Why Searchers Switch: Understanding and Predicting Engine Switching Rationales

Qi Guo[1], Ryen W. White[2], Yunqiao Zhang[2], Blake Anderson[2], and Susan T. Dumais[2]
[1] Mathematics and Computer Science, Emory University, Atlanta, GA 30322, USA
[2] Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA
qguo3@emory.edu, {ryenw,yuzhan,blakean,sdumais}@microsoft.com

ABSTRACT

Search engine switching is the voluntary transition between Web search engines. Engine switching can occur for a number of reasons, including user dissatisfaction with search results, a desire for broader topic coverage or verification, user preferences, or even unintentionally. An improved understanding of switching rationales allows search providers to tailor the search experience according to the different causes. In this paper we study the reasons behind search engine switching within a session. We address the challenge of identifying switching rationales by designing and implementing client-side instrumentation to acquire in-situ feedbacks from users. Using this feedback, we investigate in detail the reasons that users switch engines within a session. We also study the relationship between implicit behavioral signals and the switching causes, and develop and evaluate models to predict the reasons for switching. In addition, we collect editorial judgments of switching rationales by third-party judges and show that we can recover switching causes *a posteriori*. Our findings provide valuable insights into why users switch search engines in a session and demonstrate the relationship between search behavior and switching motivations. The findings also reveal sufficient behavioral consistency to afford accurate prediction of switching rationale, which can be used to dynamically adapt the search experience and derive more accurate competitive metrics.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*selection process, search process.*

General Terms

Algorithms, Measurement, Experimentation, Human Factors.

Keywords

Search engine switching, predicting switching rationales.

1. INTRODUCTION

Search engines facilitate rapid access to information on the Web. A user's decision regarding which engine to use most frequently (their primary engine) can be based on factors including reputation, familiarity, effectiveness, interface usability, and satisfaction [18], and can significantly impact their levels of search success [22][24]. Similar factors can influence a user's decision to switch engines, either for a particular query if they are dissatisfied with results or seek broader topic coverage, or for specific task types if another engine specializes in such tasks, or more permanently due to unsatisfactory experiences or relevance changes [23]. The barrier to switching Web search engines is low and multiple engine usage is common. Indeed, prior work in this area suggests that 70% of Web searchers use multiple search engines [24].

Previous work on switching has examined promoting multiple search engine use [24], characterizing both short- and long-term engine switching behavior [22][23], predicting when users will switch [12][15][22], developing metrics for competitive analysis of engines using estimated user preference and user engagement [13], or building conceptual and economic models of search engine choice [17][21]. Despite the economic significance of engine switching to search providers, and its prevalence among search engine users, little is known about users' rationales for switching, the search behaviors that may relate to different rationales, or the features that are most useful in automatically predicting different switching causes. This is in part due to difficulty in reliably identifying switching rationales.

Improved understanding of switching rationales is important and has several applications. If search engines could predict that a searcher was about to switch along with the reason for that switch, they could adapt the search experience accordingly. For example, if a user searching for [*butterfly flight patterns*] is about to switch because they could not find what they were looking for (i.e., the *reason* is dissatisfaction), then the search engine can intervene and provide help such as additional query support. Conversely, if a user searching for [*bellevue apartments*] is about to switch to check for additional offerings (i.e., the *reason* is topic coverage or verification), an intervention could be annoying. Even in offline settings, search providers can improve the search experience through a better understanding of switching rationales. For example, queries resulting in dissatisfaction-related switches can be identified and analyzed to improve search quality.

In this paper, we focus on understanding users' motivations for search engine switching within a single search session. We address the challenge of identifying switching motivations by implementing and deploying a browser extension to capture search interaction and acquire explanations from searchers in situ (i.e., at the time they switch search engines). The in-situ explanations provide first-hand insights about the reasons for switching, and can be used to understand the relationships between patterns of search interaction and the switching causes. We investigate the effect of different interaction features (derived from the query, pre-switch user behavior, and post-switch behavior) on the accuracy of models to automatically predict the reason for a switch.

The main contributions of this paper include:

- The implementation and deployment of client-side instrumentation to collect in-situ searcher explanations that enables the study of session-level switching rationales.
- An in-depth analysis of session-level engine switching rationales and their correlations with behavioral signals.

SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07...$10.00.

- An investigation and evaluation of predicting switching causes with various session-level behavioral features.

This study is the first research to understand switching causes in-situ and to predict the reasons for search engine switches, laying the groundwork for more extensive future work on engine switching and search satisfaction.

The remainder of this paper is structured as follows. Section 2 outlines previous work on predicting query difficulty, user satisfaction and frustration, and characterizing search engine switching behavior. Section 3 provides an overview of the study. Section 4 describes the design and implementation of our client-side instrumentation for collecting in-situ explanations. In Section 5 we characterize the session-level switching causes, and their correlations with search behavior. In Section 6, we investigate prediction of switching rationales, varying features used. In Section 7, we explore using editorial judgments to approximate the in-situ explanations. We discuss our findings and their implications in Section 8 and conclude in Section 9.

2. RELATED WORK

Three lines of work are most relevant to our research: (i) predicting query performance, (ii) user satisfaction and frustration, and (iii) characterizing and predicting engine switching behavior.

Research on *predicting query performance* has been conducted to understand differences in the quality of search results for different queries. Such predictions can be used to devote additional resources or use alternative methods to improve search results for difficult queries. While it has been shown that using different query representations [4] or retrieval models [3] improves search performance, it is more challenging to accurately predict which methods to use for a particular query.

Measures such as query clarity [7], Jensen-Shannon divergence [6], and weighted information gain [1] have been developed to predict performance on a query (as measured by average precision, for example). Guo *et al.* [10] used interaction features, including switching features, to predict query performance. 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. Hassan *et al.* [11] developed methods to predict search success on a session-level. Feild *et al.* [8] developed methods to predict user frustration, and showed that features capable of accurately predicting switching events were also highly predictive of frustration.

Some research has examined *search engine switching behavior*. Early research by Mukhopadhyay *et al.* [17] and Telang *et al.* [21] has 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. Juan and Cheng [13] 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.

Heath and White [12] 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 [12] 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.

White *et al.* [24] 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 can 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. White and Dumais [22] characterized search engine switching through a large-scale survey and built predictive models of switching based on features of the pre-switch query, session, and user. White *et al.* [23] modeled long-term engine usage over a six-month period, and identified three user classes: (i) those who do not switch, (ii) those who switch at some time, and (iii) those who switch back and forth between different search engines.

We extend previous work on search engine switching by focusing on *understanding the rationales* for search engine switching in a single search session in depth, and on *predicting switching rationales* given features of the query and pre-/post-switch behavior.

3. STUDY OVERVIEW

We begin by providing definitions used throughout the paper. We then provide an overview of the reasons for engine switching that were identified in previous work, and present the research questions that we address in this paper.

3.1 Definitions

Some terms are formally defined as below.

DEFINITION 1. A *search session* is a sequence of user activities that begins with a query, includes subsequent queries and URL visits, and ends with a period of inactivity. URL visits include both clicks on the search engine result page (SERP clicks) and post-SERP navigation. We used 30 minutes of user inactivity to mark session boundaries; similar timeouts have been used previously to demarcate search sessions in Web log analysis [22].

DEFINITION 2. A *search engine switching event* is a pair of consecutive queries that are issued on different search engines within a single search session. In our definition of a switching event, navigational queries for search engine names (e.g., search on Yahoo! for [*google*], [*google.com*], etc.) are regarded as part of the act of switching and not as the pre- or post-switch query. For example, if a Yahoo! user searches for [*snowshoes*], then for [*google*] and switches to Google, and then searches for [*snowshoe merchants*], then the pre-switch query is [*snowshoes*] (and not [*google*]), and the post-switch query is [*snowshoe merchants*].

DEFINITION 3. A *search goal* is an atomic information need, resulting in one or more related search queries issued to accomplish a single discrete task [11]

3.2 Reasons for Search Engine Switching

In earlier research on characterizing and predicting search engine switching behavior, White and Dumais [22] surveyed 488 users regarding their experiences with search engine switching. They asked respondents to provide retrospective rationales for switching by selecting at least one explanation from a list of possible reasons. For reader reference, in Figure 1 we present the response breakdown reported in their study.

As we can see from the figure, there are three classes of reasons: dissatisfaction (DSAT) with the results in the original engine (dissatisfaction, frustration, expected better results; shown in red in Figure 1 and totaling 57%), verifying or finding additional information (coverage/verification, curiosity; in green and referred to as "Coverage", totaling 26%), and user preferences (destination preferred, destination typically better; in blue, totaling 12%).

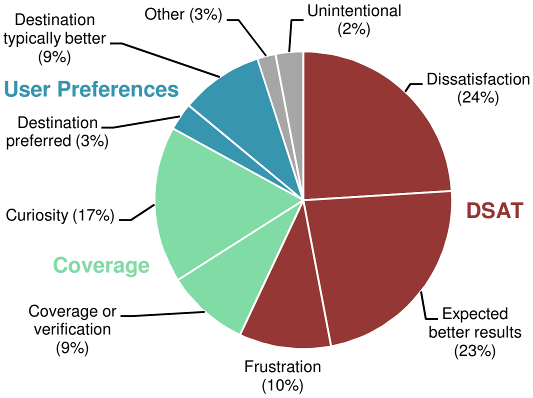


Figure 1. Reasons for search engine switching.

In this study, we collected data about switching rationales at the time the switches happen and develop models to distinguish between DSAT- and coverage-related switches using session-level evidence. We focus on these classes since they are most common, are directly actionable by search providers, and unlike preference-related switching, do not require knowledge of users' long term behavior, which may not be available to the engine at query time.

3.3 Research Questions

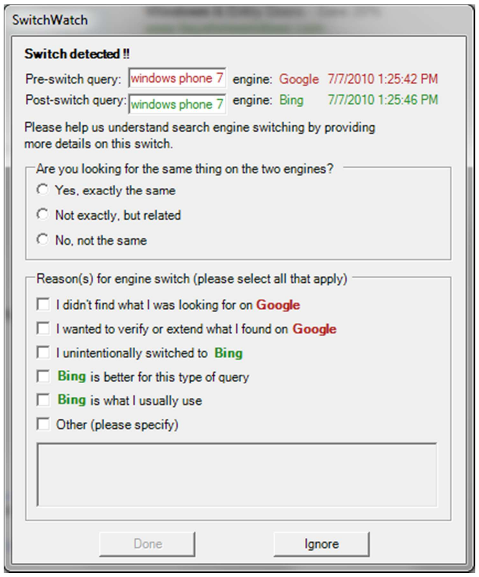
Our study answers a number of research questions:

1. Why do searchers switch search engines?
2. Which behavioral signals are associated with different causes?
3. How accurately can we predict the causes of engine switching?

Answers to these questions can help us better understand switching and help search providers improve the user experience in accordance with searchers' motivations for switching or derive more sophisticated competitive metrics for comparing search engines.

4. IDENTIFYING RATIONALES

The first challenge we addressed was collecting switching rationales and their associated search behaviors. One possibility is to design difficult search tasks and ask participants to try to complete these tasks (e.g., [1][8]). For this approach, the challenge lies in the design of tasks that will induce switching behavior naturally. Since switching behavior is rare [22], it is difficult to collect a sufficient amount of switching data in a laboratory setting if participants are not instructed to switch but this will bias the data. Another possibility is to sample from search engine logs and recruit human judges to label sessions. Previous research has demonstrated feasibility in asking human judges to label search success [11]. However, switching rationales are more subjective and personal since they relate not only to searcher goals and result relevance, but also to other factors such as long-term searcher preferences and browser settings. As a result, it is challenging to identify real switching rationales by examining log data alone. Another option is to ask searchers why they switch search engines. White and Dumais [22] asked people to summarize their reasons for switching using a retrospective questionnaire. While this provides some interesting insights, retrospective surveys do not always align with actual behavior and the corresponding behavioral data is not available. Thus, we chose to ask searchers about their switching behavior *in situ* when they switch engines. To obtain these *in-situ* switching assessments we implemented and deployed a browser add-on, called *SwitchWatch*, which presents a short questionnaire to the user at the time of a switch between Google, Yahoo!, or Bing. This questionnaire elicits switching rationales directly from the user at switch time. Figure 2 shows the questionnaire which contains questions about the search task and switching rationale (as described in more detail below).



Switch detected !!

Pre-switch query: windows phone 7 engine: Google 7/7/2010 1:25:42 PM

Post-switch query: windows phone 7 engine: Bing 7/7/2010 1:25:46 PM

Please help us understand search engine switching by providing more details on this switch.

Are you looking for the same thing on the two engines?

○ Yes, exactly the same
○ Not exactly, but related
○ No, not the same

Reason(s) for engine switch (please select all that apply)

□ I didn't find what I was looking for on **Google**
□ I wanted to verify or extend what I found on **Google**
□ I unintentionally switched to **Bing**
□ **Bing** is better for this type of query
□ **Bing** is what I usually use
□ Other (please specify)

Figure 2. Example *SwitchWatch* questionnaire, where the pre- and post-switch queries are [*windows phone 7*].

This approach allows us capture the subjectivity or variation in switching rationales among different users. Not surprisingly, the main challenge of this method is the slow pace with which labels are obtained since switching occurs infrequently [22]. The pop-up survey may also be irritating, and lead to users ignoring the survey, or terminating participation. However, since switching does not occur often, this is not a big concern. Further, we allow people to ignore the questionnaire (by clicking on the *Ignore* button).

We deployed *SwitchWatch* within Microsoft Corporation for one month to collect sufficient data. We now provide details on its implementation and deployment.

4.1 SwitchWatch Implementation

SwitchWatch was implemented as an add-on for the Internet Explorer browser. Once installed, it started automatically every time a new browser tab was opened and recorded browser activity in that tab to a remote server. For each visited Web page, we recorded the URL, the timestamp, a unique browser window identifier, and an anonymous user identifier. A separate instance of *SwitchWatch* ran for each active browser tab and tab focus/blur events were recorded, allowing us to accurately identify multi-tab usage and Web page dwell times. Intranet and secure (https) URL visits were excluded to help maintain user privacy.

Once an engine switch between two of Google, Yahoo!, and Bing was detected (per definition 2), *SwitchWatch* displayed a questionnaire in the center of the screen occluding part of the active browser window. The questionnaire is shown in Figure 2 for a switch between Google and Bing on query [*windows phone 7*].

As we see in Figure 2, there are two questions shown to users: (i) whether the user changed their search goal (to help determine the extent to which switching was related to goal shifts), and (ii) the cause of the switch. For the first question, the user can pick one response from “exactly the same,” “related,” and “different” via the radio buttons. For the second question, the user can select multiple reasons that apply. The options include dissatisfaction, coverage, and unintentional. In addition, two options about user habit and preferences are included, namely, “the post-switch engine is better for this type of query” and “the post-switch engine is what I usually use.” These response options were selected based on the survey responses in [22], and summarized in Figure 1.An *Ignore* button is also provided to allow users to skip the pop-up if they do not want to interrupt their current task to answer.

4.2 SwitchWatch Deployment

We distributed an invitation to deploy the *SwitchWatch* add-in via email to approximately 2,200 employees within Microsoft Corporation, including colleagues in affiliated groups, interns and 1,000 randomly-selected full-time employees from across the organization. Invitations were sent to employees with a diverse range of occupations, from software engineers and testers, to program managers, paralegals, and administrative staff. 216 employees participated in the study by installing the add-on on their machine, for a response rate of around 10%. Privacy restrictions prevented us from determining the identities or occupations of participants who accepted the invitation to participate. We ran this study for approximately four weeks. In each week, we randomly selected a participant with *SwitchWatch* installed for the week and rewarded them with a 50 USD gift card. There were no other usage requirements to be considered for the prize drawing. Our goal was to retain participants in our study, and make sure that they behaved normally, without forcing them to switch search engines as they may not do normally in their daily search routine.

5. CHARACTERISTICS OF SWITCHING

We now discuss findings on switching characteristics. First, we present the overall summary of search engine switching in the in-situ logs. Second, we discuss how frequently users change search queries and search goals when they switch engines, and the relationship between changing search queries and different switch causes. Finally, we characterize the different switch causes and their associated search behaviors in the session.

5.1 Definitions

We start by providing some definitions used in this section:

- **Same Query (*SQ*):** Identical pre- and post-switch queries.
- **Related Queries (*RQ*):** Pre- query and post-switch query share at least one query term that is not a stop word, but are not *SQ*.
- **Different Queries (*DQ*):** Pre-switch query and post-query do not share any (non-stop-word) terms.
- **Same Goal (*SG*):** User has the same search goal.
- **Related Goals (*RG*):** Search goals related but not same.
- **Different Goals (*DG*):** Search goals are totally different.
- **Ignored:** User dismissed the *SwitchWatch* questionnaire without providing feedback by clicking *Ignore* button.

5.2 Data Overview

We now provide more details on the data that were gathered as part of our experiment. We begin by describing relevant features of the log data gathered by the *SwitchWatch* add-in.

In the in-situ log, 20,554 queries were issued on Google, Yahoo! or Bing by our participants in the four-week duration of the study. Among all the queries, we observed 1029 switches. We excluded 25 switches that were suggestive of users testing *SwitchWatch* (e.g., assessments with queries [*test*], [*hello world*], etc.). As a result, we considered 1,004 instances of search engine switching events, of which 562 (56%) received in-situ assessments. In the remaining 45% of switches, participants clicked *Ignore*, indicating that they did not wish to offer a reason for the switch at that time.

The 1,004 session-level switches in our set comprise only 4.2% of all queries, while 107 (49.5%) of the 216 users who installed the add-in switched search engine within a session at least once (similar to switching rates reported in [22] which was for a much larger and more heterogeneous sample of search engine users).

5.3 Switch Causes and Goal Changes

In this section, we investigate how the changes of search queries and search goals relate to the underlying switching rationales. The inspiration of this analysis came from the observed high percentage of query changes in engine switching events.

5.3.1 Query Changes

We begin by analyzing the breakdown of query changes during engine switching. As described earlier, query change measured by the overlap between pre- and post-switch queries. In our logs, only around 32% of the engine switching events observed had an identical pre- and post-switch query (*SQ*), and approximately 50% of the query pairs shared at least one query term (*SQ+RQ*), the remaining 50% of query pairs comprised different queries *(DQ)*. This raises a question regarding search goal inconsistency during engine switching, something that we explore next.

5.3.2 Goal Changes

We now examine the relationship between changes of search queries and changes of search *goals* (i.e., characterized by *SG, RG,* and *DG* as defined earlier). Table 1 shows the breakdown of user-reported goal changes with respect to query changes. The table shows the percentage of switches where participants clicked *Ignore* and the fraction of remaining judged (non-ignored) queries that received each goal change label.

Almost all (98%) of the switches with the same pre- and post-switch query *(SQ)* share the same search goal (*SG*). In contrast, only 20% switches with related pre- and post-switch queries (*RQ*) are reported as having related goals (*RG*), and only 65% of switches with different pre- and post-switch queries (*DQ*) are reported as having different search goals *(DG)*. This difference is mainly due to the contributions of *RQ* (77% *SG*) and *DQ* (23% *SG*) to switches with same goal. This seems reasonable since users might change the query terms slightly or perhaps make typos when switching with the same search goal. Nevertheless, *SQ* is the best proxy of *SG*, while *RQ* could also be a proxy for *SG* to increase coverage, and *DQ* could be a reasonable proxy of *DG.*

Table 1. Breakdown of goal and query changes.

Query change	Ignored	Goal change [% judged (non-ignored) queries]		
		SG	*RG*	*DG*
All	45%	65%	9%	25%
SQ	27%	98%	1%	1%
RQ	39%	77%	20%	3%
DQ	60%	23%	12%	65%

There are differences in how often the *SwitchWatch* dialog is ignored (by selecting the *Ignore* button) for different query changes. The more related the queries, the more likely users are to provide feedback. The ignore rate of *SQ* switches is 27%, while the ignore rate of *DQ* increases to 60%. Note that we cannot compute the ignore rates for goal changes (*SG*, *RG*, *DG*) since we do not have the search goal information for ignored switches because it was captured explicitly from users by *SwitchWatch*.

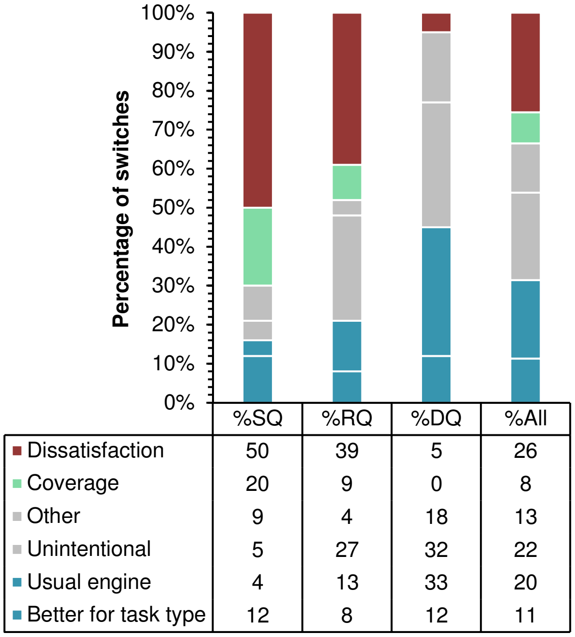
5.3.3 Impact on Switch Causes

Now, we study how the change of search queries (as a proxy of search goal change) is related to the underlying switch causes.

As mentioned earlier, the causes were collected via the *SwitchWatch* questionnaires presented to participants when they were about to switch search engines. The candidate set of causes was derived from previous work and included: (i) dissatisfaction with current engine, (ii) additional topic coverage or verification, (iii) unintentional, (iv) post-switch engine was better for this type of task, (v) returning to usual engine, and (vi) other.

In Figure 3 we present the six causes of search engine switching for each class of query changes (*SQ*, *RQ*, *DQ*), and across all switches (*All*). The colors correspond to those used in Figure 1, but unlike Figure 1, these explanations were captured at switch time rather than based on users' recollection of switching events. Focusing on the distribution for *All* in Figure 3, we compare the in-situ reasons and the retrospective reasons reported in Figure 1. There are differences in the figures. For example, dissatisfaction and coverage were less common in-situ than retrospective, and reasons associated with user preferences and unintentional or other reasons were more common in-situ than retrospective. There are some differences in the questions used in the two cases, which make it challenging to map directly between them, as well as a different set of respondents. However, there are some interesting relations, e.g., *dissatisfaction* in Figure 1 includes "dissatisfied" (24%), "expected better" (23%), and "frustrated" (10%), but in Figure 3, dissatisfaction is only "dissatisfied" (26%) and corresponds well with the 24% from the retrospective study. Alternatively, the *SQ* results in Figure 3 correspond well with results in Figure 1. One explanation could be that when people consider switching retrospectively, they focus on *SQ* instances.

There are also different distributions of reasons for switching for the different query relations. The reasons for *SQ* switches are primarily dissatisfaction and coverage, whereas preferences (*usual engine* and *better for task type*) and other (*unintentional* and *other*) are the primary reasons for *DQ*. The reasons for *RQ* lie between these two extremes. Interestingly, as shown in the figure, most of the *DQ* switches are unintentional or indicate a return to respondents' preferred search engine, and only 5% of them were caused by dissatisfaction. This suggests that *DQ* switches may be misleading if used in deriving performance-related metrics, although they are interesting to understand general preferences.



	%SQ	%RQ	%DQ	%All
🟥 Dissatisfaction	50	39	5	26
🟩 Coverage	20	9	0	8
⬜ Other	9	4	18	13
⬜ Unintentional	5	27	32	22
🟦 Usual engine	4	13	33	20
🟦 Better for task type	12	8	12	11

Figure 3. Relationship between query changes and switch causes. Values are shown in the bars in the same order as in the table below the chart.

In contrast, *SQ* has the greatest fraction of queries associated with intentional causes such as dissatisfaction and coverage. Switches caused by dissatisfaction are perhaps the most interesting to a search provider; pre-switch queries are those that the origin engine can improve on in order to retain users, while post-switch queries are important for the destination engine to perform well on in order to gain users. Therefore, we focus on *SQ* switches for later analysis in predicting switching rationales.

In this section we have studied the change of switching queries, search goals, and their connections with the causes of engine switching. We now turn our attention to interaction behaviors associated with different switching motivations. Knowledge of these behaviors may help us predict the reasons behind engine switches given observed session-level search interactions.

5.4 User Behaviors for Same-Query Switches

We study the behaviors surrounding engine switching using the following three groups of features: (i) query, (ii) pre-switch behavior, and (iii) post-switch behavior. These features are useful in characterizing switching and may be useful for predicting switching rationales. All of the features used are described in Table 3, along with the mean and the standard deviation feature values from the in-situ log data for switches associated with dissatisfaction (*DSAT*), *Coverage*, and all other reasons (*Other*). For each feature we perform one-way independent measures analyses of variance (ANOVA) between the three causes of switching, and indicate in the table which paired differences are significantly different with Tukey *post-hoc* testing, for significance with $p <$ .01 and $p <$ .05. We now describe each feature group in more detail, with reference to Table 2 where appropriate.

Query features derived from the query string itself, include the query length in words and characters, and the time between the pre- and post-switch queries. The longer the query, the more likely the searcher is to be dissatisfied with the search results. Table 2 shows that the number of words in queries associated with dissatisfaction is approximately four, while number of words for switching queries for coverage and other are around three, although the results are not reliable statistically. This is consistent with previous work showing that query length can be an important determinant of search success in Web search [1]. Also included is the time between the pre-switch and post-switch query. The time between the queries is longer for switches associated with dissatisfaction, perhaps because the user is considering all top-ranked search results before making the decision to switch engines.

Pre-switch behavior features from the search interaction before the engine switch include the number of queries (total, unique, and reformulations), the number and rate of clicks broken down by satisfaction and bounce, and the length of the post SERP-trail. Satisfaction with a clicked result (denoted Sat in Table 2) is defined to be a dwell for longer than 30 seconds, as in previous work in this area by Fox *et al.* [9]. Bounces are defined as clicks with a dwell time on the landing page of under 15 seconds, where the searcher returns to the SERP after viewing the landing page. The number of queries is higher for switches associated with coverage, significantly so for the number of queries (*pre_q*). In coverage scenarios, users visit more search results (*pre_c*) and tend to dwell longer on these results (*pre_c_SAT*). We would also expect that the smaller the number of the satisfied clicks and the more *bounces*, the more likely the user is dissatisfied. Trends are in this direction, significantly between the number of bounce clicks (*pre_c_Bounce*) for *DSAT* switches and *Other*. Reformulation rates (*pre_reformRate*) are also somewhat higher for *DSAT* versus *Other*. Also of interest are the number of pages on trails following the SERP click and the number of pages on those trails with the query in their title, the latter of which (*pre_c_containsQ*) appears slightly higher for coverage switches, although not significantly.

Post-switch behavior features correspond to the pre-switch behaviors described earlier in this section. Post-switch behaviors can provide insight into the nature of the search task, and hence potentially the reason for the search engine switch. Interestingly, none of the differences for post-switch behavior were statistically significant. This suggests that post-switch behavior may be less useful for differentiating between switching motivations, something that we will return to later in the paper when we discuss prediction. That said, trends in the findings revealed some similarities to pre-switch behaviors. For example, in switches associated with coverage, there are more queries in the session (*post_q*), and more query reformulation for *DSAT* switches (*post_reform*).

There are noticeable (and some significant) differences in the interaction behavior associated with different switching rationales. In addition to characterizing the interactions associated with the different rationales, we were also interested in whether we could *predict* the reasons behind switching given evidence of searcher interaction behavior within a search session. We now describe our work on predicting engine switching rationales using interactions.

6. PREDICTING SWITCHING CAUSES

We built and evaluated classifiers to predict the reasons for search engine switching. For each reason, we formulate the prediction task as binary classification, where the goal is to predict whether an observed switch is attributable to the reason of interest. We also experimented with multi-class (tertiary) prediction, where the goal was to correctly attribute one of three switching explanations —*DSAT*, *Coverage*, and *Other* (everything else)—to an observed switching event. In this section we describe the results from our

Table 2. Features of search behavior per switching cause. Bolded features exhibit statistically-significant differences. *Dissatisfaction (DSAT)* versus *Other*: ^{n} *p*<.05; $^{\circ}$ *p*<.01; *Coverage* versus *Other*: ○ *p*<.05; ● *p*<.01.

Feature	Mean (stdDev)		
	DSAT	*Coverage*	*Other*
Query features			
q_charLength: Num. chars in switching query	27.3 (±22.5)	18 (±13.5)	18.2 (±12.5)
q_wordLength: Num. words in switching query	4.2 (±3.2)	3.2 (±2.4)	2.9 (±2.0)
q_timeDiff: Time in seconds between the pre-switch and post-switch queries	80.9 (±134)	52.5 (±98.9)	46.7 (±80.1)
Pre-switch features			
pre_q●: Num. queries in session	2.7 (±2.0)	4.0 (±3.5)	2.1 (±1.8)
pre_uniqQ: Num. unique queries in session	2.0 (±1.5)	2.1 (±1.9)	1.4 (±1.0)
pre_reform: Num. query reformulations	1.0 (±1.5)	1.2 (±1.9)	0.4 (±1.0)
pre_uniqQRate: *pre_uniqQ* / *pre_q*	0.8 (±0.3)	0.7 (±0.3)	0.8 (±0.3)
pre_reformRate$^{\circ}$: *pre_reform* / *pre_q*	0.2 (±0.3)	0.2 (±0.3)	0.1 (±0.2)
pre_c: Num. SERP clicks for related queries	2.0 (±2.9)	2.0 (±2.5)	0.7 (±1.1)
pre_c_Sat: Num. satisfied SERP clicks for related queries	0.8 (±1.7)	1.0 (±1.7)	0.4 (±0.9)
pre_c_Bounce$^{\circ}$: Num. bounce SERP clicks for related queries	0.8 (±1.2)	0.6 (±1.0)	0.2 (±0.5)
pre_cRate: *pre_c* / *pre_q*	0.9 (±2.1)	0.4 (±0.5)	0.3 (±0.4)
pre_c_SatRate: *pre_c_Sat* / *pre_q*	0.2 (±0.4)	0.3 (±0.4)	0.2 (±0.4)
pre_c_BounceRate: *pre_c_Bounce* / *pre_q*	0.3 (±0.4)	0.2 (±0.3)	0.2 (±0.4)
pre_t: Num. pages on click trail	1.4 (±2.1)	1.6 (±2.1)	0.5 (±0.9)
pre_c_containsQ: Num. SERP clicks on a search result with title containing at least one non-stop-word query term	0.6 (±2.9)	1.7 (±4.4)	0.4 (±1.8)
Post-switch features			
post_q: Num. queries in session	2.5 (±1.9)	2.9 (±3.7)	2.2 (±2.5)
post_uniqQ: Num. unique queries	1.8 (±1.3)	1.4 (±1.1)	1.3 (±0.8)
post_reform: Num. query reformulations	0.8 (±1.3)	0.5 (±1.1)	0.4 (±0.8)
post_uniqQRate: *post_uniqQ* / *post_q*	0.8 (±0.3)	0.7 (±0.3)	0.8 (±0.3)
post_reformQRate: *post_reform* / *post_q*	0.2 (±0.3)	0.1 (±0.2)	0.1 (±0.2)
post_c: Num. SERP clicks for related queries	2.3 (±2.4)	2.1 (±2.8)	1.6 (±2.0)
post_c_Sat: Num. satisfied SERP clicks for related queries	0.9 (±1.2)	0.8 (±1.2)	0.7 (±0.9)
post_c_Bounce: Num. bounce SERP clicks for related queries	1.0 (±1.2)	1.0 (±1.6)	0.7 (±1.0)
post_cRate: *post_c* / *post_q*	1.0 (±0.4)	0.8 (±0.5)	0.9 (±0.5)
post_c_SatRate: *post_c_Sat* / *post_q*	0.3 (±0.4)	0.3 (±0.3)	0.3 (±0.4)
post_c_BounceRate: *post_c_Bounce* / *post_q*	0.3 (±0.4)	0.3 (±0.4)	0.3 (±0.4)
post_t: Num. pages on click trail	2.0 (±2.3)	1.8 (±2.6)	1.2 (±1.8)
post_c_containsQ: Num. SERP clicks on a search result with title containing at least one non-stop-word query term	1.6 (±3.9)	1.3 (±5.9)	3.5 (±8.9)

experiments. We begin by describing the classification algorithm used in the prediction, then describe the evaluation metrics, the models compared in the study, and then the prediction findings.

6.1 Classifiers

We used features described in Table 2 and experimented with a variety of different classification algorithms for predicting switching causes, including decision trees, logistic regression, and naïve Bayes [5], which were the three best performers. The performance of all three methods was similar, and we report on the results of the logistic regression classification here.

6.2 Evaluation Metrics

In evaluating the performance of our predictions, we measure precision (the fraction of predicted instances that were correctly predicted) and recall (the fraction of all true instances that were correctly predicted). We report on the F_β measure, with β set to 0.5, which gives twice as much weight to precision than to recall. Precision is very important in application scenarios for a predictor of switching rationales. In an online scenario, we would want to be highly confident before adapting the search experience based on switch rationale predictions. In an offline scenario, such as studying dissatisfied switches in log data, we need to obtain a set of dissatisfied switches for further analysis. Since there are many switching events in logs, we do not need to classify all switches (have high recall) as long as we can precisely label some.

6.3 Methods Compared

We compare a number of different methods for predicting switching rationales. We used two strong baselines which leverage the marginal distribution and use rules derived manually from a visual inspection of switching events in the logs. The baselines are:

- **Baseline (Prior):** Bases predictions on the class distribution.
- **Baseline (Rule):** Uses rules derived from inspection of switching events in logs. Predict *DSAT* if there is no click before the switch and one or more clicks after the switch; predict *Coverage* if there are clicks before and after the switch; predict *Other* (i.e., all reasons other than dissatisfaction or coverage) if neither of the above rules are triggered.

In addition to these baselines, we also trained binary and tertiary classifiers on varying sets of features described in Table 2:

- **All:** Classifiers trained on all features.
- **Query:** Classifiers trained only using query features.
- **Pre-switch:** Classifiers are trained only using pre-switch behavioral features. These are the features available before engine switching and could be used in combination with a switch predictor (such as that described in [22]) to predict the reason for an anticipated switch.
- **Post-switch:** Classifiers trained only using post-switch behavioral features. This could help the destination engine predict the reason for the incoming switch and adjust the search experience accordingly (e.g., provide diverse results if reason is coverage).

We now present findings on prediction effectiveness using the different feature classes. Given the importance of *SQ* switches to search providers (as discussed earlier in Section 5.3.3), we elected to focus on the set of 354 in-situ *SQ* switches, and used them for training and testing. This is a relatively small set given the fairly intensive *SwitchWatch* deployment effort, primarily because engine switching is a rare event and to maintain reaslism, we did not artificially promote switching in our study. We compare the models via ten-fold cross validation, across 100 randomized experimental runs, and report averages across all runs and folds.

6.4 Binary and Tertiary Predictions

We begin our analysis by comparing the performance of each of the binary classification algorithms with the two baselines, for each of the three classes: *DSAT*, *Coverage*, and *Other*. Table 3 reports the average $F_{0.5}$ values for each switching explanation versus baselines for all features listed in Table 2, and the results of paired t-tests between the models and the baselines.

Table 3. Binary prediction performance (measured via $F_{0.5}$) of *DSAT*, *Coverage*, and *Other*. Significance of differences between models and baselines is marked: Baseline (Prior): △ p<.05, ▲ p<.01; Baseline (Rule): ○ p<.05, ● p<.01.

Method	*DSAT*	*Coverage*	*Other*
Base (Prior)	72.40	27.12	17.40
Base (Rule)	48.84	24.19	20.20
All Features	85.69▲●	47.84▲●	29.01△

The findings presented in Table 3 above show that the prediction model trained with all features significantly outperforms both baselines in the prediction of *DSAT* and *Coverage*, and marginally outperforms the baselines in the prediction of *Other*. The observed gains over the baselines are strong given the limited amount of data available and suggest that there is good predictive signal in the features, espcially for *DSAT* predicitons which are of great interest to search engine providers. Prediction of the *Other* class appears more challenging, perhaps because this class includes several different switching motivations, each of which may have its own associated behavioral patterns.

In addition to the binary classification, we also experimented with multi-class prediction of switch explanation among *three* reasons: *DSAT*, *Coverage*, and *Other*. Multi-class prediction is important because it allows search providers to use a single predictor of switching reasons, potentially reducing training and deployment overhead. Findings from prediction experiments conducted in the same way as above (but this time with three-level judgments) show that the $F_{0.5}$ of a logistic regression model (74.58) exceeded both of the baselines (Prior: 59.46, Rule: 52.80) at $p < .01$. This suggests that multi-class prediction using in-situ data is feasible with these data. However, further work is needed to establish whether a multi-class predictor would outperform a combination of binary classifiers, such as those described above, and study the cost-benefit tradeoffs of each solution.

6.5 Feature Group / Feature Performance

In Table 4 we present the average $F_{0.5}$ metric for how well each model predicts *DSAT* for different sets of features: all features, pre-switch features, query features, and post-switch features.

Table 4. Feature group performance (measured via $F_{0.5}$) for in-situ assessment (*DSAT*).

Group	*All*	*Pre-switch*	*Query*	*Post-switch*
$F_{0.5}$	85.69	81.12	74.28	78.99

All differences between the four feature groups and the baselines were statistically significant using paired t-tests at $p < .05$. The best predictive performance was attained when all features were used. The most important features learned for those predictions (in descending order of importance) and directionality of their DSAT relationship were *pre_c_SatRate* (–), *pre_c* (+), *pre_uniqQRate* (+), *post_c_containsQ* (+), and *post_reformRate* (–). SERP clicks and query reformulation behavior on both the pre- and post-switch engines may be predictive of user satisfaction, a claim supported by [11]. Factors such as low satisfaction on the pre-switch engine were also good predictors of dissatisfaction-related switching.

Pre-switch features appeared to provide more predictive signal than post-switch features, as was suggested in our earlier analysis of those search behaviors (see Section 5.4). Pre-switch interaction behavior reveals more about searchers' behavior leading up to the switching event, and therefore might provide better quality evidence of the reason behind the switch. In addition, it may also be that there is more variance in what users do following a search engine switch, and that makes features of post-switch behavior less reliable indicators of engine switching rationale.

In-situ assessment affords the capture of switching explanations at switch time from the searcher, and should be an accurate elicitation method for switching rationales. However, it may not always be desirable to deploy such a tool to searchers, especially on a large scale, given interruption costs to users, privacy implications, and the infrastructure required to store large volumes of behavioral data. Therefore, we also explored the use of editorial assessments by third-party judges performing manual log analysis to identify the reasons for search engine switching. Including editorial assessments in our study helps us better understand the judgment correspondence between judges and switching users.

7. EDITORIAL ASSESSMENTS

The main advantages of editorial assessments lie in the large volume of switching data available for analysis and the fact that labels can be obtained faster than the in-situ method. In addition, sessions drawn from logs may be more representative of general Web search activity than those from a subset of searchers who elect to participate in in-situ switch monitoring. The disadvantages of editorial assessment include the possible misinterpretation of switching rationales by third-party human assessors and the cost of human labor involved in performing the assessments. Therefore, in this section we consider the relation between *in situ* judgments by the searcher and those by third-party assessors.

7.1 Log Data

We randomly selected 100 search sessions containing at least one same-query engine switch from the in-situ log data used so far in the paper. Each of these switches was judged by two human assessors and a switching reason was assigned to the first same-query switch in the session. We selected sessions from the in-situ logs so we could directly compare the two judgment methods on the same set of switches. As we will describe later in the section, the judgment task was very intensive. To maintain judgment quality we focused on same-query switches to allow judges to focus on identifying the switching *reason* (rather than goal changes, etc.), restricted judgments to the first same-query switch in each session, and limited the number of judgments per judge to 100.

7.2 Judges

Two human judges performed the editorial assessment task. Both judges are Web search researchers who are familiar with engine switching and search log data. Initial training and discussion was conducted as a pair to help ensure consistency in the labeling. The group examined a few example sessions containing switching events and discussed likely switching motivations.

7.3 Editorial Guidelines

Judges were presented with a spreadsheet containing the 100 sessions to be judged. Each row contained a unique session identifier and the URL of the page visited. If the URL was a search engine result page, the query and engine name were also shown, as were timestamps, and browser and tab identifiers, to help track multitasking. Judges answered a few intermediate questions about the search sessions before identifying the reasons for engine switching in order to get acquainted with the searcher's intent and overall experience with the search engine. The intermediate questions included: determining the search goal of the user (based on the taxonomy of search goals proposed by Rose and Levinson [19]), whether the information need requires multiple sources to fulfill, overall success, pre-switch success, and post-switch success. To answer these questions, judges used the information about landing page content and how well it matched query terms as well as the actual sequence of query and click patterns on search results in a session. When determining the reason for the engine switch, the response options included: *Dissatisfaction, Coverage,* and *Other*. Note that we regard unintentional, user habit or user preferences (highlighted as being an important reason in Figure 1) as a subset of *Other* here, since judges may not be able to assess them from log data based solely on a single-session evidence.

Answer options for all questions were presented to judges as drop-down lists in the judgment spreadsheet. Space was also provided for additional comments, although this was seldom used in practice. Judges performed their judgments in isolation and then met to discuss and resolve inconsistencies.

7.4 Judge Agreement

Each judge assessed the same 100 search sessions and each session contained at least one *SQ* switch (that first of which was labeled). The Cohen's kappa (κ) between the two judges for judging switching reasons as *Dissatisfaction*, *Coverage*, and *Other* was 0.78 while κ between judges for *Dissatisfaction* and *Other*, where we merged *Coverage* and *Other* classes, was approximately 0.88. This signifies "substantial" judge agreement for tertiary labeling and "almost perfect" judge agreement for binary labeling [14].

7.5 Switch Causes for Same-Goal Switches

In the case of editorial assessments, we regarded switches with the same query as having the same search goal (given that the analysis in the previous section showed that *SQ* was a reasonable proxy for *SG*). As mentioned earlier, in this analysis we only focus on *Dissatisfaction*, *Coverage*, and *Other* (i.e., all reasons that are not dissatisfaction or coverage).

Overall, there was an 83% agreement between the reasons provided by the third-party judges and the switchers over the same 100 switches used in the analysis presented in this section. The typical disagreements lay in *Coverage* and *Other*, with it being most challenging to differentiate coverage and preference-based switches, which were present in *Other* but not explicitly labeled by judges.

There appears to be reasonable agreement between in-situ and editorial sources on the three main reasons for switching for the *SQ* switches. The editorially-assessed data has a similar distribution to the in-situ data: the percentage of switches associated with dissatisfaction is similar but slightly lower (45% vs. 49%) and the percentage of switches associated with coverage is similar but higher (33% vs. 21%). The differences in coverage estimates may be real, or given fewer assignable switching reasons (three as opposed to six in the in-situ survey), our judges may have overestimated the amount of switching associated with topic coverage.

7.6 Predicting Switching Causes

Prediction models were constructed using the data provided from the editorial assessment process. In Sections 7.6.1 and 7.6.2 we train *and* test on editorial judgments. However, in Section 7.6.3, we train on editorial judgments and test on in-situ judgments.

7.6.1 Binary and Tertiary Predictions

Baselines were updated to reflect the distributions in the editorial data set. Table 5 shows the average obtained $F_{0.5}$ values. The performance of the prediction model based on all interaction features is reasonable, outperforming both baselines for *DSAT* and *Coverage*, and performing marginally better for *Other*. Performance on predicting DSAT exceeds that of predicting coverage, but the observed gains over the baselines in both cases are lower than we observed for in-situ assessments. One reason for this difference is that less data were available for training and testing the predictive models (i.e., 100 editorial judgments vs. 354 in-situ judgments). To test the extent of this effect, we trained and tested predictive models on the same 100 in-situ switches (using ten-fold cross validation and 100 runs, as before), and observed small differences: the *DSAT* $F_{0.5}$ value with 100 in-situ judgments was 79.43 vs. 85.69 with 354 in-situ judgments. Another reason for the lower performance with editorial assessment could be noisy labels from third-party judges, who may incorrectly interpret logs and would make it challenging to associate reasons with actions.

Table 5. Binary prediction performance (measured via $F_{0.5}$) of *DSAT*, *Coverage*, and *Other*. Train on editorial, test on editorial. Symbol meaning same as Table 3.

Method	*DSAT*	*Coverage*	*Other*
Baseline (Prior)	50.41	35.54	30.16
Baseline (Rule)	54.88	41.90	30.98
All Features	66.16▲●	47.44▲	33.86

In addition to the binary classification, we again experimented with multi-class predictions, this time with the editorial data. The findings showed significant gains in $F_{0.5}$ over baselines (Logistic regression: 64.40, Prior: 50.10, Rule: 45.45) at $p < .01$, suggesting that tertiary prediction using editorial data may also be feasible.

7.6.2 Performance of Feature Groups and Features

In a similar way to Section 6.5, we now examine performance on predicting *DSAT* this time using editorial assessment data. Table 6 summarizes performance for all features, for only query features, and for only pre- and post-switch interaction features.

Table 6. Feature group performance (measured via $F_{0.5}$) for editorial assessment (*DSAT*).

Group	*All*	*Pre-switch*	*Query*	*Post-switch*
$F_{0.5}$	66.16	64.50	57.69	60.12

Once again, prediction performance is worst when using query-only and post-switch features, suggesting that they may be least useful for reliably predicting switch causes. We have already noted that post-switch behavior can be highly variable, making predictions based on it challenging. In addition, in post-assessment debriefings we discovered that judges generally ignored post-switch behavior; something that needs to be resolved in future studies, perhaps by modifying judge instructions.

7.6.3 Predicting In-Situ Judgments

In addition to using the editorial judgments to predict editorial judgments, we can also use the editorial judgments to predict the in-situ judgments, under the assumption that the in-situ judgments are the real ground truth. We re-ran our experiments, again using ten-fold cross validation over 100 runs, but instead of using the editorial judgments for the test fold, we used the associated in-situ judgments. The performance findings are reported in Table 7.

Table 7. Binary prediction performance (measured via $F_{0.5}$) of *DSAT*, *Coverage*, and *Other*. Train on editorial, test on in-situ. Symbol meaning same as Table 3.

Method	DSAT	Coverage	Other
Baseline (Prior)	47.83	32.00	25.05
Baseline (Rule)	50.65	37.17	27.71
All Features	62.76▲●	44.33▲	31.92

The findings show that the performance of the editorial judgments in predicting the in-situ judgments is slightly lower than that obtained when predicting editorial judgments. One explanation is differences in the criteria used by the judge and by the switcher; such inconsistency would lead to poor predictive performance on this task. More work is needed to understand these and other differences noted in this section, primarily because third-party labeling of switching episodes would likely be used in practice.

8. DISCUSSION AND IMPLICATIONS

We have presented an investigation into the causes of search engine switching and the automatic prediction of switching rationales as identified via in-situ and editorial assessments. The study has provided valuable insights into the reasons behind search engine switching and shown that we can predict the motivation behind engine switching with only limited interaction evidence. We found that a large percentage of *DQ* switches are unintentional or preference-related. Therefore, it could be misleading if these switches were included in performance-related analysis and metric derivation. As expected, search goals typically remain constant during *SQ* search engine switches, there are some behavioral patterns (especially in pre-switch behavior) that can reveal different motivations for *SQ* switching, and affording the accurate prediction of different switch rationales. Our findings provide better understanding of switching, and help search engines improve their user experience or derive more accurate competitive metrics.

The findings of the prediction experiments revealed that using the behaviors both before and after the switch lead to the most accurate predictive performance, with accuracy ranging from 65-85% depending on the source of the judgment data. The analysis also showed that the most predictive subset of features were those from interaction preceding the switch. This is promising for the development of real-time support for dissatisfied users, and is in line with the findings of Feild *et al.* [8], who demonstrated value in using recent behavior (including those associated with switching) to predict frustration. In future work, we will experiment further with query features such as type or aggregate clickthrough rate, which have been shown to be effective in predicting query performance and may also be useful in this context [10]. Given that dissatisfaction with the pre-switch engine was the dominant switching rationale, we will also explore the reasons that underlie user dissatisfaction (e.g., lack of diversity, obsolete results, etc.).

The participants in our study were all Microsoft employees. Although we showed some similarities between the switching behavior of these users and those reported in previous work [22], a larger deployment of *SwitchWatch* beyond our organization to a more diverse cohort is needed to further generalize our findings.

This work has a number of practical implications for the design of user-facing search technology. First, predictions of the reasons for switching can be used to dynamically adapt the search experience. Previous work has shown that we can accurately predict *when* a user will switch engine [12][15]. In this research we have shown that we can predict the *reason* behind such switching with good accuracy. Combining these methods would enable us to predict in real-time when a user is going to switch engines because they are dissatisfied. Over time, dissatisfaction-related switches can potentially erode user confidence in the search engine, and ultimately lead to permanent switching. Advanced warning of when switches are likely to occur enables search engines to intervene with an improved experience or offer new capabilities to candidate dissatisfied switchers. Such advanced capabilities include real-time chat with a domain expert or more powerful (but also more computationally-costly) search technologies. Conversely, in cases where the switch is detected by the post-switch engine, perhaps through a toolbar or inspecting the referrer URL, the engine could predict the reason for the switch to them based on post-switch behavior. For example, for incoming switchers dissatisfied with the pre-switch engine, more computational resources could be devoted to ranking. For incoming switchers seeking topic coverage or verification, emphasis could be put on search result diversity.

In addition, search engine companies extensively analyze log data to identify opportunities for improving their service. The ability to predict the reasons for switching allows search engines to compute more accurate competitive metrics that target for improvement of queries frequently leading to dissatisfied switching.

9. CONCLUSIONS AND FUTURE WORK

We have presented a study of the reasons that people switch search engines within sessions. We capture the reasons for switching and associated search behaviors *in-situ* and use the data to develop and evaluate models to automatically predict switching motivations using features of the switching queries and pre- and post-switch behavior. Our findings offer insight into searchers' decision-making processes and demonstrate the relationship between behaviors and switching causes. The findings also reveal sufficient consistency in search behaviors to afford accurate prediction of switching reasons. This could be useful for search providers to improve the search experience for users and derive more accurate competitive metrics. Future work involves studying the different types of dissatisfaction-related switching in more detail, exploring the use of additional features for predictions, creating refined metrics, and deploying switch rationale predictors on a search engine to tailor search experiences to switch rationales.

REFERENCES

[1] Aula, A., Khan, R. M., and Guan, Z. (2010). How does search behavior change as search becomes more difficult? *Proc. CHI*, 35-44.

[2] Bailey, P., White, R.W., Liu, H., and Kumaran, G. (2010). Mining past query trails to label long and rare search engine queries. *ACM TWEB*: 4(15).

[3] Bartell, B.T., Cottrell, G.W., and Belew, R.K. (1994). Automatic combination of multiple ranked retrieval systems. *Proc. SIGIR*, 173−181.

[4] Belkin, N., Cool, C., Croft, W.B., and Callan, J. (1993). The effect of multiple query representations on information retrieval system performance. *Proc. SIGIR*, 339−346.

[5] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

[6] Carmel, D., Yom-Tov, E., Darlow, A., and Pelleg, D. (2006). What makes a query difficult? *Proc. SIGIR*, 390−397.

[7] Cronen-Townsend, S., Zhou, Y. and Croft, W. B. (2002). Predicting query performance. *Proc. SIGIR*, 299−306.

[8] Feild, H., Allan, J., and Jones, R. (2010). Predicting searcher frustration. *Proc. SIGIR*, 34–41.

[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] Guo, Q., White, R.W., Dumais, S.T., Wang, J., and Anderson, B. (2010). Predicting query performance using query, result, and user interaction features. In *Proc. RIAO*.

[11] Hassan, A., Jones, R., and Klinkner, K.L. (2010). Beyond DCG: User behavior as a predictor of a successful search. *Proc. WSDM*, 221−230.

[12] Heath, A.P. and White, R.W. (2008). Defection detection: Predicting search engine switching. *Proc. WWW*, 1173−1174.

[13] Juan, Y.F. and Chang, C.C. (2005). An analysis of search engine switching behavior using click streams. *Proc. WWW*, 1050−1051.

[14] Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159−174.

[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] Mukhopadhyay, T., Rajan, U., and Telang, R. (2004). Competition between internet search engines. *Proc. HICSS*.

[18] Pew Internet and American Life Project. (2005). *Search Engine Users*. Accessed December 15, 2008.

[19] Rose, D. and Levinson, R. (2004). Understanding user goals in Web search. *Proc. WWW*, 13−19.

[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. and Dumais, S.T. (2009). Characterizing and predicting search engine switching behavior. *Proc. CIKM*, 87−96.

[23] White, R.W., Kapoor, A., and Dumais, S.T. (2010). Modeling long-term search engine usage. *Proc. UMAP*, 28−39.

[24] 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.

[25] Zhou, Y. and Croft, W.B. (2007). Query performance prediction in Web search environments. *Proc. SIGIR*, 543−550.