History Repeats Itself: Repeat Queries in Yahoo's Logs

Jaime Teevan
MIT, CSAIL
Cambridge, MA 02138 USA
teevan@csail.mit.edu

Eytan Adar
University of Washington, CSE
Seattle, WA 98195 USA
eadar@u.washington.edu

Rosie Jones and Michael Potts
Yahoo! Research
Sunnyvale, CA 94089 USA
{jonesr,mpotts}@yahoo-inc.com

ABSTRACT

Thanks to the ubiquity of the Internet search engine search box, users have come to depend on search engines both to find and re-find information. However, re-finding behavior has not been significantly addressed. Here we look at re-finding queries issued to the Yahoo! search engine by 114 users over a year.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.*

General Terms

Measurement, Experimentation, Human Factors.

Keywords

Query log analysis, Web search, re-finding, repeat queries.

1. INTRODUCTION

Though much attention has been paid towards understanding Web search behavior and optimizing search engines to support finding behavior, little attention has been paid to re-finding. The lack of support for this activity has led to user frustration. In a study of Web users [4], 17% of those surveyed reported "Not being able to return to a page I once visited," as one of "the biggest problems in using the Web." While many search engines have begun to address the issue by, for example, caching query history, these efforts are just a beginning.

Using a log of the queries and result clicks issued by the anonymous users of 114 Web browsers over a period of 365 days, we explored issues of re-finding. Because we were not interested in short-term query repetitions, we considered all instances of the same query string that occurred within thirty minutes to be a single query. In total, we observed 13,060 queries and 21,942 clicks. The data is comparable in basic statistics to other recent studies (*e.g.*, the average query length is 2.7 words).

Studies of how people re-find have tended to be small-scale laboratory [3] or interview-based studies [2][7]. Log analysis [1][5][6][8] allows researchers to observe a greater variety of behavior than laboratory and observational studies. Surprisingly, little analysis of re-visitation and re-finding has been done of Web logs. Studies of browser logs have found Web site re-visitation is common [8], but query log studies that have looked at search trends for individual users have focused on search sessions [5][6].

2. PREDICTING RE-FINDING

While log studies give a realistic picture of users' actions, they give no insight into underlying motivation. To study re-finding behavior through log analysis, it was necessary for us to try to glean from the data which queries were intended to re-find information rather than find new information. We attempted to capture re-finding intent by looking for repeated clicks on the same search result(s) in response to queries issued by the same user at different times (the query used to find the same result may or may not be the same). A behavior was considered re-finding if a person ran a search with the query "KHTS" and clicked on the results http://www.channel933.com, and later clicked on the same result while search for "channel 933".

Forty percent of all observed queries (5216/13,060) led to a click on a result that was also clicked during another query session by the same user. Of the 21,942 total clicks observed in the data set, 6145, or 28%, of them were clicks on URLs that were clicked by the same user more than once. In contrast, only 1435, or 7%, were clicks on URLs that were clicked by multiple users. People were clearly much more likely to click on results they themselves had seen before.

As re-finding behavior appears to be very common, it would be useful for a search engine to be able to predict when a user is re-finding, because this information could affect the best results to display or the best manner in which to display them. Below, we look at predicting whether a previously viewed result will be clicked based on the query string and past clicks.

Note that often queries that led to repeat clicks also involved clicks on results that had not been clicked before. Of the queries that led to a repeat click, 14% also involved a click on at least one new result. Clearly even if a search engine is able to accurately predict when previously viewed information is being sought, new information is likely to still be beneficial for those searches.

2.1 Repeat Queries

The query a person issued was a good indicator of whether the searcher was going to click on a previously viewed result or not. Approximately 71% (3692/5216) of the queries that resulted in repeat clicks involved the same query string (*e.g.*, a person searched for "oklahoma city fairgrounds", clicked on a result, and then later searched with the same query and clicked on the same result). Not all identical queries led to a repeat click, but 87% (3692/4256) did. It was significantly less common for searches with the same query string to result in clicks on different results (1632 or 38%). It was not always the case that the searcher only clicked on results that were common between two identical searches, or only on results that were unique, as 25% of the searches, or 1070, involved both a repeat click and a unique click.

2.2 Repeat Queries and Past Clicks

We found that we were sometimes able to very accurately predict the likelihood of a repeat click using clicked results from past queries. Navigational query behavior was particularly easy to predict. We called repeat queries where the user entered the same query string, and always clicked on one and only one (and always the same) result, *navigational queries*. Forty-seven percent of the unique repeat queries were labeled navigational. Such queries tended to be shorter in length (13.6 characters v. 16.4 characters), repeated more often (4.0 times v. 3.8 times) and repeated at less frequent intervals (22 days v. 20 days) than other repeat queries.

It was easy to predict whether or not a query was navigational given two previous instances of the same query as training data. By doing this, we were able to automatically labeled 1841, or 12%, of all observed searches as navigational. For these searches, we could predict with 96% accuracy one of the URLs that was clicked. When restricted to predicting the first URL clicked, accuracy only dropped slightly, to 95%, and if we predicted that only that URL was clicked, accuracy dropped slightly more, to 94%. It was less easy to identify a navigational query using only one previous query instance. While doing so covers more of the data (2955, or 23%, of the searches), the prediction was right only 87% of the time. This was not surprising given 87% of all repeated queries have result clicks in common.

3. CHANGE AFFECTS RE-FINDING

We also looked at how changes to result ranking affected people's ability to re-find. Result lists can change due to personalization, relevance feedback, or improvements made to the search engine's underlying index and algorithms, as evidenced by the fact that 27% of the results that were clicked more than once by an individual were not actually in the same rank for each click. We found that changes to result ranking reduced the likelihood of a repeat click and slowed repeat clicks when they happened.

3.1 Change Reduces Likelihood of Re-Finding

We compared the probability that any given click would be a repeat click for re-finding searches under two conditions: (i) when a change in rank was observed among one of the common clicks and (ii) where no rank change was observed. We found that repeat clicks were significantly more likely to occur when there was no observed change. Eighty-eight percent of the clicks were repeat clicks if there was no change in rank, while only 53% of the clicks were repeat clicks if there was a change in rank. It is not immediately obvious whether a decreased likelihood of re-finding reflects a positive or negative influence of result list changes on user experience. It could be that the changes interfered with re-finding, or it could be that the searcher found new and better information in the new result set.

3.2 Change Slows Re-Finding

To get a better idea of whether changes interfered with re-finding, we looked at repeat queries where we were certain that information was being re-found, as evidenced by a repeat click. We measured time from when a query was issued until the common URL was clicked for queries where the finding and re-finding queries were distinct. Table 1 shows the average number of seconds it took to click a URL that was clicked during the initial session when that URL was (i) shown at the same rank it

Table 1. Time to click (seconds) as a function of rank change.

Query type	Mean	Median	StdDev
Rank the same (i)	94	6	234
Rank changed (ii)	192	26	365

originally appeared, and (ii) shown at a different rank. If the rank of the result was unchanged, the second click occurred quickly, while if the rank changed, it took significantly ($p<0.01$) longer.

4. CONCLUSION AND FUTURE WORK

We have looked at re-finding behavior through analysis of the queries issued to the Yahoo! search engine by 114 users over the course of a year. We have observed that re-finding behavior is common, and shown that repeat clicks can often be predicted based on a user's previous queries and clicks. Changes to result ordering appear to slow re-finding. It is our hope that the results of this study will encourage search engines to take a more active roll in supporting information re-finding.

One thing that is evident from our study is that it is important for search engines to return results that match their users' expectations. Teevan [9] has found that when people interact with previously viewed information, it is important to understand what aspects of the original information are memorable before allowing the information to change. This understanding can then be used to highlight important changes (by having changes occur to memorable aspects of the information) or to hide unimportant changes (by only allowing changes to occur to unmemorable aspects of the information). A solution such as this may help search engines provide the best new results to their users while still supporting re-finding.

5. REFERENCES

[1] Anick, P. Using terminological feedback for Web search refinement: A log-based study. In *Proceedings of SIGIR '03*, 2003, 88-95.

[2] Aula, A., Jhaveri, N., and Käki, M. Information search and re-access strategies of experienced Web users. In *Proceedings of WWW '05*, 2005, 583-592.

[3] Capra, R. and Pérez-Quiñones, M.A. Using Web search engines to find and refind information. *IEEE Computer*, 38 (10), 2005, 36-42.

[4] Graphic, Visualization, and Usability Center. GVU's Tenth WWW User Survey, October 1998.

[5] Jones, R. and Fain, D. C. Query word deletion prediction. In *Proceedings of SIGIR '03*, 2003, 435-436.

[6] Kamvar, M. and Baluja, S. A large scale study of wireless search behavior: Google mobile search. In *Proceedings of CHI '06*, 2006.

[7] Komlodi, Anita (2004). Task management support in information seeking: A case for search histories. *Computers in Human Behavior*. 2004 (2), 163-184.

[8] Tauscher, L. and Greenberg, S. How people revisit Web pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 47 (1), 1997, 97-137.

[9] Teevan, J. The Re:Search Engine: Helping people return to information on the Web. In *Proceedings of SIGIR '05*, Doctoral Consortium, 2005.