Characterizing Typical and Atypical User Sessions in Clickstreams

Narayanan Sadagopan
Yahoo!
2821 Mission College Blvd
Santa Clara, CA
narayans@yahoo-inc.com

Jie Li
Yahoo!
2821 Mission College Blvd
Santa Clara, CA
lijie@yahoo-inc.com

ABSTRACT

Millions of users retrieve information from the Internet using search engines. Mining these user sessions can provide valuable information about the quality of user experience and the perceived quality of search results. Often search engines rely on accurate estimates of Click Through Rate (CTR) to evaluate the quality of user experience. The vast heterogeneity in the user population and presence of automated software programs (bots) can result in high variance in the estimates of CTR. To improve the estimation accuracy of user experience metrics like CTR, we argue that it is important to identify typical and atypical user sessions in clickstreams. Our approach to identify these sessions is based on detecting outliers using Mahalanobis distance in the user session space. Our user session model incorporates several key clickstream characteristics including a novel conformance score obtained by Markov Chain analysis. Editorial results show that our approach of identifying typical and atypical sessions has a precision of about 89%. Filtering out these atypical sessions reduces the uncertainty (95% confidence interval) of the mean CTR by about 40%. These results demonstrate that our approach of identifying typical and atypical user sessions is extremely valuable for cleaning "noisy" user session data for increased accuracy in evaluating user experience.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Algorithms,Experimentation,Theory

Keywords

Web Search, Clickstream Analysis, Outlier Detection

1. INTRODUCTION

Search engines provide easy access to vast information resources available on the Internet. Their usage has been steadily increasing over the last few decades. This has also resulted in a burgeoning online advertising industry worth billions of dollars. Due to the revenue implications, it is important for these search engines to constantly improve the quality of user experience. Superior user experience will attract more users to their site and consequently lead to greater revenue. This has led to an increasing interest in mining user sessions to evaluate the user experience quality and incorporating the user feedback to improve the relevance of search results [11] [1]. Click Through Rate (ratio of clicks to pageview requests) computed from user sessions is one of the most frequently used metrics to evaluate the user experience. Hence an accurate estimation of CTR is crucial especially for comparing the fielded system to their beta counterparts [2].

Estimates of user experience metrics like CTR tend to have a high variance due to a user population that is extremely heterogenous along several dimensions such as demography, age, Internet familiarity, interests, etc [23]. Moreover Cove *et al.* show that even a single user adopts different interaction modes that include goal oriented search, general purpose browsing and random browsing [8]. Presence of bots adds another dimension of complexity to the estimation problem. Bots are automated software programs that issue queries to search engines while performing data mining tasks such as inferring index size, finding out the position of a particular ad (for Search Engine Optimization purposes), etc. Some of these programs can also spam the search engines by issuing several query requests or by producing excessive clicks (click fraud) [22].

Past studies have shown that bot detection is important for increasing the robustness of data mining techniques applied to web logs [19] [20]. In this study we argue that due to the vast heterogeneity in the user population it is equally important to identify typical and atypical user sessions. This identification can help in improving the estimation accuracy of user experience metrics like CTR. Due to an extremely heterogeneous user population it is not clear what constitutes typical or atypical user behavior. We adopt a simple approach of relating the rarity of a user session to the probability of its clickstream characteristics. Sessions with low rarity are considered typical while those with high rarity are considered atypical. We approximate the rarity of a session's clickstream characteristics using Mahalanobis distance. The higher the Mahalanobis distance, higher the rarity of the session. Our user session model incorporates several features including a novel conformance score obtained by Markov Chain analysis.

We illustrate the utility of our proposed approach by analyzing the sessions belonging to the tail 1% of the Mahalanobis distance distribution. The analysis reveals that these sessions indeed have rare clickstream behavior. Editorial evaluation shows that our approach of identifying typical and atypical sessions has a precision of about 89%. Moreover filtering out these atypical sessions reduces the uncertainty (95% confidence interval) of the mean CTR by 40%, indicating that these sessions indeed add "noise" to the CTR estimation. These results show that our approach of identifying typical and atypical user sessions is extremely valuable in cleaning "noisy" user session data.

The rest of the paper is organized as follows: section 2 discusses the related work. The clickstream data and modeling of user ses-

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

sions are described in section 3. Section 4 presents a detailed characterization of the atypical sessions in our clickstream data. Section 5 presents the evaluation of our approach while section 6 concludes the study and discusses future directions.

2. RELATED WORK

Over the last decade there has been a growing interest in mining World Wide Web (WWW) data. Kosala *et al.* classify these studies very broadly into Web Content Mining, Web Structure Mining and Web Usage Mining [14]. Web Content Mining focuses on analyzing the structure of individual documents on the WWW, while Web Structure Mining analyzes the link structure between the individual documents. Web Usage Mining studies the interaction of the user with the WWW by analyzing user sessions. These user sessions provide valuable information about user experience and perceived relevance of search results [1] [11].

User interaction can be influenced by a myriad of factors including web page content, the link structure of the web and user specific aspects (such as interests, ethnography and web familiarity) [10] [7]. Some of the approaches for modeling user specific aspects include browser usage studies [8] [6] and website navigation pattern studies. Our approach is similar to the latter and hence in the first part of this section, we discuss studies that analyze the navigation pattern of a user session. Several studies use Markov Chains to predict the next action in a sequence of user actions [17] [16]. Yates *et al.* model the interaction between the number of user clicks and the number of query formulations as a Markov Chain [4]. They also study the time distributions of the transitions between the states. Their analysis reveals several interesting aspects of user behavior: users tend to formulate short queries, click on few pages and majority of the users refine their initial query in order to retrieve relevant documents. Similar conclusions are also obtained in a study by Kammenhuber *et al.* that models user clicks and pages visited as a Markov Chain [13]. Borges *et al.* model user navigational patterns as an N-grammar in which the next page visited by a user depends on the previous N pages visited. They propose an efficient algorithm to mine preferred trails of a user that correspond to higher probability strings generated by the grammar [5].

All these studies show that web usage mining holds a lot of promise in understanding and personalizing user experience on the WWW. Menascé, Almeida *et al.* argue that understanding the nature of workloads is crucial for evaluating and improving the level of user experience [15] [3]. They use a hierarchical approach by analyzing the arrival process and usage statistics across different levels such as session level, application level and HTTP request level. Using this approach they characterize the workload at an online bookstore and an electronic auction site. Their study reveals the presence of bots in the workload. These bots are identified using some a priori rules of thumb such as access of the "robots.txt" file, session requests not following a logical sequence, etc. Several other studies use supervised machine learning to characterize bots [19] [20] [9] [18]. The above studies show that failure to detect web bots can significantly undermine attempts to develop models of user experience. Bots not only consume valuable bandwidth and web server resources but also decrease the robustness of applying Web Mining techniques on the Web logs.

In this study we argue that the vast heterogeneity in the user population makes it equally important to detect atypical user behavior to improve the effectiveness of web log mining. To characterize atypical and typical sessions, we relate the extent to which a user session is typical to the probability of its clickstream characteristics. One of the characteristics we use is a conformance score obtained by a "path analysis" of the user session using a Markov Chain model. Our analysis of the low probability atypical sessions reveals that these sessions indeed have rare clickstream behavior. Editorial assessments show that our approach identifies typical and atypical sessions with a preceision of about 89%. Our evaluation shows that filtering out these sessions reduces the uncertainty of mean CTR by about 40%, thus improving its estimation accuracy.

To our knowledge this is the first study that characterizes typical and atypical sessions in clickstream data and analyzes its effect on the estimation accuracy of user experience metrics. Our study is closely related (and can be used in conjunction) with Unexpected Browsing Behavior (UBB) mining [21]. While UBB mining is a supervised approach that uses a pattern matching algorithm to detect deviations from labeled examples of expected behavior, we adopt an unsupervised probabilistic approach for detecting atypical (or unexpected) sessions.

3. MODELING USER SESSIONS

As mentioned in section 1, there are several factors that contribute to the vast heterogeneity in user sessions. In this study our goal is to identify typical and atypical sessions. We use the following intuitive definitions:

1. Typical: The clickstream follows a logical sequence of events and there is no reason to think that it is abnormal, machine generated or abusive.
2. Atypical: The clickstream does not seem to follow a logical sequence of events. It is full of tasks that a normal user would rarely do. It appears to be mechanical, performing repetitive, nonsensical tasks.

Our approach towards this goal is based on detecting rare (outlier) user sessions. We first describe our clickstream data in section 3.1, which motivates our model for user sessions.

3.1 Clickstream Data

The clickstream data used in this study is a random sample of a single day's data obtained from the Search Results Page (SERP) of a major search engine. The SERP is a web page where a user can submit a query and interact with the returned search results. A user session is identified by a unique (user, query) pair. Our study is based on analyzing the aggregated data consisting of 2.4 million user sessions.

Each session consists of page requests and 0 or more clicks on search results. A page request occurs when a user submits a query to the search engine. If a user clicks on the back button in the browser window to review a page, then it is very likely that the browser renders the page from its cache. Thus this click does not generate a page request. However a click on "Reload/Refresh" button of the browser will lead to another page request for the same query. Once the search results are available, the user can click on "Web" (algorithmic) results, "Sponsored" (bidded) results, or "Next" (leading to visiting subsequent pages of search results). A user session consists of the following types of events:

1. Page request (denoted by P).
2. Click on the SERP. This click can be any of the following types:
   - Web click (denoted by W).
   - Sponsored click (denoted by O).
   - Next click to navigate through multiple pages of the SERP (denoted by N). This click can either be on the "Next" link or on the link for a specific page number displayed at the bottom of the SERP.

- Click that is not one of W, O or N. Such a click is called Any click (denoted by A). Some examples of Any clicks are clicks on “Search" button, “Also Try” (query refinement suggestion from the search engine), “Video" or “Images" tab (for obtaining video or image search results), etc.

Thus our clickstream data consists only of page requests and click events for the SERP. Unlike earlier studies, this data does not include clicks on pages referred by the search results and clicks on the browser interface [8] [6] [13]. Each event is also associated with a page number. Using this information, we define an Event-Locality Pair (ELP) as follows:

DEFINITION 1. ***Event-Locality Pair (ELP):*** *It is an ordered pair of (event, pageNumber) where*
event $\in \{P, W, O, N, A\}$ *and pageNumber* $\in \mathbb{N}$.

Consider the following example interaction listed in chronological order after the user enters “Flowers" in the search box & clicks the “Search" button:

1. A page request is sent to the search engine, leading to the SERP (with search results) being rendered in the user browser.

2. User clicks on one of the Web Results on page 1.

3. User clicks on the page 2 (next) link. This click also results in a page request for the 2’nd page of search results.

4. User clicks on a Sponsored Result on page 2.

The above interaction consists of the following sequence of ELPs: $(P, 1), (W, 1), (N, 1), (P, 2), (O, 2)$.

As a first step towards identifying rare user sessions, we seek to assign a score to each session based on its ELP sequence. It is desirable for this score to be higher for sessions whose ELP sequence is more conforming with normal (or popular) usage. To obtain such a score, we model a user session as a Markov Chain that is described in section 3.2.

3.2 Markov Chain Model

As shown in section 3.1, a user session can be viewed as a sequence of ELPs. It is reasonable to assume that within a session, the next event is most impacted by the previous event. This leads to a Markovian model for user sessions. The state space of the resulting Markov Chain model consists of every ELP that occurs in the clickstream data. Also we represent the start of a user session by the state ‘S’. Thus the state space of the Markov Chain is given by $\{S\} \cup \{\{P, W, N, O, A\} \times \mathbb{N}\}$. The transition probability $Pr(i, j)$ from state i to j is estimated as follows:

$$Pr(i, j) = \frac{Q_{i,j}}{Q_i} \quad (1)$$

where $Q_{i,j}$ = Number of instances where state i is followed by state j in the ELP sequences of all the user sessions[1]. $Q_i = \sum_j Q_{i,j}$. For the clickstream data used in our study, there are about 4300 possible transitions. Some of the transitions and their associated probabilities are illustrated in figure 1. Transitions with high probability can be associated with normal behavior, while transitions with low probability can be associated with rare behavior. For example, transition $S \rightarrow (P, 1)$ has a probability of almost 1 as it is normal for most user sessions to start on the first page of the SERP after submitting a query. Any other transition from state S has a negligibly small probability. Interestingly this observation indicates that not all user sessions start at the first page of the SERP. This abberation can be caused due to any of the following reasons: users accessing the SERP using cached copies (or bookmarks), user session straddling multiple days[2], bots, etc.

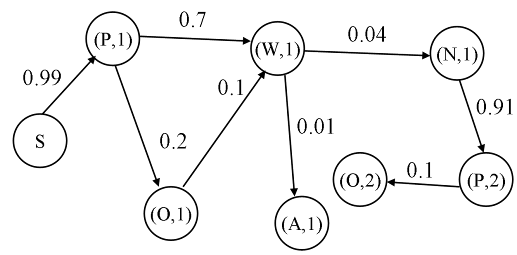

Figure 1: Markov Chain with some transitions and their associated probabilities. The probabilities are slightly modified for proprietary reasons.

After the transition table is computed from the clickstream data, each user session is assigned a likelihood score based on its ELP sequence. For example, consider the user session having the ELP sequence $(P, 1)$,$(W, 1)$,$(N, 1)$,$(P, 2)$,$(O, 2)$. The likelihood score (ϕ) for this session is computed as follows:

$$\begin{aligned}\phi &= Pr((P,1)|S) \times Pr((W,1)|(P,1)) \times Pr((N,1)|(W,1)) \times \\ &\quad Pr((P,2)|(N,1)) \times Pr((O,2)|(P,2)) \\ &= 0.99 \times 0.7 \times 0.04 \times 0.91 \times 0.1 = 0.00252\end{aligned}$$

Since the likelihood score is obtained by multiplying the probabilities of the individual state transitions, a user session with a longer ELP sequence will get a smaller score. Hence we take a log of the likelihood score and normalize it by the number of transitions (events) to obtain the average Markovian LoglikeHood (MLH$_{avg}$). In the above case, MLH$_{avg} = \frac{\ln(0.00252)}{5} = -1.2$.

One can think of MLH$_{avg}$ as a measure of conformance of the session’s ELP with normal (or popular) usage. A higher value indicates that the majority of transitions within the session are popular. A lower value indicates that the majority of transitions are rare.

For our clickstream data, MLH$_{avg} \in [-14, 0]$. 99.6% of the user sessions have their MLH$_{avg} \in [-2, 0]$. Moreover we also observe that about 99.8% user sessions with MLH$_{avg} \geq -2$ have $S \rightarrow (P, 1)$ as their first transition, while about 50% of the user sessions with MLH$_{avg} < -2$ **do not** have $S \rightarrow (P, 1)$ as their first transition.

MLH$_{avg}$ is a very important characteristic to determine whether a user session is typical or atypical. However sessions with similar values of MLH$_{avg}$ can have qualitatively different characteristics. For example consider the following two sessions from our clickstream data having a MLH$_{avg}$ value of around -0.8:

1. q_1: S, (P,1), (W,1), (N,1), (P,2), (N,2), (P,3), (W,3)

2. q_2: S, (P,1), (W,1)

[1] In our clickstream data, it is quite common for a single user session to contribute multiple instances of (i, j) transitions for a given pair i, j.

[2] The clickstream data used in this study consists of events from 12 am of one day to 12 am of the next day.

Session q_1 appears more typical than q_2 as the second session is repeating the pattern of (P,1), (W,1) 19 times. While the transition $(P,1) \rightarrow (W,1)$ is highly probable, the fact that it is repeated often seems to make session q_2 less typical. Thus in order to identify typical and atypical sessions, we need to incorporate other clickstream characteristics apart from MLH_{avg}. This motivates our multidimensional model for user sessions described in section 3.3

3.3 Multidimensional Session Model

The examples shown at the end of section 3.2 motivate the need for incorporating multiple clickstream characteristics into our model for user sessions. For a user session q let

- P_t: Number of Page Requests.
- W_t: Number of Web clicks.
- O_t: Number of Sponsored clicks.
- N_t: Number of Next clicks.
- A_t: Number of Any clicks.
- $E = P_t + W_t + O_t + N_t + A_t$: Total number of events.

We propose a 7 dimensional model in which we associate the *means to the end* i.e. the different types of events in the ELP sequence to the MLH_{avg} score. A user session q is represented as

$$q \equiv (\text{MLH}_{avg}, E, P_f, W_f, O_f, N_f, A_f)$$

where $P_f = \frac{P_t}{E}$, $W_f = \frac{W_t}{E}$, $O_f = \frac{O_t}{E}$, $N_f = \frac{N_t}{E}$ and $A_f = \frac{A_t}{E}$.

Thus in this model, the sessions q_1 and q_2 shown in section 3.2 are represented as $q_1 \equiv (-0.8, 7, \frac{3}{7}, \frac{2}{7}, 0, \frac{2}{7}, 0)$ and $q_2 \equiv (-0.8, 38, \frac{19}{38}, \frac{19}{38}, 0, 0, 0)$. We use this multidimensional model of user sessions to characterize typical and atypical sessions in section 4.

4. CHARACTERIZING OUTLIERS

As described in section 3, our definition of typical and atypical session is related to the rarity of a session's clickstream characteristics. For this purpose we approximate the probability measure around a point q (corresponding to a user session) in the 7D space by the Mahalanobis distance, which is defined as

$$d = \sqrt{(q-\mu)\Sigma^{-1}(q-\mu)^T} \quad (2)$$

where μ is the mean row vector and Σ is the covariance matrix for the clickstream characteristics in the 7D space. For a multivariate Gaussian distribution, the distance d is directly related to the probability density around point q. Higher the distance, lower the density and hence higher the rarity of q.

We apply the standard log transform technique to every dimension of point q so that the resulting data is closer to a multivariate Gaussian distribution[3]. We compare the quantiles of the following two distributions: the squared Mahalanobis distance computed from the clickstream data and χ^2_7 (the Chi Square distribution with 7 degrees of freedom) [4]. Figure 2 shows that the Q-Q (Quantile-Quantile) plot departs from the line $y = x$ indicating that the log transformed data does not exactly fit a multivariate Gaussian distribution. However the two sets of quantiles are strongly correlated

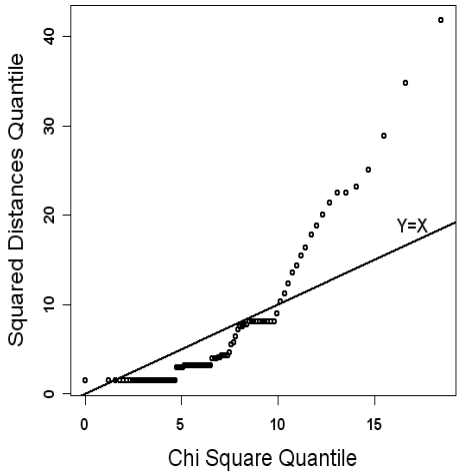

Figure 2: Quantile-Quantile plot of the squared Mahalanobis distance computed from the clickstream data and the χ^2_7 distribution.

with a correlation coefficient of 0.9. Thus it is reasonable to use the Mahalanobis distance as an approximate measure of the probability in the proposed 7D space of user sessions.

We consider the user sessions belonging to the tail $x\%$ of the Mahalanobis distance distribution as atypical. x is a parameter and can be varied. For illustrative purposes, we choose 2 values for x: 0.5 and 1. We conduct a detailed evaluation of the user sessions belonging to the tail 0.5% and 1% of the Mahalanobis distance distribution. Both the outlier categories exhibit similar behavior. Hence for the remaining part of the section, we focus most of the analysis on the 1% outlier set.

A detailed observation of the user sessions in the outlier set reveals the following 4 non-overlapping classes:

Page Requesters: These user sessions consist of only pageview requests, mostly for the same page. Their ELP sequence consists entirely of pairs of the form (P,i), where i is some page number. About 55% of the outlier sessions belong to this category.

As shown in figure 3, the outlier set exhibits a nice monotonic property beyond a certain value β of pageview requests i.e. if all user sessions with β pageview requests are considered as outliers, so are all sessions with greater than β pageview requests[5]. In this case $\beta = 2$ (for the 0.5% outlier set, $\beta = 3$). In addition to monotonicity, the outlier set also exhibits a sharp phase transition at the value $\beta = 2$ i.e. a very small percentage ($< 5\%$) of user sessions with less than 2 pageview requests are considered as outliers, while 100% of the user sessions with 2 or more pageview requests are considered as outliers. The user sessions that have less than 2 pageview requests and considered as outliers are the ones that do not have $S \rightarrow (P,1)$ as their first transition.

One plausible way of generating a clickstream consisting completely of pageview requests is by repeatedly clicking on the "Refresh" button of the browser (possibly due to general purpose browsing or random clicking). For Page Requesters in our clickstream data, $Pr(P_t \geq 2) = 0.05$ indicating that it is quite rare for users

[3]Prior to applying the transform, the 0 values are replaced by small positive values and the MLH_{avg} is converted to its absolute value.

[4]The squared Mahalanobis distance of a multivariate Gaussian distribution follows a χ^2 distribution with degrees of freedom equal to the dimensionality of the distribution [12].

[5]This is an artifact of using the Mahalanobis distance for identifying the outliers. In the case of Page Requesters, the user sessions have the same value for all coordinates except the number of events. If a session with β events is considered an outlier, so will a session with greater than β events. This is because the Mahalanobis distance for the latter session will be at least as large as the former.

to generate sessions with 2 or more pageview requests without clicks. We also observed that sessions with higher pageviews seem to generate these requests at regular intervals compared to those with lower pageviews. Figure 4 shows that for a given number of pageview requests, the standard deviation of the inter arrival times can range from being close to 0 to several hundred minutes. However as the pageview requests increase, the range of the standard deviations becomes narrower and closer to 0.

Next Clickers: These user sessions have clicks only on the Next Link. About 8% of the outlier sessions belong to this category.

As shown in figure 5, the outlier set exhibits a monotonic behavior beyond a value β similar to figure 3. Here $\beta = 10$ (for 0.5% outlier set, $\beta = 14$). To the left of β the behavior is non-monotonic. This is due to the difference among the sessions in the conformance of their ELP sequences with popular usage. For example consider 2 Next Clicker sessions q_1 and q_2 having 7 Next clicks given by the following ELP sequences:

- q_1: $(P,1)$,$(N,1)$,$(P,2)$,$(N,2)$,$(P,3)$,$(N,3)$,$(P,4)$,$(N,4)$, $(P,5)$, $(N,5)$, $(P,6)$, $(N,6)$, $(P,7)$, $(N,7)$,$(P,8)$.
- q_2: $(N,1)$,$(P,1)$,$(P,2)$,$(N,1)$,$(P,3)$,$(N,1)$,$(P,4)$,$(N,1)$, $(P,1)$,$(P,5)$,$(N,1)$,$(P,2)$,$(N,1)$,$(P,3)$,$(N,1)$,$(P,4)$,$(P,5)$.

The ELP sequence of q_1 is more conforming with popular usage (browsing the SERP in an increasing order of page numbers) and gets a higher MLH_{avg} score of -0.67 compared to q_2's score of -5.98. This results in session q_2 belonging to the set of outlier sessions.

Several users act as Next Clickers when they might not find an interesting result in response to their query. For the Next Clickers, $Pr(N_t > 10) = 0.045$ indicating that it is indeed very rare for a user session to consist of 10 or more clicks only on the Next Link. Moreover the clicks on the Next link seem to arrive at more regular intervals as compared to the page requests in the case of Page Requesters. Figure 6 shows that for a given number of Next clicks, the range of standard deviation is much lower than in the Page Requester case. Most of the user sessions have a standard deviation of less than 2 minutes and a handful of them have a standard deviation of above 20 minutes which indicates that the Next clicks arrive at reasonably regular intervals.

Repeated Web Result Clickers: These user sessions consist of multiple clicks on the same web result. 6.3% of the outlier sessions belong to this category.

For each of the user sessions in our clickstream data, we compute the Web Repeat Coefficient (w_{rc}) as a ratio of the number of web clicks to the number of web clicks on distinct results. Thus w_{rc} represents the average repeats per every distinct web result clicked. As shown in figure 7, the outlier set exhibits a monotonic behavior beyond $\beta = 25$ (for 0.5% outlier too, $\beta = 25$). However for $w_{rc} < 25$, the outliers exhibit a non monotonic behavior similar to the case of Next Clickers. As before this is due to the difference among the sessions in the conformance of their ELP sequences with popular usage.

In several cases we found that users click multiple times on the same web result when they use the search engine to reach a particular web site (such as "www.myspace.com") rather than directly typing the web site URL in the address bar of the browser. In our clickstream data $Pr(w_{rc} \geq 25) = 0.001$ indicating that it is very rare for a user to click 25 times or more on the same result. Also these clicks do not follow any specific arrival pattern. Figure 8 shows that for a given w_{rc}, the standard deviation can vary from 0 to a few hundreds minutes. Similar to the case of Page Requesters, as w_{rc} increases the range of the standard deviation gets smaller indicating that the clicks arrive at more regular intervals.

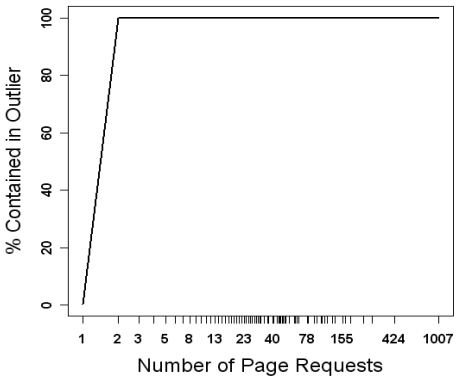

Figure 3: Page Requesters contained in the outliers.

Repeated Sponsored Result Clickers: These user sessions consist of multiple clicks on the same sponsored (ad) result. About 3.5% of the outlier sessions belong to this category.

For each user session we compute the Ad Repeat Coefficient (a_{rc}) as a ratio of the number of ad clicks to the number of clicks on distinct ads. As shown in figure 9, the outlier set exhibits a monotonic behavior beyond a value β (as for the other classes). Here $\beta = 7$ (for 0.5% outlier set, $\beta = 9$). While the behavior of the outlier set is very similar to the case of Page Requesters, the transition leading to β is smoother.

As in the case of web results, users tend to click on the same sponsored result multiple times (at different time instants). For our clickstream data $Pr(a_{rc} \geq 7) = 0.002$. Figure 10 is more sparse than figure 8 which indicates that the tendency to click repeatedly on a sponsored result is lower than that for a web result. For a given a_{rc}, the standard deviation can vary from 0 to a few hundred minutes. However as a_{rc} increases the variation in the standard deviation gets smaller.

The above 4 types of sessions account for about 73% of the atypical sessions. The remaining outlier user sessions fell into several small classes: sessions not starting on the first page, sessions that act as Next Clickers most of the time and click on a web result or sponsored result once in a while, sessions that click on all web results in sequence for the first few pages, sessions browsing and clicking beyond page 10, sessions with only 'A' (any) clicks, sessions only with 'O' (sponsored) clicks, etc. It is interesting to note that the outlier detection based on Mahalanobis distance in the proposed 7D space identifies several session classes that indeed have rare clickstream characteristics. In section 5, we use editorial judgements to validate our characterization of typical and atypical sessions and illustrate the importance of this characterization for computing user experience metrics like CTR.

5. EVALUATION

The lack of publicly available data sets of typical and atypical clickstream sessions led us to evaluate the precision of our approach using an editorial test. For this test the panel consisted of 3 unbiased human judges who had experience in analyzing clickstreams. Given the constraints on editorial resources we submitted 100 sessions (half of them sampled randomly from typical sessions and the other half sampled randomly from atypical sessions) for evaluation. The sessions consisted of the ELP sequence along with time stamps and the clicked urls. Our guidelines for judging typical and atypical sessions were exactly the same as the definitions given in section 3. While our approach of detecting typical and atypical

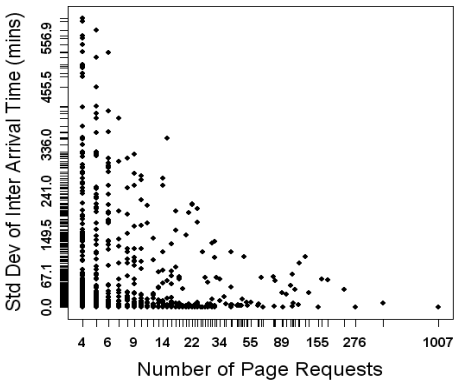

Figure 4: Standard deviation of the inter arrival times (of page requests) for the Page Requesters. The y-axis is plotted on a log scale.

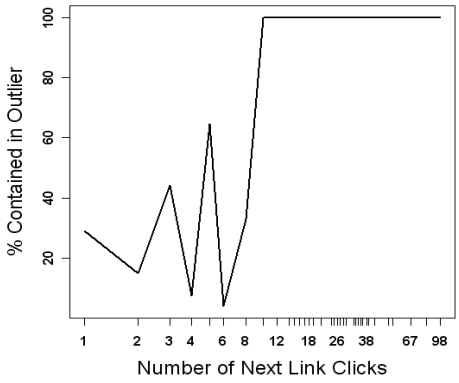

Figure 5: Next Clickers contained in the outliers.

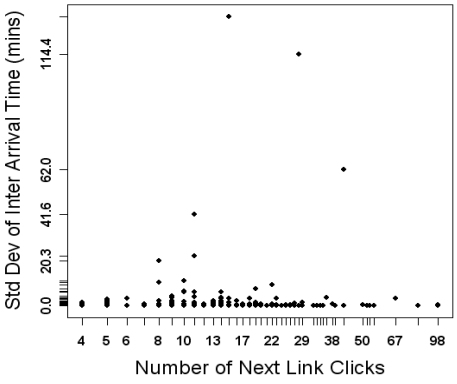

Figure 6: Standard deviation of the inter arrival times (of Next clicks) for the Next Clickers. The y-axis is plotted on a log scale.

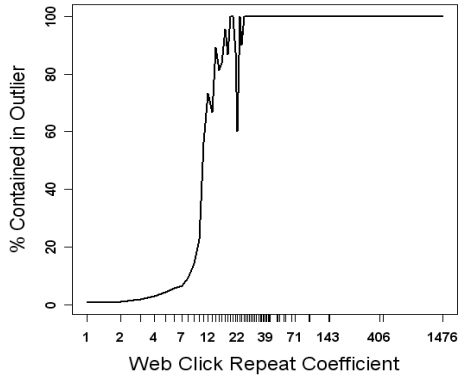

Figure 7: Repeated Web Result Clickers contained in the outliers.

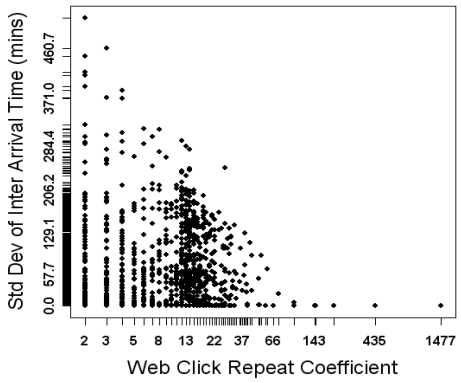

Figure 8: Standard deviation of the inter arrival times (of web results clicks) for the Repeated Web Result Clickers. The y-axis is plotted on a log scale.

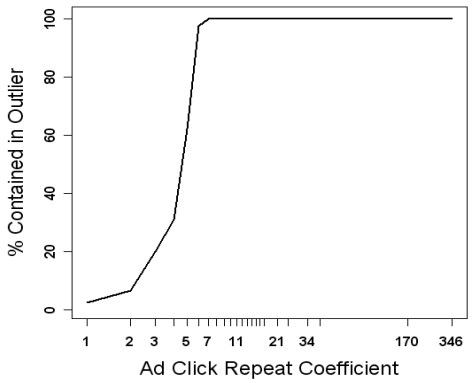

Figure 9: Repeated Sponsored Result Clickers contained in the outliers.

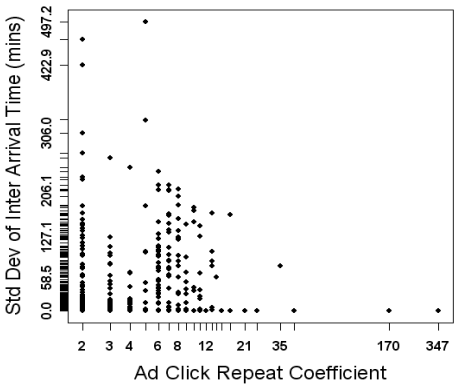

Figure 10: Standard deviation of the inter arrival times (of sponsored results clicks) for the Repeated Sponsored Result Clickers. The y-axis is plotted on a log scale.

sessions was oblivious of the semantics of the query and the clicked urls, the editors found this information to be useful in judging if the observed clickstream followed a logical sequence for a given query.

The editors' judgments were fairly consistent, all of them having the same judgement for about 72% of the sessions. As shown in figure 11, for 89% of the sessions our proposed approach of identifying typical and atypical sessions concurred with the majority of the editors. The precision for identifying typical sessions was about 88% while that for atypical sessions was 90%. The error percentage was quite small with 10% Type I and 12% Type II errors. We

	Typical (judged by majority of editors)	Atypical (judged by majority of editors)
Typical	44	6 (Type II Error)
Atypical	5 (Type I Error)	45

Figure 11: Validation of typical and atypical sessions using editorial judgements.

briefly discuss some representative examples of each type of error and the insights gained from editorial assessments.

1. Type I error: These are errors where our approach found a session to be atypical while the editors judged it as typical. 3 out of the 5 sessions under Type I error were the ones with several clicks only on sponsored results. A few representative examples in this category were "cruises from new orleans" $\equiv (-1.64, 12, 0, 0.92, 0, 0,$ $0.08)$ and "McFarlene toys" $\equiv (-1.25, 10, 0, 0.9, 0, 0, 0.1)$. Table 1 shows the detailed clickstream for query term "McFarlene toys". Most of the user sessions tend to have at least one click on a web result. However depending on the type of query, sponsored results can be more click-attractive than the web results. These queries usually are commercial in nature where a user is probably interested in buying the product. Since our approach did not account for query categories, it identified such sessions as atypical.

   The remaining 2 sessions belonging to Type I error were those that had only only 'A' (any) clicks. Examples in this category were "Jessica Alba" $\equiv (-1.07, 16, 0,$ $0, 0, 0.88, 0.12)$ and "America's next top model" $\equiv$ $(-1.06, 12, 0, 0, 0, 0.92, 0.08)$. For these queries, it appeared that the user was interested to see results of different modalities such as videos, images, etc. This led to multiple clicks only on the "Video" and "Images" tabs on the SERP. Table 2 shows the detailed clickstream for query term "America's next top model". As part of future work, it would be interesting to incorporate query categories (intent) into our approach for identifying atypical sessions.

2. Type II error: These are errors where our approach found a session to be typical while the editors judged it as atypical. 4 of 6 sessions under Type II error were those that had repeated clicks on the same location (or url). An example of a session belonging to this category was "www.t.o.k.music.com" $\equiv (-1.58, 16, 0, 0, 0, 0.44, 0.56)$. Subsequently we found that for "www.t.o.k.music.com", the search engine returned 0 results which possibly led the user to click multiple times on the "Search" button as shown in table 3. Another example of a session under Type II error was "work at home" $\equiv (-1.08, 30, 0.4, 0, 0.27, 0, 0.33)$. According to the editors, the only abnormality in the session was that it started on page 4 of the SERP.

Time (sec)	ELP	Clicked Url
0	(P,1)	-
6	(O,1)	toyrocket.com
161	(O,1)	store.yahoo.com
165	(O,1)	toyrocket.com
473	(O,1)	toywiz.com
650	(O,1)	store.yahoo.com
845	(O,1)	bigreds.com
898	(O,1)	iconusa4.com
967	(O,1)	davidsdepot.com
1000	(O,1)	hometeams.com

Table 1: Type I error example: commercial intent query ("McFarlene toys").

Time (sec)	ELP	Clicked Url
0	(P,1)	-
0	(A,1)	"Video" tab
1	(A,1)	"Video" tab
1	(A,1)	"Video" tab
2	(A,1)	"Video" tab
3	(A,1)	"Video" tab
50	(A,1)	"Images" tab
51	(A,1)	"Images" tab
52	(A,1)	"Images" tab
287	(A,1)	"Shopping" tab
288	(A,1)	"Shopping" tab
292	(A,1)	"Shopping" tab

Table 2: Type I error example: multi-modal intent query ("America's next top model").

The editorial assessments not only illustrate the effectiveness of our approach of identifying typical and atypical sessions but also indicate aspects of the methodology that can be improved. Incor-

Time (sec)	ELP	Clicked Url
0	(P,1)	-
0	(P,1)	-
637	(A,1)	“Search" button
638	(A,1)	“Search" button
638	(A,1)	“Search" button
638	(A,1)	“Search" button
638	(P,1)	-
638	(P,1)	-
638	(P,1)	-
639	(A,1)	“Search" button
639	(P,1)	-
639	(P,1)	-
641	(A,1)	“Search" button
641	(P,1)	-
643	(A,1)	“Search" button
643	(P,1)	-

Table 3: Type II error example: 0 result query (“www.t.o.k.music.com"). This query results had repeated clicks on the “Search" button.

porating query category (intent) is an interesting future direction we wish to explore. It is important to note that for illustrative purposes we identified the sessions belonging to the tail 1% of the Mahalanobis distance distribution as atypical. In general atypical sessions are those with rare clickstream characteristics. In this study, we used the Mahalanobis distance as an *approximation* of this rarity. Alternatively we could use non parametric outlier detection schemes to identify rare clickstream behavior. Such schemes could help in reducing Type II as well as Type I errors.

As mentioned in section 1, our motivation for identifying typical and atypical sessions is mainly to improve the robustness of data mining techniques applied to user sessions. As an example we show that filtering out the atypical sessions improves the estimation accuracy of CTR, an important metric for user experience evaluation. CTR is defined as the ratio of total clicks to total pageview requests. A high CTR corresponds to better user engagement. Quite commonly the estimation accuracy of a metric is measured by the width of the 95% confidence interval ($\approx 2\times$ sampling error). The smaller the width, the greater the accuracy of the estimation. To construct a sampling distribution of CTR we randomly assigned user sessions to N bins and computed the mean and the 95% confidence interval. Table 4 shows the mean (μ_1), confidence interval (CI_1) before filtering out the atypical sessions and mean (μ_2), confidence interval (CI_2) after filtering. These results show that filtering the atypical sessions leads to a substantial reduction (around 40% on average) in the uncertainty of the mean CTR. This increased estimation accuracy results in an increased sensitivity to small yet statistically significant changes in CTR, which is very important for an accurate evaluation of user experience. Interestingly the mean CTR remains unaffected indicating that the atypical sessions add “noise" to the CTR estimation.

Our evaluation shows that the identification of typical and atypical sessions using the Mahalanobis distance (as an approximate probability measure) in the 7D space of clickstream characteristics is extremely promising for cleaning the “noisy" user session data. We describe some interesting future directions in section 6.

N	μ_1	CI_1	μ_2	CI_2	% CI Diff
50	0.89	0.87-0.91	0.89	0.88-0.90	50
300	0.89	0.84-0.94	0.89	0.86-0.92	40
600	0.89	0.83-0.95	0.89	0.85-0.93	33.33
800	0.89	0.81-0.97	0.89	0.84-0.94	37.5
1000	0.89	0.80-0.98	0.89	0.84-0.94	44

Table 4: Comparison of the mean and 95% confidence intervals of CTR before and after filtering the atypical sessions. Total number of sessions before filtering is 2.4M and that after filtering is 2.38M.

6. CONCLUSIONS & FUTURE WORK

The last few decades have seen a substantial increase in the use of search engines for retrieving information from the Internet. This had led to an increasing interest in mining user sessions to evaluate the quality of user experience and use it to improve the relevance of search results. Quite often search engines rely on accurate estimates of CTR to evaluate the quality of user experience and compare fielded systems to their beta counterparts. However user experience metrics like CTR tend to have a high variance due to the vast heterogeneity in the population and presence of bots masquerading as genuine users. Past studies have shown that bot detection is important for increasing the robustness of data mining of web logs. In this study we advocate that due to a vastly heterogeneous user population it is equally important to identify typical and atypical sessions for improving the estimation accuracy of user experience metrics.

The vast heterogeneity in the user population makes it challenging to define a notion of what is typical and what is not. Our approach is based on relating the extent to which a user session is typical to the probability of the session's clickstream characteristics. Our session model used several features including a novel conformance score obtained by Markov Chain analysis. User sessions having high Mahalanobis distance in this multidimensional space are considered atypical. Our analysis showed that these sessions indeed exhibit rare clickstream behavior. Editorial results showed that our approach identified typical and atypical sessions with a precision of about 89%. The fact that these atypical sessions indeed contribute to “noise" is illustrated by our observation that filtering out these sessions reduces the uncertainty of the mean CTR by about 40%. These results show that our approach of identifying typical and atypical user sessions is extremely valuable in cleaning “noisy" user session data.

To our knowledge this is the first study that characterizes typical and atypical sessions in clickstream data and analyzes its effect on user experience metrics. With an appropriate definition of the state space for the Markov Chain, this approach can be extended to several other e-commerce web sites such as online bookstores and electronic auction sites. While we approximated the probability measure by the Mahalanobis distance (which relies on the assumption of multivariate normality), we would like to explore other measures that are non parametric. It would also be interesting to incorporate query categories (or intent) in our approach for detecting typical and atypical sessions.

7. ACKNOWLEDGEMENTS

We would like to thank Seokkyung Chung for his valuable feedback on earlier versions of this work. We would also like to thank Rajesh Shenoy and his team of editors for the careful human evaluation of typical and atypical sessions.

8. REFERENCES

[1] E. Agichtein, E. Brill, and S. T. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pages 19–26, 2006.

[2] K. Ali and M. Scarr. Robust methodologies for modeling web click distributions. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 511–520. ACM Press, 2007.

[3] V. Almeida, D. A. Menascé, R. H. Riedi, F. Peligrinelli, R. C. Fonseca, and W. M. Jr. Analyzing robot behavior in e-business sites. In *SIGMETRICS/Performance*, pages 338–339, 2001.

[4] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret. Modeling user search behavior. In *LA-WEB '05: Proceedings of the Third Latin American Web Congress*, page 242, 2005.

[5] J. Borges and M. Levene. Data mining of user navigation patterns. *Web Usage Analysis and User Profiling, Springer-Verlag as Lecture Notes in Computer Science*, 1836:92–111, 1999.

[6] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[7] L. Clark, I. Ting, C. Kimble, P. Wright, and D. Kudenko. Combining ethnographic and clickstream data to identify user Web browsing strategies, Information Research, 11(2) paper 249, 2006.

[8] J. F. Cove and B. C. Walsh. Online text retrieval via browsing. *Information Processing and Management*, 24(1):31–37, 1988.

[9] M. D. Dikaiakosa, A. Stassopoulou, and L. Papageorgioua. An investigation of webcrawler behavior: characterization and metrics. *Computer Communications*, 28(8):880–897, 2005.

[10] C. Holscher and G. Strube. Web search behavior of internet experts and newbies. In *Proceedings of the 9th international World Wide Web conference on Computer networks*, pages 337–346, 2000.

[11] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, 2005.

[12] R. A. Johnson and D. W. Wichern, editors. *Applied multivariate statistical analysis*. Prentice-Hall, Inc., 1988.

[13] N. Kammenhuber, J. Luxenburger, A. Feldmann, and G. Weikum. Web search clickstreams. In *Proceedings of the 6th ACM SIGCOMM on Internet measurement (IMC)*, pages 245–250, 2006.

[14] Kosala and Blockeel. Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, 2, 2000.

[15] D. A. Menascé, V. Almeida, R. H. Riedi, F. Ribeiro, R. C. Fonseca, and W. M. Jr. In search of invariants for e-business workloads. In *ACM Conference on Electronic Commerce*, pages 56–65, 2000.

[16] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty. Modeling online browsing and path analysis using clickstream data. In *Mining Business Databases*. Joint Statistical Meetings (JSM), 2003.

[17] R. R. Sarukkai. Link prediction and path analysis using markov chains. *Computer Networks*, 33:377–386, 2000.

[18] A. Stassopoulou and M. D. Dikaiakos. Crawler detection: A bayesian approach. In *International Conference on Internet Surveillance and Protection (ICISP)*, 2006.

[19] P. Tan and V. Kumar. Modeling of web robot navigational patterns. In *Proc. ACM WebKDD Workshop*, 2000.

[20] P. Tan and V. Kumar. Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, 6:9–35, 2002.

[21] I. Ting, C. Kimble, and D. Kudenko. UBB mining: Finding unexpected browsing behaviour in clickstream data to improve a web sites design. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 179–185, 2005.

[22] D. Vise. Clicking to steal. Washington Post Magazine, April 17 2005.

[23] H. Weinreich, H. Obendorf, and E. Herder. Data cleaning methods for client and proxy logs. In *WWW Workshop Proceedings: Logging Traces of Web Activity: The Mechanics of Data Collection*, 2006.