An Analysis of Search Engine Switching Behavior using Click Streams

Yun-Fang Juan
Yahoo! Inc
701 First Ave
Sunnyvale CA 94089 USA
+1 408 349 6880
yunfangjuan@yahoo.com

Chi-Chao Chang
Yahoo! Inc
701 First Ave
Sunnyvale CA 94089 USA
+1 408 349 7611
chichao@yahoo-inc.com

ABSTRACT

In this paper, we propose a simple framework to characterize the switching behavior between search engines based on click streams. We segment users into a number of categories based on their search engine usage during two adjacent time periods and construct the transition probability matrix across these usage categories. The principal eigenvector of the transposed transition probability matrix represents the limiting probabilities, which are proportions of users in each usage category at steady state. We experiment with this framework using click streams focusing on two search engines: one with a large market share and the other with a small market share. The results offer interesting insights into search engine switching. The limiting probabilities provide empirical evidence that small engines can still retain its fair share of users over time.

Categories and Subject Descriptors
Search Engines, Data Mining

General Terms
Sequence, Session, Markov Chain, Principal Eigenvectors

Keywords
Search Engines, Clustering, Switching Behavior, Transition Probability Matrix, Limiting Probabilities

1. INTRODUCTION

Web search has become a very competitive field in recent years. With virtually zero switching cost and large revenue, web search engines are trying hard to expand their market share. Our research aims to characterize the web search competition with a set of metrics on *user share*, *user engagement* and *user preference*. We focus on *interaction* metrics: the main statement is about the probability of users switching from one engine to another over a specific time period. From there, we also paint a picture of the ultimate market share of search engines when the web search competition reaches equilibrium. Although the web search competition is highly non-stationary, these numbers offer a distilled view of the current competitive landscape and can be used as an objective to optimize.

We assume that quality of web search results affects people's choice of search engines. We identify the queries searched on both engines by the segment users who switch from one engine to the other as *potentially problematic queries.* The initial results have shown that these queries are indeed more problematic than queries searched on both engines by all the users.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. Framework

We partition click streams into *user sequences*. First, click streams are divided into individual sessions, each session being assigned a representative timestamp. Each session will then be characterized according to its usage across search engines and be assigned a label. After labeling, we specify two adjacent time periods t and $t+1$. We then construct two sequences of labeled sessions (S_t, S_{t+1}) for each user according to the session timestamps where S_t represents the sequence during time period t; S_{t+1} represents the sequence during time period $t+1$. These *user sequences* (S_t, S_{t+1}) are the input to our framework.

We estimate the transition probabilities P_{ij} from usage class i to usage class j from time period t to time period $t+1$. To define the usage classes, we segment the users by apply clustering procedure to the user sequences on both t and $t+1$ and find K clusters. The resulting clusters are interpretable and the clusters representing loyalists for individual search engines and switchers that frequently switch between search engines inter- and intra- search sessions are identified.

Each user will be assigned two cluster memberships $C_t = f(S_t)$ and $C_{t+1} = f(S_{t+1})$ where f is the model generated from the clustering procedure with C_t and $C_{t+1} \in \{1, 2,\ldots, K\}$. We construct the frequency table of the number of users who transition from class i to class j from t to $t+1$. Let $\boldsymbol{F_{ij}}$ denote the number of users who transition from class i to j from t to $t+1$. Let $\mathbf{P}$ denote the transition probability matrix and each element $\boldsymbol{P_{ij}}$ denote the conditional probability that a user will be in class j during time period $t+1$ given that she is in class i during time period t. That is,

$$P_{ij} \equiv \Pr(j \text{ at } t+1 \,|\, i \text{ at } t)$$

$\boldsymbol{P_{ij}}$ can be estimated as follows

$$\hat{P}_{ij} \equiv \frac{F_{ij}}{\sum_{j=1}^{K} F_{ij}}$$

$\mathbf{P}$ describes the search engine switching behavior, or *trend*, of the underlying population from time period t to time period $t+1$. From $\mathbf{P}$, we can make inferences about how loyal the users are with respect to individual search engines. We can also infer if a particular engine is losing users to another search engine.

We can also forecast the transition probabilities from time t to $t+s$ as $\mathbf{P}^s$. When s approaches infinity and assuming $\mathbf{P}$ is aperiodic, $\mathbf{P}^s$

will converge to $\mathbf{P^*}$ with all the rows equal to the vector of limiting probabilities. Let $\mathbf{\Pi^T} = (\pi_1, \pi_2, .. \pi_K)$ denote the vector of limiting probabilities where $\sum_i \pi_i = 1$. $\mathbf{\Pi}$ is the principal eigenvector of $\mathbf{P^T}$ since $\mathbf{P^T\Pi} = \mathbf{\Pi}$ with eigenvalue 1. The limiting probabilities are the ultimate user share at steady state assuming current trends hold.

3. Preliminary Results

The raw data used here is the complete click streams from an ISP. 12 weeks of data is used. The click stream is in the following format:

(USER ID , TIMESTAMP, URL VISITED)

The data is sessionized and the sessions with nonzero usage of either A or B are kept. Each search session is given a label as follows:

A: *if only engine A is searched in this session*

B: *if only engine B is searched in this session*

C: *if both engine A and B are searched in this session*

We select out the users who have at least 5 search sessions in the first six weeks and at least 5 search sessions in the second six weeks for our analysis.

We use the percentage of sessions on A, B and C as the clustering features. The clustering results of K-means are shown in Table 1. Note that the 5th to 7th columns are the centers of the clusters representing the mean percentage of sessions on A, B and C. The last column is our interpretation of each cluster by comparing the cluster centers and examining the members in each cluster. The cluster IDs are sorted by "%B".

ID	Limiting Probs	1st 6-week population	2nd 6-week population	% A	% B	% C	Cluster Interpretation
1	1.89%	2.14%	2.03%	95.00	2.71	2.29	A Loyalists
2	1.11%	1.08%	1.08%	67.86	23.21	8.93	A Primary
3	0.39%	0.27%	0.33%	30.45	29.00	40.55	Switcher I
4	1.31%	1.18%	1.19%	40.89	53.37	5.74	Switcher II
5	1.05%	0.67%	0.90%	12.04	62.17	25.79	B Primary
6	1.92%	1.72%	1.75%	20.00	76.68	3.32	B Principal
7	2.25%	1.63%	2.04%	2.36	81.52	16.13	B Principal using A as Backup
8	2.88%	2.73%	2.73%	9.13	89.72	1.15	B Loyalists checking out A
9	4.98%	3.92%	4.73%	0.45	92.38	7.18	B Loyalists using A as backup occasionally
10	82.22%	84.66%	83.21%	0.03	99.92	0.05	B Purists

Table 1: Cluster Memberships

The 10 clusters are further grouped into 3 main categories: *prime-A* (cluster 1 and 2), *prime-B* (cluster 5-10) and *switchers* (cluster 3, 4) to make the interpretation easier and construct the transition probability matrix as shown in Table 2. From Table 2, we can clearly see that non-switchers tend to stay in the same group between the two time periods and engine B has a much more cohesive user base than engine A. The *switchers* are more likely to switch to other groups and become either *prime-A* or *prime-B* users. User engagement and user preference can be inferred from the transition probability matrix and will be discussed in section 4.

From \ To	Prime-A (1,2)	Prime-B (5-10)	Switchers (3,4)
Prime-A (1,2)	78.67%	9.76%	11.57%
Prime-B (5-10)	0.27%	98.93%	0.80%
Switchers(3,4)	21.95%	51.17%	26.87%

Table 2: Transition Probability Matrix

Table 3 summarizes the mean number of sessions consumed during the 12-week period by each cluster of users. An interesting fact is that the off-diagonal cells suggest that people who transition from one group to another consume less sessions. It suggests if a search engine can make users search more, the chance of losing users to another search engine will be lower.

From \ To	Prime-A (1,2)	Prime-B (5-10)	Switchers (3,4)	Row Mean
Prime-A (1,2)	**26.66**	22.32	24.53	25.99
Prime-B (5-10)	21.98	**38.92**	23.90	38.75
Switchers (3,4)	24.26	25.03	**28.22**	25.72
Column Mean	26.03	38.75	25.16	

Table 3: Mean Number of Sessions

4. Key Metrics

User Share. We define user share of A during time *t* as the share of *prime-A* users during time *t.* From Table 1, the user share of A is 3.22% (cluster 1,2) during the 1st 6-week period whereas the user share of B(cluster 5-10) is 95.33%.

User Engagement. We define user engagement of A as the probability that users remain in *prime-A* during the second period. From Table 2, the engagement of A is 78.67% whereas the engagement of B is 98.93%.

User Preference. We define user preference of B to A as the odds ratio of switchers-to-prime-B over switchers-to-*prime-A*. We construe preferences as a choice made after an evaluation process. We consider that the *switchers* in the first period as users who are in the process of evaluation. From Table 2, the preference of B to A is 3.73, which means *switchers* are over three times more likely to prefer B to A.

Trends. We define trend of A as the current share of *prime-A* users and the share of *prime-A* users at steady state. From Table 1, the trend for A is -6.8%.

5. Conclusions

We present a simple framework to characterize the switching behavior between search engines based on click streams. Our findings indicate that such simple framework can generate insightful competitive metrics. The metrics about user preference and user engagement can be derived from the transition probability matrix. The user share describes the current market share and the limiting probabilities offer a distilled view of the current trend. We also infer that engines with small market share can retain its fair share since the limiting probability is non-zero and some users actually prefer small engines to big engines. Finally, user engagement, user preferences, market share and number of search sessions consumed are all positively correlated with one another. It provides empirical evidence that to increase market share, search engines should work to improve user engagement and preference scores.

A working version of the full paper is available at

http://www.geocities.com/yunfangjuan/www2005named.pdf