Examining repetition in user search behavior

Mark Sanderson[1], Susan Dumais[2]

[1]Department of Information Studies, University of Sheffield, Sheffield, S1 4DP, UK
[2]Microsoft Research, Redmond, Redmond, WA, USA
m.sanderson@shef.ac.uk, sdumais@microsoft.com

Abstract. This paper describes analyses of the repeated use of search engines. It is shown that users commonly re-issue queries, either to examine search results deeply or simply to query again, often days or weeks later. Hourly and weekly periodicities in behavior are observed for both queries and clicks. Navigational queries were found to be repeated differently from others.

1.   Introduction

With the advent of large scale logging of user's activities on search engines, analysis of those logs has produced much valuable information. Early examples of such work come from Broder (2002), who highlighted the differing forms of queries that users of Web search engines issue. He classified queries into *navigational queries* (where the goal is to find a particular web site); *informational queries* (where the user is seeking information on a particular topic) and *transactional queries* (where the user is looking to find sites, which themselves have to be searched to locate required information).

In almost a decade of research in web search engine query logs much of the published work has analyzed small samples of logs often from a single day. The many works of Jansen & Spink (summarized in their 2006 paper) cover no more than a sample of a days worth of activity from a particular search engine. Although such analyses provide insights about short-term interactions with search engines, they do not shed light on longer-term patterns, which are of interest in this paper.

An early log study by Silverstein et al. (1998) summarized characteristics of almost a billion queries collected over a 43 day period of time. However, the authors did not specifically look at temporal effects or individual usage over time. A temporal analysis of queries covering a week's worth of data was recently reported by Beitzel et al. (2004). Their analyses focused on daily periodicities for queries in different topical categories. More recently, Teevan et al. (2006) examined the search behaviors of 114 anonymized users over the course of one year. In this work, users' repetition of queries and items clicked on in search result lists were of interest. Teevan et al. found that across the year, 33% of user queries were repetitions of queries previously issued by the same user. Repetition across users was lower at around 18%. Teevan et al. also separated out navigational queries, which they defined as queries issued at least twice and where the same URL was clicked in the result list for each query. They found that 71% of repeated queries were navigational; if duplicate repeat queries were eliminated, this number fell to 47%. They also examined *clicks*: the item a user chose

in the search result list. They found that 29% of clicks recorded in the logs had been clicked on by the same user before. Teevan et al. identified different patterns of repetition and described types of user search behavior that fitted the patterns.

From this study, it was clear that search repetition is common. A more detailed study of repetition was undertaken in this paper to better understand search behavior over time with the goal of improving the search experience. We started with simple measurements of the repetition of queries and clicks, but it became clear that many forms of repetition in user behavior exist. Therefore the analysis was broadened to explore repetition and periodicities in general.

2.   Data set examined

The query log analyzed was gathered from a major Web search engine for a 3 month period, from 9th Jan. – 13th Apr. 2006, consisting of approximately 3.3 million queries and 7.7 million search result clicks gathered from 324,000 unique users. Users were selected from a voluntary opt-in feedback system. The query text and the date/time when the query was received by the search engine were recorded. In addition, the URL of items in the search result list that were clicked on, the query that generated that result list, and the rank position of the clicked item were recorded. When users retrieved additional search results for a query, such events were treated as a separate query. This differs from Teevan et al.'s work where all identical queries issued by a user less than thirty minutes apart were treated as the same query. Their approach attempted to capture the notion of a search session. Any such approximation of sessions is prone to error, so we choose to examine the query data in its rawer form.

Users were identified by an anonymised ID associated with a user account on a particular PC. As is the case with most log analyses, if a user has more than one computer each with the opt-in feedback system working, they have multiple IDs. Conversely, if more than one person used the same account on a PC, they were amalgamated into a single user.

## 3.   Initial analysis of query and click repetitions

The first analysis conducted on this data replicated the initial analyses of Teevan et al. determining the level of repetition in queries and in clicked results. Of the 3.3 million queries submitted, 1.62 million were unique to a single user (although many queries were repeated across users); the rest (1.68 million) were submitted more than once by a user. Repeat queries represented a little over 50% of all the queries submitted. This compares to the 33% observed by Teevan et al. We speculate that the different proportion of repeat queries is due to the difference in definition of what counts as a repeat query. The repeated queries were examined to determine how many were navigational queries. Based on Teevan et al.'s definition (same query and same URL click), around 80% of the 1.68 million repeat queries were navigational queries. This compares with the 71% observed by Teevan et al. From this analysis and those published before, it is clear that users repeat queries often on a search engine.

The search result clicks of users were also examined. Of the 7.6 million clicks recorded in the three month period, 1.3 million (17.5%) were found to be clicks accessed more than once by individual users. Within that group of repeat clicks, 83% were from the same query, and 17% were from different queries. A similar ratio of repeat clicks from same or different queries was found by Teevan et al.

We next examined temporal differences in click patterns over time, which Teevan and colleagues did not analyze. Specifically, we examined the number of repeat clicks over time as well as whether repeat clicks were more or less likely to come from the same query as the time between repeat clicks increased.

3.1.   Change in repetition for varying differences in time

Figure 1 shows a histogram of the counts of repeated clicks as a function of the number of days between the two clicks. The numbers of same click pairs steadily declines as the difference in time between the click events grows – searchers are more likely to click on the same URL in close temporal proximity. The curve drops off smoothly for several months and then more suddenly around 90 days due to *windowing effects* in the query logs which cover only 94 days.
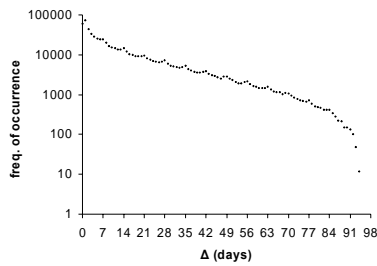
Figure 1: Histogram of counts of same click events in the query log binned by the Δ in days between the two events. Note in all graphs in the paper, the number of paired events = the number of events-1.
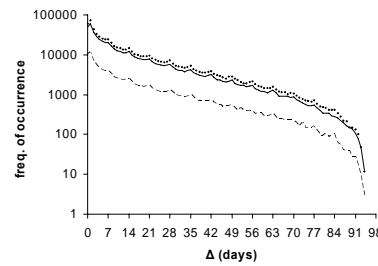
Figure 2: As Figure 1, with the addition of: same click pairs resulting from the same query (middle curve); same click pairs resulting from different queries (lower curve).

The repeat click data was sub-divided into two sets, click pairs resulting from the same query and click pairs resulting from different queries. Figure 2 shows this breakdown. The upper of the two new lines shows click pairs resulting from the same query; the lower line shows clicks resulting from different queries. As can been seen in the graph, the number of repeat clicks from different queries is substantially lower than from the same query. The gap between the two lines decreases slightly as the difference in time between clicks grows. Table 1 charts the proportional difference between the two curves, calculated by the following formula

$$\frac{(Same\,Queries - Different\,Queries)}{Total\,Queries}$$

As the Δ between the paired events of a user clicking on a particular search result URL grows, the user is more likely to reach that URL via a different search query.

Δ (days)	0	14	28	42	56	70
Relative difference	80%	79%	77%	76%	77%	74%

Table 1: Table of the percentage relative difference between the same query and different query lines in Figure 2.

3.2. Periodicities in repetitions

It can also be seen in the histograms in Figures 1 and 2 that there is a seven day periodicity in the data. From this data we can infer that if a user uses a search engine on a particular day of the week, they are more likely to re-use the engine on the same day in the following weeks. The weekly periodicity is observable for pairs of click events that occur months apart. A more detailed analysis of the log data revealed that the periodicity was due to a *weekend effect*. Users who access the search engine on a weekend are more likely to use the engine again on a weekend than on a weekday. It was found that if one observes a user event on a weekend, the probability of that user's next event also happening on a weekend was 55% (by chance, the probability was 28.6%, $^2/_7$). For weekdays, the probability of the next event also occurring on a weekday was 81%, by chance it was 71% ($^5/_7$). The frequency of occurrence of search events on the engine was also different for each day (see Table 2). This combination of factors leads to the observed periodicity.

Sun	Mon	Tues	Wed	Thurs	Fri	Sat
14%	16%	15%	15%	14%	13%	13%

Table 2: Distribution of queries by days of the week.

Recalling the definition of user set out in Section 2 (an ID associated with a user account on a particular PC), we conclude that accounts on a PC used for searching on a weekday tend to be used more for searching on other weekdays and that accounts used on a weekend tend to be used more on other weekends.

4. Further temporal analysis

From the histogram in Figure 1, it was clear that a number of factors were influencing the shape of the graph: the 94 day windowing effect (which caused the slope and sharp tail off) and the weekend effect (which caused the 7 day periodicity). Therefore, we further analyzed the data using a series of normalizations to remove such effects.
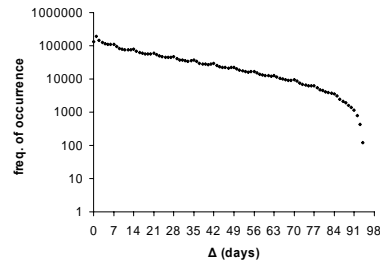
4.1. Normalizing the data

In order to examine just the windowing effect in the data (independent of any repetition), query events in the log were randomly paired ignoring which user or

query they came from. A histogram of the events binned by the number of days between the two events is shown in Figure 3. The windowing effect is clear. We can now use this curve to remove the windowing artifact from other analyses. Queries issued by the same user were randomly paired. A histogram of this plot is shown in Figure 4: both the windowing and weekend effects are present. To remove the windowing effect, the data in Figure 4 was normalized using the randomly paired data in Figure 3 producing the graph in Figure 5. The normalization formula is shown below: the count c at a certain Δ_i (expressed as a fraction of the total counts) is divided by a similarly calculated fraction from the normalizing data.

$$\left(c(\Delta_i) \middle/ \sum_{t=0}^{94} c(\Delta_t) \right) \middle/ \left(cn(\Delta_i) \middle/ \sum_{t=0}^{94} cn(\Delta_t) \right)$$
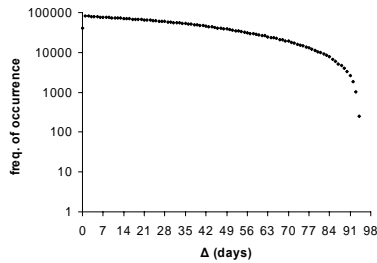

Figure 3: Histogram of randomly paired events.

Figure 4: Histogram of randomly paired events from the same user.

In this normalized view, the horizontal line in Figure 5 crossing the y-axis at 1 is where the data points in Figure 3 would be plotted. Anything appearing above or below the line constitutes a deviation from the norm. Plots above the line are events occurring more often than found in the normalizing data; in contrast, anything plotted below occurs less often.
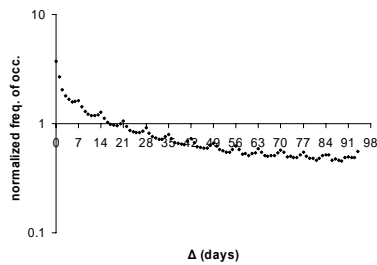

Figure 5: Normalized histogram of events from the same user (Figure 4 normalized by data in Figure 3). Y-axis is a normalized count of the number of events in each bin.

From these graphs it can be seen that if a user issues a query to a search engine, the chances of them issuing another query on the same day ($\Delta=0$) is 3-4 times more likely than would be expected by chance. Users are more likely than chance to re-use the search engine after issuing a query for a period of up to 20-21 days. Two search

events by the same user with a difference greater than 21 days are less likely to occur than would be expected by chance. This graph shows that on average, users' search engine use tends to be *bursty*. If we observe users searching, we are likely to observe them searching again relatively soon, probably within the next three weeks; beyond that time, however, there is an increasing chance they may not be observed again.

4.2. Analyzing query repetition

Replicating the methodology used above, we examined user behavior with repeated queries. Events in the logs from the same user issuing the same query were randomly paired and a histogram of those events binned by the time difference in days was plotted. The graph produced was normalized by the data in Figure 4, so as to eliminate the windowing & weekend effects as well as the burstiness of user search behavior. The results are shown in Figure 6, where it can be seen that users re-issuing of queries to a search engine is more bursty than user search engine re-use. If a user issues a particular query, they are likely to re-issue that query again within the following 7 days. After that, however, the chance of observing the same query from the same user reduces. Users appear to have a limited interest in pursuing a particular query. Whether this is because the user's information need was satisfied or because the user gave up is left for study in future work.
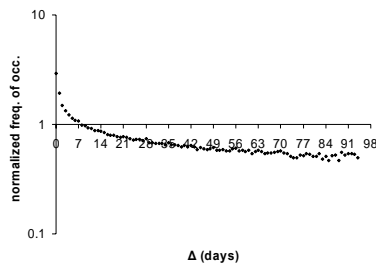
Figure 6: Normalized histogram of counts of the same user issuing the same query.
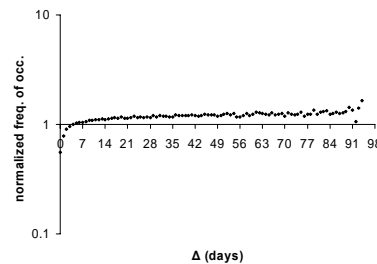
Figure 7: Normalized histogram of repeat navigational queries issued by the same user.

4.2.1. Examining query types
We examined whether the pattern of query repetition observed so far varied for different query types. Based on heuristics like those identified by Teevan et al. and by Lee et al. (2005), we generated a list of navigational queries. Queries in the logs that matched items in this list were randomly paired and normalized by the data used to generate the graph in Figure 6. The results are shown in Figure 7. As can be seen, if a user issues a navigational query, we are less likely to observe the same query being issued again within a few days than would be expected from general repeat query behavior by users (Figure 6). Thus the burst of repeat queries observed in Figure 6 appears due to non-navigational queries, which tend to be more information seeking focused. From this examination of the data sets, we conclude that repeat query behavior is different depending on the nature of the users' query. Navigational queries

are less likely to be repeated by users within a few days than queries with a more information seeking focus, and navigational queries are more likely to be repeated at later points in time.

A final aspect of repeat searching behavior was examined: queries repeatedly submitted to a search engine by different users. For this analysis, search requests from different users were examined. It was found that certain such queries occurred in bursts. Within this set, queries that were topical for the time period covered by the query logs such as "April fools day" and "spring cleaning tips" were found as well as novelty or news queries such as "33-pound cat", "Grammy music awards", "testing of a scramjet engine", etc. It was found that such queries had a higher than expected frequency of occurrence for around 1-2 weeks.

5.   Hourly analysis of user queries

The presence of hourly periodicities in user behavior was also examined. The data used to produce Figure 6 was re-binned to hourly differences between events to produce the results in Figure 8. As can be seen, users' use of search engines follows a strong 24 hour periodicity. Users who query at a particular time on one day are likely to query at that same time on a different day. The data in Figure 6, which shows repeated queries from the same user, was similarly re-binned (shown in Figure 9). Remembering that the data in Figure 6 was normalized to remove windowing & weekend effects as well as user search periodicities, it is striking that users issuing the same query seem to show a stronger 24 hour periodicity in their behavior than is observed in general user search behavior.
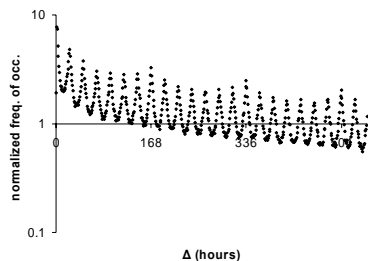
Figure 8: Histogram of the first 22 days of randomly paired query log events from the same user. Note points on the x axis mark out weeks.
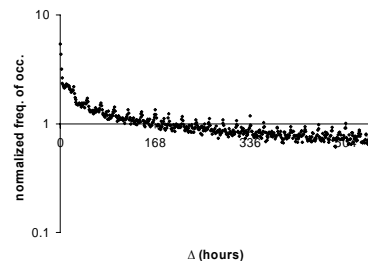
Figure 9: Histogram of the first 22 days of randomly paired events from the same user issuing the same query.

It is not entirely clear why users repeatedly searching with the same query would be more likely to do so at the same time of day. A preliminary examination of these regular queries revealed that some queries seemed to be associated with a particular time of day, such as queries related to a TV show (e.g. "deal or no deal", "american idol"). Others appeared to be queries that a user issued regularly to monitor a particular event or topic (e.g. queries for lotteries; see also Kellar et al., 2006 for a description of monitoring queries). Although there was no quality to these queries that in themselves would indicate they should be issued at a common hour, there was an

indication in the data that the peaks were more likely to occur on weekdays than on weekends. One might speculate that during the week, times when search engines are used are regulated by the structure of people's work and school lives. The exact reason for this periodicity is left to future work.

6. Conclusions and future work

This paper presented an analysis of repetitions in user search behavior. Many queries and URL clicks are repeated over time. Users show both seven day and 24 hour periodicities in their use of search engines and these periods repeat and stay consistent over many weeks. Use of search engines is also bursty with users tending to repeatedly use search engines within a short period of time (e.g. a few weeks). Users re-issuing queries displayed an even stronger burstiness, although for navigational queries, the opposite was observed with users unlikely to re-issue navigational queries within a few days of first issuing the query. Queries that are repeatedly issued by different users were also examined and found to be related to temporally varying events or news (see also Vlachos et al., 2004).

The work in this paper constitutes a preliminary analysis of the topic of repetitions in user interactions with search engines. All analyses presented in this paper, described the general behavior of a large user population. We have not yet examined the variation within the averages and the degree to which individuals deviate from the norm. It is also unclear to what extent the periodicities that we have observed are related uniquely to search engine use or are a reflection of general use of the Web or even of general computer use. An examination of such behavior would be one avenue of future work.

7. References

Andrei Broder. A taxonomy of web search. *SIGIR Forum 36(2), Fall 2002*, 3-10, 2002.

Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David A. Grossman and Ophir Frieder: Hourly analysis of a very large topically categorized web query log. ACM *SIGIR 2004*, 321-328, 2004.

Bernard J. Jansen, Amanda Spink: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management, 42(1),* 248-263, 2006.

Melanie Kellar, Carolyn Watters and Michael Shepherd. A goal-based classification of web information tasks. *ASIST*, 2292-2315, 2006.

Uichin Lee, Zhenuy Lui and Junghoo Cho. Automatic identification of user goals in web search. *WWW 2005*, 391-400, 2005.

Craig Silverstein, Monica Henzinger, Hannes Marais and Michael Moricz. Analysis of a very large web search engine query log. *DEC SRC Technical Note 1998-014*, 1998.

Jaime Teevan, Eytan Adar, Rosie Jones and Michael Pott. History repeats itself: Repeat queries in Yahoo's logs. ACM *SIGIR 2006*, 703-704, 2006.

Michail Vlachos, Christopher Meek, Zografoula Vagena, Dimitris Gunopulos. Identifying similarities, periodicities and bursts for online search queries. ACM *SIGMOD 2004*, 131-142, 2004.