A Large-Scale Analysis of Query Logs for Assessing Personalization Opportunities

Steve Wedig
UCLA Computer Science Department
4732 Boelter Hall
Los Angeles, CA 90095, USA
swedig@cs.ucla.edu

Omid Madani
Yahoo! Research
3333 Empire Ave
Burbank, CA 91504, USA
madani@yahoo-inc.com

ABSTRACT

Query logs, the patterns of activity left by millions of users, contain a wealth of information that can be mined to aid personalization. We perform a large-scale study of Yahoo! search engine logs, tracking 1.35 million browser-cookies over a period of 6 months. We define metrics to address questions such as 1) How much history is available?, 2) How do users' topical interests vary, as reflected by their queries?, and 3) What can we learn from user clicks? We find that there is significantly more expected history for the user of a randomly picked query than for a randomly picked user. We show that users exhibit consistent topical interests that vary between users. We also see that user clicks indicate a variety of special interests. Our findings shed light on user activity and can inform future personalization efforts.

Categories and Subject Descriptors: H.4 [Information Systems Applications]: Miscellaneous

General Terms: Algorithms, Experimentation.

Keywords: Personalization, Query Logs, User History, User Interests, Categorization, Clustering.

1. INTRODUCTION

Personalization holds much promise for dramatically increasing the impact of online services such as web search. Query logs, the patterns of activity left by millions of users, contain a wealth of information that can be mined to aid personalization. We perform a large-scale analysis of Yahoo! search engine logs, tracking 1.35 million browser-cookies over a period of 6 months. We define and track measures to address three types of questions: 1) The extent of short and long term history available, 2) Consistency and convergence rate of users' general topical interests as reflected by their queries, and 3) Finer grain information available on user interests derived from users' clicks on search results. We find that there is significantly more expected history for the user of a random query than for a random user. We show that users exhibit consistent topical interests that vary between users. We also see that user clicks can reveal users' special interests. Such information makes personalized applications possible, including web search, targeted advertising, and recommendations. We briefly describe an important application next: web search.

Interacting with search engines has traditionally been an impersonal affair, with the returned results a function only of the query entered. Unfortunately the average query length is consistently reported to be around two, so many queries are too short to disambiguate the user's information need. Moreover, users often view only the first page of results, which makes precision critically important. These limitations have motivated researchers to look beyond the query and consider how a search's context can provide further evidence about the user's information need. A broader context could include features such as the user's geographic location, the time and date, and the user's previous interactions with the search engine [9]. A user's interactions with a search engine consist of performing searches and clicking on result pages. A user's prior actions make up her *history*. We define *contextualization* as integrating a user's history into the results ranking. We break contextualization into two types: 1) *personalization* is considering a user's long-term interests, and 2) *adjustment* is reacting to the user's short-term action history. Personalization and adjustment are complementary approaches to integrating user history. With adjustment, a search engine could quickly react to a user's actions. This would require very little prior history, perhaps starting as soon as the second query. As a polysemous word, "jaguar" is an inherently ambiguous query. In an effort to cater to all users, major search engines distribute their top 10 results among the various meanings (car, cat, football, osx, etc). More relevant results could be provided on the first page if such a query was disambiguated. Adjustment could make this possible. For example, "jaguar" is no longer ambiguous when immediately preceded by the queries "bmw" and "mercedes". In contrast, personalization requires having some length of history available. However integrating long-term interests could address the "cold-start" situation when a user searches after a period of inactivity, or begins a new search need. For example, we suspect our readers might appreciate it if results from Citeseer's domain tended to be ranked higher for them. Since web search sessions are typically short, a significant portion of queries will fall into this cold-start situation.

1.1 Related Work

There have been numerous analyses of query logs [1, 5, 15, 6]. Much has been published with relation to the distribution of queries according to aspects such as query length and query frequency [13, 5, 6], and query type and topicality [2, 15]. Other work has focused on user behavior at the query session level, showing aspects such as reformulation rates can be relatively high [16]. Our study complements previous research by focusing on users that issue the queries

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.
Copyright 2006 ACM 1-59593-339-5/06/0008 ...$5.00.

	Complete Sample	High Activity	Ratio
# Cookies	1,377,271	113,324	0.08
# Searches	26,468,452	14,723,431	0.5
# Clicks	20,231,315	11,531,001	0.5
Avg Search / Cookie	19	130	6.7
Avg Click / Cookie	15	102	6.9

Figure 1: Query Log Sample Statistics.

over a long period of time. For the purposes of exploring personalization (or more generally contextualization) opportunities, we group the queries by users that issue them. There is also much work on personalization in general, as well as personalization to aid web search in particular. To achieve personalization user profiles are created from explicit participation and/or implicit (behavioral) user feedback [10, 4, 3, 8, 14], and researchers have identified various distinctions such as short versus long term profiles, incorporating context, and different types of profiles such as content-based versus collaborative information [7, 9]. The use of learning from impressions and clicks to improve the ranking is also an exciting direction [10, 11, 9]. Our present study evaluates personalization opportunities via users' queries and clicks (implicit feedback) and should inform future personalization research.

2. DATA

Our data source was six months of query logs from the Yahoo! search engine. We tracked two types of actions: searches and clicks. A search is followed by zero or more clicks on results pages. All actions record the timestamp, browser cookie, IP address, and the associated query. If a user was logged in, then a unique identifier corresponding to their Yahoo! account was also recorded. Click actions also store the destination url, and that url's ranking within the search results

User tracking lays the foundation for contextualization. We must be able to record the user's actions. Websites generally have three identifiers available for tracking users: IP address, browser cookie, and login account. Each of these tracking methods represents a tradeoff between coverage, accuracy, and persistence. A complete personalization system could potentially use a combination of these tracking methods. In this study we track users with browser cookies. Cookies seem to provide the most appealing tradeoff between coverage and accuracy. Persistence is a concern for long-term personalization, and we further examine this in the next section. Tracking by cookies also offers the best chance that this study's results will be generally applicable. The cookie behavior of Yahoo users should reflect what many websites experience. This would not be true for user accounts.

Our data was a random sample of the cookies active on the Yahoo! search engine over a period of six months. We used our original sample and a subsample of it in this study. The original sample contained about 1.35 million cookies, 26 million searches, and 20 million clicks. Some of our tests required considering only users with a minimal amount of action history. For these tests, we defined a high activity user subsample, which contains only the users who submitted 20 or more searches, and whose activity ranged over 30 or more days. There were 235,183 cookies with more than 20 searches, and 190,445 cookies activity spanning 30 days. The intersection formed our high activity sample, containing about 115,000 cookies, 14 million searches, and 11 million clicks. Figure 1 compares the samples. The 8% of cookies in the high activity sample account for 50% of all user activity.

3. MEASURING HISTORY

Contextualization requires a record of the actions a user has performed. Within search engines this is a log of queries and clicks. We examine the amount of available user history in three ways. First we measure cookie persistence to give us an idea of how much time history is available. Second, we look at the user activity distribution, observing how much action history can be expected. Third, we consider the interplay between time and actions. We look at how much action history is available within fixed time windows.

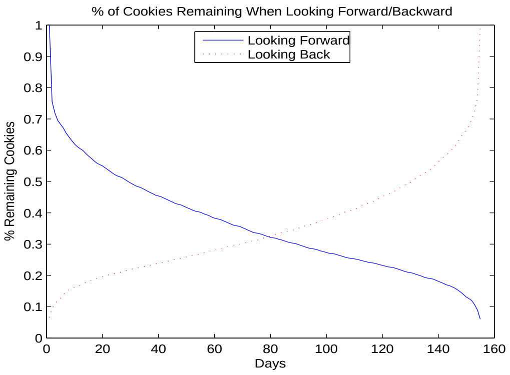

Figure 2: Cookie Persistence.

Percent of Users With $\geq$ N Days	Looking Back N Days
0.1	154
0.2	33
0.3	85
0.4	49
0.5	25
0.6	11
0.7	4
0.8	1
0.9	1
1.0	1

Figure 3: % of Users Active When Looking Back N Days.

3.1 User Persistence

Having chosen cookies as our method for tracking users, our first concern is cookie expiration. The amount of historical data we can collect is proportional to cookie persistence. For example, if 75% of cookies expire after one day, then adjustment will be the only contextualization option for a majority of users. To measure cookie persistence over time we look at how many days a cookie remains active. A cookie is active between its first and last observed actions. For the cookies active on the first day of our sample, we measure how long they remained active over the following six months. We also did the equivalent test for the set of cookies active on the last day, looking backwards in time. There were 33,836 cookies active on the first day, and 31,048 on the last.

Figure 2 plots the days in our sample on the x-axis, and the y-axis shows the percent of active users remaining. The first thing to notice is the steep initial drop of about 30% after the first day. After this the curve quickly becomes a linear decay function. Note that the curves artificially converge to zero at the ends. This is due to the finite nature of our sample. Most of those users were probably still active, but were cut off by the end of the time frame. We expect the linear attrition rate would continue without this horizon limitation. While looking forward gives us a feeling for cookie persistence, it isn't the view available during contextualization. Instead a system has to react at query time, looking backwards through the user's history. The figure shows that the forward and backward cookie

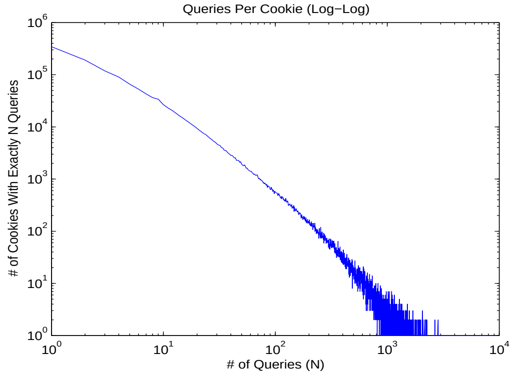

Figure 4: Queries Per Cookie.

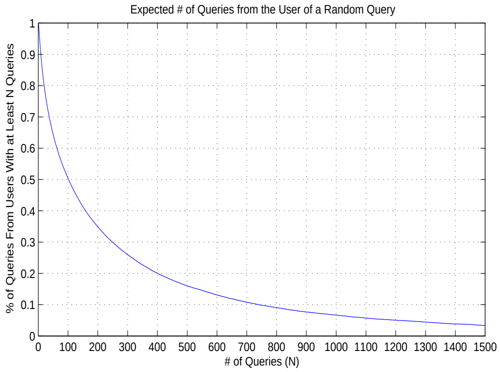

Figure 5: Queries Per Query.

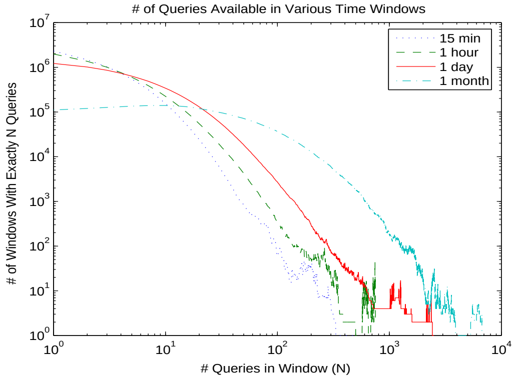

Figure 6: Number of Queries Contained in Time windows.

decay functions are symmetric. Figure 3 shows the percent of users that will still be active if we look back N days. For example, of the users active on the last day, about 40% of cookies were at least a month old. This diagram shows that a notable portion of the cookies persist over significant periods of time. Although 30% expire after the first day, over 40% of cookies persist for at least a month.

3.2 User Activity

We have seen that some cookies persist over months. However, time alone cannot help our contextualization efforts. Instead we must learn from the user's actions: searches and clicks. We know there is an average of 19 searches per cookie. In this section we take a more granular look at the distribution of how much history users accumulate before their cookie disappears. Since there is a fairly consistent ratio between searches and clicks, we only refer to searches here. The results will apply equally to both. Figure 4 is a plot showing the number of searches N on the x-axis, and the number of users who performed exactly N searches on the y-axis. It is a log-log plot, so the curve fits a power law well. This means that a significant majority of users perform very few searches over the lifetime of a cookie. In fact, 25% of observed cookies only had one search, 59% had five or less, 82% did fewer than the average of 19 searches, while 17% did more. How will this power-law activity distribution impact our personalization aspirations? The situation may seem grim at first. It is probably safe to say that five queries over six months are insufficient for personalization. Alternatively, we could ask what percent of searches come from users with sufficient history. In other words, what percent of the query stream can we personalize?

Figure 5 again shows the number of searches N on the x-axis, this time with the y-axis showing the percent of queries coming from users who performed *at least* N searches. This is like picking a random search from the data sample and finding the probability that the search's user performed at least N searches overall. For example, we can see that about 50% of the query stream comes from users who performed at least 100 queries over the 6 month period (see also Figure 3). The amount of user history looks much more promising from this query stream perspective. Figure 4 equally weighs every cookie, where Figure 5 equally weighs every search. We believe the second view is more important for evaluating personalization potential. Due to the power-law distribution, 50% of the searches come from only 3.5% of observed cookies. Most cookies don't come back. Either the users were fleeting visitors, or their cookie was invalidated for some reason.

3.3 History at Query Time

Recent history is used for adjustment, and long-term history is needed for personalization. In this section we look at how much history is available at query time, considering time windows of various sizes. As in the previous section, we will only discuss searches; the results apply equally to clicks. Assume you used a search engine once on a particular day. Over a period of 10 minutes you submitted five searches. For each search submission, a contextualization system would look back in time at your prior actions. Assume the system only considers the last 30 minutes for adjustment. Then there were five 30 minute time windows, each ending at the time of a search. The window for the first search contained zero prior searches, and the last window contained four. On that day you had an average of three searches per 30 minute time window. We measured the activity distribution within time windows of varying sizes: 15 minutes, one hour, one day, and one month. Naturally the shorter windows are more relevant for adjustment, and the one month window is more relevant for personalization. We ran these tests on our high activity user sample. Each search has four associated time windows: one of each size. To collect the data, we skipped the first month of searches. This was necessary because at the start of our sample time period we don't have a month's worth of historical data. For the remaining 5 months, we recorded the query distributions for every time window.

Figure 6 is a log-log plot with the x-axis showing the N, the number of searches, and the number of windows containing exactly N searches. The window sizes shown are 15 min, one hour, one day, and one month. You can see that it fits a power-law distribution reasonably well. This means that most sessions contain few queries. This is what we might expect. The windows with very high query counts were probably the result of automated searches.

Figure 8 shows a cumulative view of this data. It shows the number of searches (N) on the x-axis, and the y-axis shows the percent of windows containing at least N searches. For example, we can see that the one month time window of 35% of searches will contain 10 or more earlier searches. Figure 7 displays this data in tabular form. For active cookies, sessions appear to be long enough for adjustment to be a viable goal.

4. QUERY TOPICS

The idea of personalization is grounded on two related assumptions about user behavior. Our first assumption is that users have reasonably consistent interests. A user's history will only be useful if her previous actions help us predict her future interests. This will be true if her interests are consistent over time. Our second assumption is that users have different interests from one another. After all, if everyone has the same interests there is no need for personalization! We explore these themes in terms of the general topics of the queries users submit. Note that knowing the range of topics that a user is primarily interested can have a number of personalization oriented applications, such as reordering returned pages as well as displaying ads of interest [9]. We employ a query categorizer to determine the topic(s) of each query. Based on query topics, we estimate an interest distribution for each user. We examine whether and how fast individuals' interest distributions converges. We find positive evidence for convergence. We next check whether interest distributions are different across users (distinctness).

4.1 Interest Distributions

We used 22 general topical categories ("Travel", "Computing",..), and utilized an automated query categorizer obtained via machine learning techniques [17, 12]. The categorizer assigned a confidence probability to each category it suggested. It dropped categories with less than 10% confidence. The categorizer is imperfect: it does not have complete coverage. Furthermore, its coverage and errors for various topics could be different. A user's interests are reflected by her queries over time. The categorizer provides a probability vector (category probability assignment) for each individual query. We summed the probability vectors of all the queries of a user, and l_1 normalized the total vector (*i.e.,* $\frac{x}{|x|_1}$, where $|x|_1 = \sum_i |x_i|$) to obtain the *final* distribution or *interest distribution* F for each user. This vector, if computed over sufficiently large number of queries, is a reflection of the proportion of time the user spends querying on each topic, or the stationary distribution of the user's topical interests. However a user's interests may drift and she may not have a stationary distribution. The population's global (interest) distribution G is the normalized sum over all queries from all users in the sample[1]. In essence, this represents the interest distribution of the entire population. Some example topics with their probabilities in G are: *(Travel,0.116)*, *(Toys and Hobbies,0.073)*, *(Health and Beauty, 0.068)*, and *(Automotive, 0.066)*.

4.2 Consistency and Distinctness

Common distance measures for distributions include KL-Divergence and those based on l_1 and l_∞ ($|x|_\infty = \max_i |x_i|$, and $d_p(x, y) = |x - y|_p$). KL-Divergence is not as easily interpretable, and punishes severely (assigns large distance) for near zero probabilities, thus we opted for the d_p distances. The d_∞ distance amounts to taking the maximum difference over the dimensions (the 22 categories), while d_1 distance amounts to taking the sum of the absolute differences over the dimensions.

If a user's queries were drawn from a fixed distribution, the observed distribution will eventually converge to this underlying distribution. In Figures 9 and 10 we plotted distribution distances to check whether this convergence occurs. The distances for two users, one with approximately 50 and another with 200 queries are shown. The Y axis shows the distribution distances (both d_1 and d_∞). One curve shows the distance between the user's *current cumulative* distribution (C) and the user's final distribution (F). The other curve shows the distance between C and the population's global (interest) distribution G. The initial drops in both distances is due to the fact that with small query samples the probabilities are skewed. If a user has interests different from TD, we should see a rapid stabilization of the distance of C from TD, and this appears to be the case for user 2. C equals F after all queries of the user have been observed (C is guaranteed to converge to F). However, if we see a relatively rapid decrease in $d_p(C, F)$ to close to 0, then that's evidence that the user has a stationary distribution and that F (for the queries we have computed) is close to it. This may be the case for user 2. We saw a similar pattern for several users' convergence profiles that we visually examined: below approximately 100 queries, there was no visual evidence of either hypothesis (difference from G or stationary F), but with more than 100 queries, there is some evidence.

Another way to test consistency involves chronologically splitting the user's queries into a first half and a second. If a user's interest distribution is consistent in time, then we expect the sets to have similar category distributions. The halves should be more similar than the halves of different users. A consistent user's two halves would also be closer to each other than to the global distribution G. Figure 11 shows the distances as a function of subpopulation of users with a threshold on number of queries. We observe that while both distance to G and between halves decrease, the difference and the ratio of the two increases significantly as we restrict to population of users with more queries. This chart presents compelling

	15 min	1 hour	1 day	1 month
% of windows	# queries			
1.0	1	1	1	1
0.9	1	1	1	10
0.8	1	1	2	19
0.7	1	2	3	30
0.6	2	2	5	44
0.5	2	3	7	61
0.4	3	4	9	85
0.3	4	6	13	118
0.2	6	9	19	171
0.1	9	14	30	285

Figure 7: % of Time Windows Containing at Least N Queries.

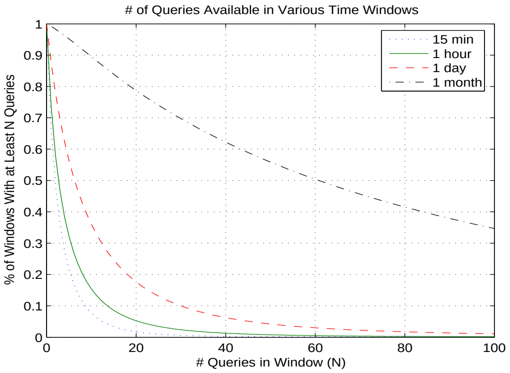

Figure 8: % of Windows Containing at Least N Queries.

[1] In computing G each query gets an equal weight. We also looked at the sum where each user gets equal weight, and the results were nearly identical.

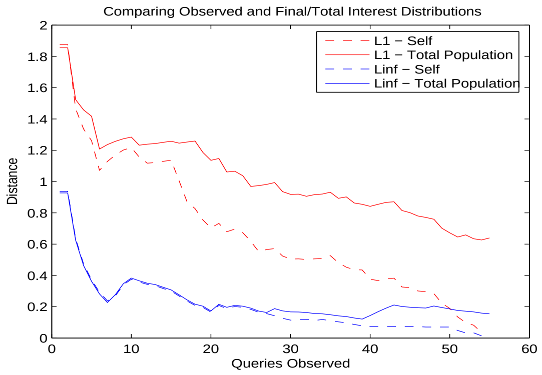

Figure 9: Evolution of the distance of the cumulative distribution for User 1 to her final distribution F and to the global distribution G. There is little sign that the user has a consistent distribution and that it is different from G. User 1 has just over 50 queries.

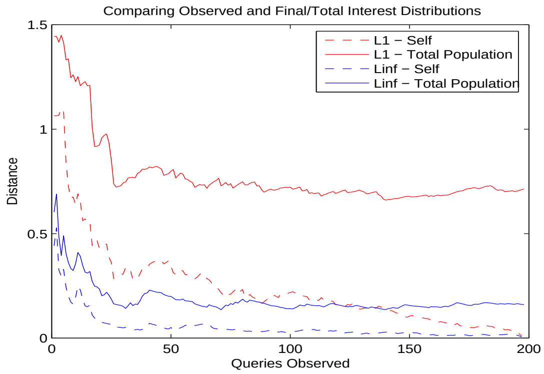

Figure 10: Evolution for User 2 with 200 queries. There is strong evidence that the user has a different interest distribution from G.

evidence that users tend to be consistent (have a stationary interest distribution), that with 100s of queries, the F we obtain is close to the user's stationary distribution, and that such F is different from G.

The goal of personalization is to provide different content to users with different interests. The value of such personalization is limited by how different users are from each other. As users appear different from global distribution G, that is good evidence that users must be different from each other (there are at least two clusters of users). We measured d_1 and d_∞ between the first-half of a pair of randomly picked users. For the populations that we computed such distances, the distances were significantly higher on average than the distance between the two halves of the same user from the same population. For example, for pairs of users with at least 80 queries each, the average d_1 and d_∞ were respectively 1.26 and 0.32 (compare with $d_1^s = 0.74$ and $d_\infty^s = 0.177$ for users with 75 queries in table 4.2), and for users with at least 500 queries, the averages were 1.076 and 0.244 (compare with $d_1^s = 0.492$ and $d_\infty^s = 0.114$ for users with 400 queries in table 4.2).

Let us briefly and informally discuss the effect of error of the query categorizer on our results. There are three type of errors that the categorizer can make: random, systematic, and malicious. Random errors include cases when an arbitrary wrong category is picked for a given query. Systematic error includes cases in which some categories may have low coverage or recall relative to others, or when a category is a false positive disproportionately higher than other categories (wrong queries are disproportionately assigned to such a category). Malicious error includes the case in which the categorizer knows what we are computing and systematically misclassifies some queries so that the errors are in a direction that we cannot anticipate. Malicious error is not possible for our categorizer. We remark that both random and systematic error tend to make users look consistent but also more like one another. Since we've seen evidence that users' interests are different from G and one another, we believe our conclusion that users tend to be consistent is sound as well.

5. PAGE CLICKS

Are some clicks more informative than others? Intuitively this appears to be the case. For example, a click on acm.org seems to tell us much more about the user's interests than a click on yahoo.com. It would be useful to have a metric quantifying how informative a click is. There is an abundance of context for every click. In this section we abstract away most contextual details and only consider the user's identity (cookie) and the domain of the clicked url. Looking at domains instead of full urls has the benefit of reducing sparsity. Our high activity user sample contains about 6 million unique urls, and only 1.5 million unique domains.

How can we quantify the amount of information in a click? The goal of contextualization is to predict a user's future interests and actions. So there are two requirements: a click is informative if 1) it presents us with new information, and 2) this information helps us predict the user's future actions. In our example, the click on yahoo.com doesn't provide much new information about the user. This is because the domain is popular. In information theory, the amount of information gained from observing an event is proportional to how surprising the event was. Applying this to clicks, the prior probability that a user will visit yahoo.com is already high, so observing that click has little impact on our existing beliefs. This reasoning leads to the idea that clicks on *rare* domains will be more informative than clicks on popular ones.

Although rarity is a necessary condition, not all clicks on rare domains are informative. Even if we weren't expecting the domain to be clicked, this new knowledge is only helpful if it improves our ability to predict the user's actions. Consider when a user clicks on a spam site. Perhaps the page title was misleading. Can we learn much from such a click? Probably not. While this is an extreme example, to differing degrees many clicks are not what the user hoped for. Click data is very noisy. We are seeking out the clicks that have predictive power. We define a click's predictive power as proportional to the probability that a user will return to the clicked domain. In other words, a click on a *sticky* domain enables us to predict that the user will come back to *that specific* domain. So the most informative clicks are on rare and sticky domains. These concepts are not orthogonal. A popular domain is inherently sticky due to the volume of traffic it receives. This suggests that domains fall into one of three general categories: 1) Popular and sticky, 2) Rare and sticky, and 3) Rare and unsticky.

5.1 Rare and Sticky Domains

Rare domains are ones that few users ever click on. Sticky domain are those that users often return to, if they visit a first time. To find domains which are both rare and sticky, we need to define a concrete measurement for each. Let $P(C_1^d)$ be the probability of a user clicking on domain d at least once. We use this value as a measurement of a domain's popularity. Let $P(C_n^d|C_{n-1}^d)$ be the probability of a user clicking on domain d an nth time, given the first $n-1$ clicks. To quantify a domain's stickiness, we simply use $P(C_2^d|C_1^d)$. These probabilities can easily be estimated from our data sample. Let $|C_1^d|$ be the number of users who clicked on a

Min	Size	d_1^s	d_1^G	δ_1	r_1	d_∞^s	d_∞^G	δ_∞	r_∞
20	74007	0.951	1.0425	0.091	0.0958	0.250	0.276	0.026	0.103
40	45276	0.846	0.945	0.0996	0.118	0.209	0.245	0.036	0.170
75	24934	0.744	0.868	0.124	0.167	0.177	0.223	0.046	0.260
250	4236	0.557	0.752	0.195	0.350	0.129	0.194	0.066	0.511
400	1545	0.492	0.724	0.232	0.472	0.114	0.188	0.074	0.650
700	367	0.421	0.719	0.298	0.708	0.0978	0.188	0.091	0.927

Figure 11: Chart of average distances between first half distribution F_1 and second half distribution F_2 ($d_p^s = d_p(F_1, F_2)$), and average distances of each half to the global distribution G, $d_p^G = \frac{1}{2}(d_p(F_1, G) + d_p(F2, G))$, and the corresponding differences and ratios ($\delta_p = d_p^G - d_p^s$ and $r_p = \frac{d_p^G}{d_p^s}$). Size refers to the number of users with the minimum number of queries categorized. While the distances decrease with more queries seen, the differences and corresponding ratios increase, indicating that users are consistent, but different from the global distribution.

domain d and let $|sample|$ be the number of users in the entire sample. Then $P(C_1^d)$ can be estimated as $|C_1^d|/|sample|$. Similarly, $P(C_n^d|C_{n-1}^d)$ can be estimated by $|C_n^d|/|C_{n-1}^d|$. Users frequently click on the same url multiple times during a search. We filtered these repeat clicks. We also filtered clicks on the same domain within a certain period of time. We wanted to find the domains that users returned after their current information need passed. We tried values of 15 minutes, one hour, and one day, with similar results for all. We applied our metrics to the high activity user sample. We classified a domain d as *rare* if $P(C_1^d) < .1$. In other words, rare domains were defined as those that were clicked by fewer than 1% of all users. Next we dropped any domains with $|C_2| \geq 10$ due to statistical significance reasons. The remaining domains formed the *candidate set* (the set of rare domains with enough data for significance). The candidate set contained 95,529 domains. We then decided that domains with $P(C_2^d|P(C_1^d) \geq 10\%$ would be labeled as sticky. This left 6,035 domains that were both rare and sticky.

Many of the results fall into what may be called *special interests*. They are not generally popular, but people have a clear reason to return. Special interest domains fall under several categories: Ethnic (persianblog.com), communities (cupid.com, christiancafe.com, tagged.com), many state lotteries (calottery.com), finance (navyfcu.org, providianonline.com), children (neopets.com, cartoonetwork.com), specialty news (drudgereport.com, rep-am.com), online games (runescape.com), specialty sales (dreamhorse.com), soap operas (soapzone.com), pornography, etc. Our click information metric could be immediately applied as a simple heuristic for result re-ranking. A search result could be bumped up from rank 100 to the front page if it was a special-domain that the user had clicked on. For example, we suspect our readers might appreciate it if results from Citeseer's domain tended to be ranked higher for them. With short-term adjustment, a single click on Citeseer could dramatically improve results for the rest of your search, even without any prior user history. Clicking on a domain often implies an interest in additional related domains. A related application of such click information would be collaborative filtering.

6. SUMMARY

We explored three main questions: the extent of history available, the convergence of users' queries to topical distribution profiles, and the information about special interest sites or stickiness of rarely clicked sites via patterns of repeated clicks. We found that users' topical interest distributions appears to become distinct from the population, and converge to a stationary distribution, but only after a few hundred queries. Future work in this direction includes using larger sets of categories, different types of classes (such as "navigational" versus "informational"), as well as exploring uses of such interest distributions for clustering the users. We showed how the pattern of repeated site clicks helps identify special interest sites. This information may improve web search ranking and in effect serve as implicit bookmarks, or help identify communities of users with special interest and inform collaborative filtering. Web search is likely to remain a major window onto users' interests and needs, and there is much more to be done in effectively utilizing query logs.

Acknowledgements

We thank Wiley Greiner and John Carnahan for assistance with query and page categorization, and Rosie Jones for discussions and useful suggestions.

7. REFERENCES

[1] S. M. Beitzel, E. Jensen, A. Chowhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *SIGIR*, 2004.

[2] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2), 2002.

[3] C. C. Chen, M. C. Chen, and Y. Sun. Pva: a self-adaptive personal view agent system. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2001.

[4] M. Claypool, P.Le, M. Wased, and D. Brown. Implicit interest indicators. In *IUI '01: Proceedings of the 6th international conference on Intelligent user interfaces*, 2001.

[5] B. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *Journal of American Society for Information Science and Technology*, 52(3), 2001.

[6] B. Jansen and A. Spink. How are we searching the web? a comparison of nine search engine query logs. *Information Processing and Management*, 42, 2006.

[7] T. Kuflk and P. Shoval. Generation of user profiles for information filtering - research agenda (poster). In *SIGIR*, 2000.

[8] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *International Conference on Information and Knowledge Management (CIKM)*, 2002.

[9] O. Madani and D. DeCoste. Contextual recommender problems. In *Utility Based Data Mining Workshop at KDD*, 2005.

[10] M. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27, 1997.

[11] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Conference on Knowledge Discovery and Data Mining (KDD), ACM*, 2005.

[12] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Survers*, 2002.

[13] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large search engine query log. *SIGIR Forum*, 33(1), 1999.

[14] M. Speretta and S. Gauch. Personalizing search based on user search histories. In *IEEEWICACM Interntional Conferenceon Web Intelligence (WI)*, 2005.

[15] A. Spink and B. Jansen. A study of web search trends. *Webology*, 1(2), 2004.

[16] A. Spink, B. Jansen, B. J. Wolfram, and T. Saracevic. From E-sex to E-commerce: Web search changes. *IEEE Computer*, 35(3), 2002.

[17] S. Wedig and O. Madani. A large-scale analysis of query logs for assessing personalization opportunities. Technical report, Yahoo! Research, 2006.