

Aggregated Click-through Data in a Homogeneous User Community

Mingfang Wu
School of CS & IT
RMIT University
Melbourne, Australia
mingfang.wu@rmit.edu.au

Andrew Turpin
School of CS & IT
RMIT University
Melbourne, Australia
andrew.turpin@rmit.edu.au

Justin Zobel
School of CS & IT
RMIT University
Melbourne, Australia
jz@acm.org

ABSTRACT

There are many proposed methods for using clickthrough data for common queries to improve the quality of search results returned for that query. In this study we examine the search behaviour of users in a close-knit community on such queries. We argue that the benefit of using aggregated clickthrough data varies from task to task: it may improve document rankings for navigational or specific informational queries, but is less likely to be of value to users issuing a broad informational query.

Categories and Subject Descriptors

H.3.3 [Information storage and retrieval]

General Terms

Experimentation, Human Factors, Reliability

Keywords

Information need, query logs, search variability

1. INTRODUCTION

Clickthrough data in search engine query logs can be utilized at an individual user level, where the information is used to infer the user's search interests and to improve the user's future searches. Such data can also be used at an aggregate level, where data for the same or similar queries from many users is pooled. Use of aggregate data avoids any privacy concerns, and is popular in the research literature. In particular, aggregated data for a query can be used to provide recommendations for other users, or to train ranking functions of the search engine and improve its ranking quality for the current query [2, 3].

The assumption underlying the use of aggregated data is that users who issue the same query might share a similar interest, and thus be interested in a similar set of documents. In this study we examine this assumption by investigating: whether a group of users who had similar background and issued the same query would click on the same set of search results; and whether users' information needs could be satisfied by a search result list of high topical relevance. We attempt to answer these questions through examining users'

click variability and query reformulation history in data obtained from a Web proxy log.

2. RESULTS

The data set used in this study was extracted from the cache log from our school's web proxy server [1]. This cache log includes 540,424 queries that were sent to a well known web search engine for the period January to October in 2006. For the purpose of our study, we selected queries that were in the computer science and information technology domain and were issued by at least ten distinct users. We know that all queries were issued by people logged in to student or staff accounts from machines in our department, and, by restricting the domain of the queries, we expect that the users define a close-knit community. It will often be the case that, when issuing a particular query, users were taking the same lecture or doing the same assignment.

We re-ran those queries from the selected domain through the same search engine and kept the queries whose ranked list would have appeared to users. In the end, we collated 135 queries and associated clicks for our study. We also collated topical relevance judgments (by one of the authors and a postgraduate student) in TREC style on the top ten pages returned for each query on a three-point scale: highly relevant (2), relevant (1) and irrelevant (0). Average precision for the top ten pages is reported in the first column of Table 1, split by categories described below. We removed those clicks that did not have associated pages in the (recently) retrieved set, and we also identified queries and clicks that followed with refined searches.

Variability in Clicks. In agreement with other search log studies [3], higher-ranked pages were clicked more frequently than lower-ranked pages. Nearly half of the clicks (46.7%) are on the top ranked page, and 12.3% of clicks are on the second-ranked page.

By examining the relevance of the clicked pages, we infer that the users were not blindly clicking the top-ranked pages. In particular, only 12% of pages clicked by users are irrelevant; and for those queries where high relevant pages were of lower rank, the majority of users ignored those higher ranked but irrelevant pages, and clicked on the lower ranked relevant pages.

Although our users demonstrated a strong tendency to click on highly ranked and relevant pages, they also tended to click on different pages for the same query. Computing the inter-rater agreement of clicks for each query gives an average over all queries is only 0.36. That is, given any pair of users that issue the same query, they will only click on

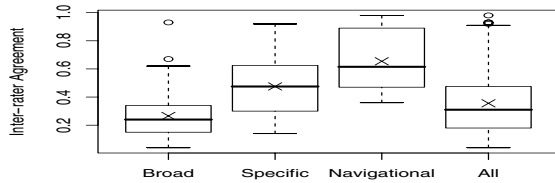


Figure 1: Inter-rater agreement per search task.

	AP@10	Number of clicks	Rank of first click	Rank of all clicks
Navigational	0.90	1.43	1.29	1.68
Specific Inf.	0.95	1.43	1.77	2.08
Broad Inf.	0.89	1.72	3.38	3.80

Table 1: Relevance and click data averaged over all queries of a given type.

the same document 36% of the time.

A possible reason for this low level of agreement might be that the same query string is used for different search tasks. We classified queries into three categories: navigational, specific informational, and broad informational [4]. The navigational queries are those intended is to reach a particular website, while the informational queries are used to learn something by viewing web pages. For the informational category we further distinguished specific queries from broad queries, depending on whether a query indicated a single unambiguous topic, or multiple interpretations.

Figure 1 shows a box-plot of inter-rater agreement per query category (boxes are the 25th and 75th percentiles, the dark line is the median, the cross mark is mean, and whiskers are extreme values). The mean inter-rater agreement of broad information queries (0.27) is significantly lower than the means of specific informational queries (0.48) and navigational queries (0.65) (un-paired two tailed t-test, $p < 0.0001$, with Cohen’s effect size $d > 1.89^1$). Users’ click patterns also vary from task to task as shown in the final three columns of Table 1. The broad informational queries have significantly more clicks than both navigational ($p < 0.03$, $d = 0.66$) and specific informational queries ($p < 0.002$, $d = 0.41$). The average ranks of first clicks and all clicks for the informational queries are significantly lower than the other two categories ($p < 0.001$, $d > 1.10$). This variation in click behaviour may indicate that users’ intents varied for broad informational queries, and that it is unlikely that an information need underlying such a query could be satisfied by pages clicked by other users with the same query.

Variability in Query Reformulation Assuming that if a user reformulates a query then most likely the user’s information need is not satisfied by the retrieved pages, we can get an indirect measure of user satisfaction. For each search with selected queries, we judged whether the subsequent queries was on the same topic as the original query: that is, the query had been reformulated for the same information need [1]. Over all the queries, 41.5% were reformulated.

¹Effect size range: $d \geq 0.8$, strong; $0.5 \leq d < 0.8$, moderate; $0.2 \leq d < 0.5$, small; and $d < 0.2$, trivial.

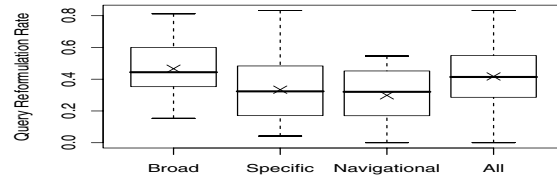


Figure 2: Proportion of queries reformulated.

Figure 2 shows a boxplot of query reformulation rate per query categories. The mean query reformulation of broad queries (0.47) is significantly higher than specific information queries (0.33, $p < 0.001$, $d = 0.98$) and navigational queries (0.30, $p < 0.0001$, $d = 4.76$). We also observe in the first column of Table 1 that the precision of the results list for a query was not related to the proportion of query reformulations.

3. DISCUSSION

As found by Teevan et al. [5] (by explicitly interviewing users), users have diverse intents even if they issue the same query; the low inter-rater agreements on users’ clicked page set and high query reformulation rate as found in our study further supports this finding. Our study also indicates that, even when result lists have high topical relevance as judged in a TREC style by external assessors, users still reformulate their queries and search for different result pages. Only one of our 135 queries was not reformulated by any users that issued that query.

Figure 1 implies the value of using aggregated clickthrough data varies for different search tasks. Using clickthrough data to alter rankings will most likely benefit specific informational search tasks and homepage finding tasks, as these tasks are precision oriented and a user’s information need can usually be satisfied by just one web page. Care should be taken when using aggregated clickthrough data in improving search quality for broad informational queries. For these queries, the top-ranked retrieved pages should be not only relevant but also be diverse to accommodate different search intentions hidden in a query.

4. REFERENCES

- [1] M. Wu, A. Turpin, and J. Zobel. An Investigation on a Community’s Web Search Variability. In *Proc. Australian Computer Science Conference*, 2008.
- [2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behaviour information. In *Proc. SIGIR2006*.
- [3] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. SIGKDD2002*.
- [4] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proc. WWW2004*.
- [5] J. Teevan, S. T. Dumais, and E. Horvitz. Beyond the commons: Investigating the value of personalizing web search. In *Proc. PLA 2005: Wks. on New Technologies for Personalized Information Access*.