Defection Detection: Predicting Search Engine Switching

Allison P. Heath
Rice University
6100 Main Street
Houston, TX 77054
aheath@rice.edu

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

ABSTRACT
Searchers have a choice about which Web search engine they use when looking for information online. If they are unsuccessful on one engine, users may switch to a different engine to continue their search. By predicting when switches are likely to occur, the search experience can be modified to retain searchers or ensure a quality experience for incoming searchers. In this poster, we present research on a technique for predicting search engine switches. Our findings show that prediction is possible at a reasonable level of accuracy, particularly when personalization or user grouping is employed. These findings have implications for the design of applications to support more effective online searching.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *search process, selection process.*

General Terms
Measurement, Experimentation, Human Factors

Keywords
Search engine switching

1. INTRODUCTION
Commercial 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 is based on many factors including reputation, familiarity, retrieval effectiveness, and interface usability [4]. Similar factors can influence a user's decision to temporarily or permanently switch search engines (*e.g.*, change from Google to Live Search). Regardless of the motivation behind the switch, successfully predicting switches can increase search engine revenue through better user retention. Previous work on switching has sought to characterize the behavior with a view to developing metrics for competitive analysis of engines in terms of estimated user preference and user engagement [2]. Others have focused on building conceptual and economic models of search engine choice [3,5]. However, these studies did not address the important challenge of switch prediction. An ability to accurately predict when a user is going to switch allows the origin and destination search engines to act accordingly. The origin, or *pre-switch*, engine could offer users a new interface affordance (*e.g.*, sort search results based on different metadata), or search paradigm (*e.g.*, engage in an instant messaging conversation with a domain expert) to encourage them to stay. In contrast, the destination, or *post-switch*, engine could pre-fetch search results in anticipation of the incoming query. In this poster we take initial steps toward predicting whether a user will switch search engines, given their recent interaction history.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. DATA REPRESENTATION
2.1 Search Logs
We analyzed three months of interaction logs obtained from a large number of consenting users through an installed toolbar. All personally identifiable information was removed from the logs prior to analysis. From interaction logs we extracted *search sessions*. Every session began with a query to Google, Yahoo!, Live Search, Ask.com, or AltaVista, and contained either search engine result pages, visits to search engine homepages, or pages connected by a hyperlink trail to a search result page. 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.*, 6]. Users with less than five search sessions per month were removed to reduce potential bias from low numbers of observed interaction sequences or erroneous log entries.

2.2 Sequence Representation
We represent each session as a character sequence. This allows for easy manipulation and analysis, and also removes identifying information, protecting privacy without destroying the salient aspects of search behavior that are necessary for predictive analyses. Downey et al. [1] 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 (see Table 1). In a similar way to [1], 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. We define a search engine switch as one of three behaviors within a session: (i) issuing a query to a different search engine, (ii) navigating to the homepage of a different search engine, or (iii) querying for a different search engine name.

Table 1. Characters assigned to actions and pages visited.

Actions		Page visited	
Q	Query issued	R	First result page (short)
S	Clicked result link	D	First result page (medium)
C	Clicked non-result link	H	First result page (long)
N	Going back one page	I	Other result page (short)
G	Going back > one page	L	Other result page (medium)
V	Navigated to new page	K	Other result page (long)
Y	Switched search engine	P	Other page (short)
		E	Other page (medium)
		F	Other page (long)

For example, if a user issued a query, viewed the search result page for a short period of time, clicked on a result link, viewed the page for a short time, and then decided to switch search engines, the session would be represented as "QRSPY". We extracted many millions of such sequences from our interaction logs to use in the training and testing of our prediction algorithm.

3. OBSERVATIONS ON SWITCHING

In the duration of the study, approximately half of the users switched search engines at least once per month. Around 8% of search sessions contained a search engine switch, a proportion which could have profound financial implications given the large number of users involved.

In this analysis, differences were observed when comparing user interaction behaviors before switches to user actions before non-switches. Users seem less likely to click on non-result links before switching. This may indicate an unsuccessful search. Additionally, users who have recently switched are more likely to switch again. This behavior may signal an inability to find a piece of information or a desire to try multiple search engines for reasons of diversity or topic coverage. Such patterns suggest that prediction of switching may be possible.

4. PREDICTION OF SWITCHING

As we stated earlier, accurate real-time prediction could be used by Web search engine companies to improve the experience of users switching to or from their search engine.

4.1 An Online Classification Method

Here we present a straightforward classification method that provides insight into the difficulty of switch prediction and directions for improvement. Other methods (*e.g.*, hidden Markov models) may be useful for predicting switches. Investigation of those methods is an important topic for future research.

Our method takes a parameter n which represents the number of past characters considered when performing prediction. A three column table is constructed containing previously seen distinct strings of length n, the number of times the next action was a switch (a positive result), and the number of times the next action was not a switch (a negative result). To perform prediction, the table is searched for the n most recent characters of a session and the ratio between positive and negative results in the corresponding row is calculated. If the ratio is larger than a supplied parameter p, the method yields a positive prediction; otherwise the prediction is negative. If no corresponding row is found in the table, a new row is added and a negative prediction is issued. Additionally, as each new action arrives, the table is updated. This allows the method to adapt to previously unseen search behaviors.

4.2 Prediction Results

Figure 1 shows the precision/recall curve produced when our method, initialized using sessions from the first half of May 2007 with n=9 and varying p, was tested with sessions from the second half of May 2007. Our data for May 2007 contains about 1.5 million users and about 26 million search sessions. Other divisions of the data produced similar results. We obtain good precision when recall is less than 0.05. Since major Web search engines have high levels of traffic, significant numbers of correct positive predictions could be made despite the low recall.

The method can be improved by partitioning the users. In Figure 2 we show precision/recall results across all users for May 2007, where the method was applied on a per user basis with n=9 and p=0.5. Each dot in the figure represents a user. The figure shows that predictability of individual users varies; partitioning users in an intelligent way should help improve the switching prediction.

There are many other potential improvements. The very low number of switches compared to non-switches is an obstacle to accurate prediction. For example, in May 2007 the most common string preceding a switch occurred about 2.6 million times, but led to a switch only 14,187 times. This suggests a need to reformulate or augment the session encoding. More characters could be added to the alphabet, or continuous variables, such as time, could be used in conjunction with the encoding.

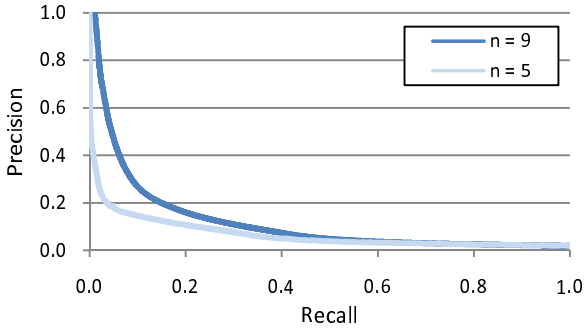

Figure 1. Precision/recall for May 2007.

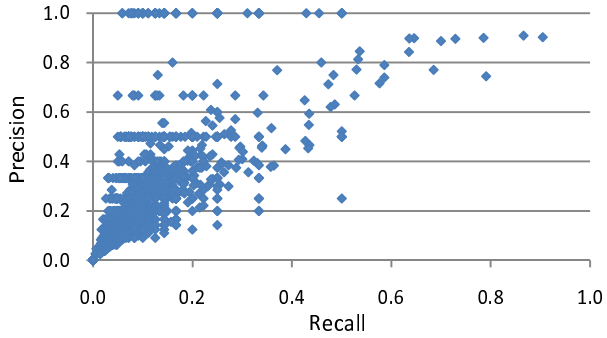

Figure 2. Precision/recall partitioning by user for May 2007.

5. CONCLUSIONS AND FUTURE WORK

In this work we analyzed several months of interaction logs to better understand the behavior of Web searchers. Based on this analysis, we took steps toward predicting when a user will switch search engines. We also discovered some important trends in search engine switching, but truly predictive patterns remain elusive and the topic deserves additional research. Future work should include ways to determine behaviors predictive of switching, improvement of the encoding, partitioning of users, and other classification methods.

6. REFERENCES

[1] Downey, D., Dumais, S. T. & Horvitz, E. (2007). Models of searching and browsing: Languages, studies and applications. In *Proc. IJCAI*, 2740-2747.

[2] Juan, Y.-F. & Chang, C.-C. (2005). An analysis of search engine switching behavior using click streams. In *Proc. WWW*, 1050-1051.

[3] Mukhopadhyay, T., Rajan, U. & Telang, R. (2004). Competition between internet search engines. In *Proc. 37th HICSS*, p. 80216a.

[4] Pew Internet & American Life Project. (2005). *Search Engine Users*. Accessed October 16, 2007. Available at: http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf

[5] Telang, R., Mukhopadhyay, T. & Wilcox, R. (1999). An empirical analysis of the antecedents of internet search engine choice. In *Proc. Workshop on Information Systems and Economics* (WISE, Charlotte NC).

[6] White, R.W. & Drucker, S.M. (2007). Investigating behavioral variability in web search. In *Proc. WWW*, 21-30.