick resolution to a user’s information need. In fact, we found that users perform more queries and ac-

tions on the destination engine, and do not appear to be more successful (as measured by *NoClick* and *SatAction*). One reason may be that the queries that users switch on are difficult, making it likely that neither engine will provide relevant search results, or that the other engines do not provide any additional information over the origin engine. Further exploration is needed in the identification of different motives for switching and dividing the analysis to determine their effect on search satisfaction.

We examined the use of several types of features for the difficult task of predicting switching during the course of a session. The findings showed that models trained using query, session, and user features performed best for all sessions and for sessions with three or more search queries. We achieved levels of performance that we believe will be useful in supporting some kinds of user assistance. For example, additional query suggestions, or richer support for sorting or filtering using metadata could be provided. We also showed that some level of prediction accuracy could be obtained by using simple features of the query (e.g., query length and average number of search results). These could be used to construct a query-only switch prediction model that is not dependent on session or user history information.

One limitation of this work is the focus on the three most popular search engines: Google, Yahoo!, and Live Search. More examples of switching behavior would be observed if additional engines were considered in the analysis. There may be noteworthy behaviors and rationale in the switches from popular engines to less popular search providers, such as vertical search engines.

7. CONCLUSIONS AND FUTURE WORK

We have presented a characterization of search engine switching behavior and an examination of several types of features for the challenging task of predicting switch search engines. We have drawn from findings from a large scale log-based analysis and a large user survey to improve our understanding of how, when, and why users switch engines. Survey findings revealed that switching is not only a result of dissatisfaction with the origin engine; it is also frequently related to user preferences and a desire to verify or find additional information. Survey respondents identified common behaviors preceding a switch that were also identified as significant in log analysis. These findings plus additional insights gleaned from the logs were used to inform feature selection for logistic regression models that let us examine predictive value of query, session, and user features. Predictive models may be useful for search engines who may want to modify the search experience if they can accurately anticipate a switch. Our findings suggest that the predictive models provide sufficient signal to provide some additional user support, especially at low recall. More importantly, we demonstrated the relative value of each feature class and highlighted individual features that may be useful predictors.

In future work, we will develop improved predictive models using new features and alternative learning algorithms. In addition, we would like to further distinguish different motivations for switching (e.g., dissatisfaction with original engine, desire to verify or diversify results) and develop models and the appropriate end-user support for each. Better understanding how to help users identify the vertical search engines or other general search engines that could provide diversity of focus, also presents an important investigative opportunity. Finally, to better understand longitudinal behaviors, we will study between-session and long-term switches.

8. REFERENCES

- [1] Bartell, B.T., Cottrell, G.W., and Belew, R.K. (1994). Automatic combination of multiple ranked retrieval systems. *Proc. SIGIR*, 173-181.
- [2] Belkin, N., Cool, C., Croft, W., and Callan, J. (1993). The effect of multiple query representations on information retrieval system performance. *Proc. SIGIR*, 339-346.
- [3] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.
- [4] Carmel, D., Yom-Tov, E., Darlow, A., and Pelleg, D. (2006). What makes a query difficult? *Proc. SIGIR*, 390-397.
- [5] Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*. New York: John Wiley and Sons.
- [6] Cronen-Townsend, S., Zhou, Y. and Croft, W. B. (2002). Predicting query performance. *Proc. SIGIR*, 299-306.
- [7] Downey, D., Dumais, S.T., and Horvitz, E. (2007). Models of searching and browsing: Languages, studies and application. *Proc. IJCAI*, 2740-2747.
- [8] Downey, D., Dumais, S.T., Liebling, D., and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. *Proc. CIKM*, 449-458.
- [9] Fox, S., Karnawat, K., Mydland, M., Dumais, S.T., and White, T. (2005). Evaluating implicit measures to improve the search experience. *ACM TOIS*, 23(2), 147-168.
- [10] Grimes, C., Tang, D., and Russell, D. (2007). Query logs are not enough. *Proc. Workshop on Query Log Analysis*.
- [11] Heath, A.P. and White, R.W. (2008). Defection detection: Predicting search engine switching. *Proc. WWW*, 1173-1174.
- [12] Hosmer, D.W. and Lemeshow, S. (2004). *Applied Logistic Regression*. New York: Wiley.
- [13] Huntington P, Nicholas D, Jamali, H.R., and Watkinson A. (2006). Obtaining subject data from log files using deep log analysis: case study OhioLINK. *J. Inf. Sci.*, 32(4): 299-308.
- [14] Juan, Y.F. and Chang, C.C. (2005). An analysis of search engine switching behavior using click streams. *Proc. WWW*, 1050-1051.
- [15] Laxman, S., Tankasali, V., and White, R.W. (2008). Stream prediction using a generative model based on frequent episodes in event sequences. *Proc. SIGKDD*, 453-461.
- [16] Leskovec, J., Dumais, S., and Horvitz, E. (2007). Web projections: Learning from contextual subgraphs of the web. *Proc. WWW*, 471-480.
- [17] Ling, C. and Li, C. (1998). Data mining for direct marketing: Problems and solutions. *Proc. SIGKDD*, 73-79.
- [18] Mukhopadhyay, T., Rajan, U., and Telang, R. (2004). Competition between internet search engines. *Proc. HICSS*.
- [19] Pew Internet and American Life Project. (2005). *Search Engine Users*. Accessed December 15, 2008.
- [20] Teevan, J., Dumais, S., and Liebling, D. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. *Proc. SIGIR*, 620-627.
- [21] Telang, R., Mukhopadhyay, T., and Wilcox, R. (1999). An empirical analysis of the antecedents of internet search engine choice. *Proc. Wkshp on Info. Systems and Economics*.
- [22] White, R.W., Richardson, M., Bilenko, M., and Heath, A.P. (2008). Enhancing web search by promoting multiple search engine use. *Proc. SIGIR*, 43-50.
- [23] Zhou, Y. and Croft, W.B. (2007). Query performance prediction in web search environments. *Proc. SIGIR*, 543-550.