Position Paper: Access to Query Logs – An Academic Researcher's Point of View

Judit Bar-Ilan
Department of Information Science,
Bar-Ilan University
Ramat Gan, 52900, Israel
972-3-5318351
barilaj@mail.biu.ac.il

ABSTRACT
Academic researchers have very limited access to query logs of major web search engines. Studying and analyzing large-scale query logs is essential for advancing Web IR. We propose setting up review boards with clear rules for appropriate conduct, and allowing researchers access to logs within this framework.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords
Log analysis, ethical problems, institutional review boards

1. INTRODUCTION
The Web with its billions of documents is today probably the largest available source of information. Because of its size and complexity search engines are needed for efficient information location.

Research related to automatic information retrieval started out in the 1960's with the Cranfield project [14]. The Cranfield Research Project [6] was based on 271 documents and 641 queries. Today, projects of such size could easily be carried out in the academic environment, but the findings would not scale up to the Web. Even using considerably larger test collections are not sufficient for studying search behavior on the Web. As Krishna Bharat stated (if I recall correctly at the "Web experiments and test collections: are they meaningful" panel at WWW2002 - http://www2002.org/panels.html#N3), academics are unable to conduct research on Web based information retrieval, because test collections are small and unrepresentative; solutions will not scale up and will not be able to handle web spam. As an example he criticized the then one of the largest available test collections, TRECs .gov collection (http://ir.dcs.gla.ac.uk/test_collections/govinfo.html) as being non-representative because of the lack of spam in .gov documents. Henzinger et al. [13] provide a clear account of the major challenges of Web search.

The research labs of the search engines are major forces in Web based information retrieval research, but naturally they are secretive of their results, and thus there is still room for pure academic research to satisfy the researchers' interest and to advance Science.

So far I discussed the need for test collections, but documents are only the items that are being retrieved as a result of queries issued by users. The users, their information needs, the way they formulate their queries, evaluate and use the results and their satisfaction with the system and the overall user experience are central "players" in Web IR systems. Now, how can we study the user-side of the retrieval system?

We can learn about users search activities through surveys (e.g. [9, 10]), aggregate reports based on logging activity of a set of users (e.g. [20]), from the results of qualitative user studies (e.g. [24, 3]) or from query log analyses (e.g. [26, 27]). Each of these methods has its advantages and shortcomings. Surveys can be sent to large populations, but usually they are comprised of closed questions and low response rate can skew the sample even if it was initially representative. They are based on recall and not on actual use. Aggregate reports provide basic statistics but are not sufficient for understanding user experience. Qualitative studies, especially when not based on recall, but on actual user monitoring supplemented by information on the user's actual information need and satisfaction can provide valuable information, but these studies involve only small and non-representative sets of participants. Query logs allow non-obtrusive monitoring of the use of search engines based on large user populations, but query logs alone are not sufficient to learn about the users' intent, satisfaction and use of the results. Thus the best way to understand how, why and when users search the Web is to combine all the above and complement them with additional methods.

In section 2 we review some published search engine log analysis studies, discuss the ethical problems and illustrate that findings that are not widely known to the public can be obtained from query log data without interfering with the privacy of the users. Next we propose to setup review boards and guidelines for accessing query logs.

2. QUERY LOGS
2.1 Previous studies
To this date we have data on a few large-scale query log analyses. One of the first published (or perhaps the first) large-scale study was carried out at and by AltaVista [26] - one billion queries in 285 million sessions collected during 43 days in 1998. The data provided by the paper includes the number of terms per queries, use of operators, most frequently used query terms, number of results pages viewed, query modification, query duplication and number of queries per sessions. These parameters have become more or less standard and are reported in the Spink and Jansen

Copyright is held by the author/owner(s).
WWW 2007, May 8--12, 2007, Banff, Canada.

studies [27] as well. Spink and Jansen and their collaborators reported query log analyses of Excite, AlltheWeb and AltaVista (the latest data is from 2002). One of the newer studies examining similar parameters for the clustering search engine Vivisimo analyzed data from 2004 [17]. In a recent article [15], Jansen outlines the methodology for this type of query log analyses. Additional analyses were carried out in [4, 22] emphasizing the temporal aspects of searching. These last two studies were based on AOL data.

Other studies analyzed the query logs of site specific search engines (e.g. [29, 5]). In a recent study Ravid et al. [23] analyzed the log of site providing information for citizens. Most of the requests to this site arrived from search engines, thus the analysis focused on the queries that drove traffic to the Web site.

2.2 Ethical Problems

The techniques and methodologies described by Jansen provide simple descriptive statistics only, however more sophisticated data mining techniques can be applied as well. The results of data mining can interfere with the users' privacy. All the major search engines (Google, Yahoo! and Windows Live [11, 28, 18]) have clear privacy policies, which allow the use of query logs for internal research purposes, but it is not clear whether academic researchers, even after agreeing to comply with the company's privacy policy may get access to these logs. Google specifically mentions that "aggregated non-personal information" may be shared by third parties [11]. Microsoft saw the importance of academic research on search [19] and allowed access to academic researchers who won the RFP awards. All principal investigators had to sign a licensing agreement.

In August 2006, AOL released to the public a very large query log. This is essentially a Google query log, since AOL searches are powered by Google. AOL users were identified in the logs with random numbers replacing actual AOL user names (see a copy of the original announcement [1]). AOL withdrew the query logs almost immediately and apologized for the release of the private data [16], but a number of mirror sites were set up and the data is still freely available as of today (for a summary of the event and issues, see [25]). Both cNet [16] and the New York Times [12] reported that academic researchers are eager to use the logs, but hesitate because of concerns about the users' privacy. Their hesitation is understandable, even though I am confident that these researchers would not be looking for embarrassing personal data, and only use the logs for pure research purposes.

2.3 What could be done with query logs?

In this section I will illustrate that findings based on the query logs can be obtained without identifying the queries and or the users. These findings are not widely known outside the search engine industry. I was interested in the distribution of the placement of the clicked-through items. The Enquiro eye-tracking study found that users concentrate on the top results [8]. Query log analyses ([26, 27]) showed that users usually consider only the first page of search results (i.e., the top-ten results). The AOL log is partitioned into ten subsets; the readme file states that "The data is sorted by anonymous user ID and sequentially arranged." The data appears in ascending order of user IDs in each file, but there does not seem to any order between the files. The distribution of the rank of the clicked items was computed separately for each part. The results show an almost identical distribution of the clicked-through items in each subset (see Figure 1). In about 2.5% of the cases the query is missing – these records were removed from the analysis. For the remaining records in each set we tabulated the rank of the clicked-through item. AOL displayed at most 500 results for a query. In each file there were a few cases (about 0.04% of the queries with click-through) where the clicked-through item was of rank 0 - it is not clear what the meaning of rank 0 is. Not surprisingly, among the clicked items, in more than 40% of the cases the users chose the top-ranking result. In about 89% of the cases the chosen item was one of the top-ten items.

More interesting to note is that in each subset for about 46% of the submitted queries, the user did not click on any of the results (see Table 1).

Why do users submit queries and then do not click on any of the results on the given results page? There could be several reasons for this:

1. The user did not find any satisfactory results on the current page and continued to the next results page
2. The user saw something in the snippets and decided to rephrase the query based on the information in the snippets
3. The user made a typo, and the spell-checker suggested a correction that was accepted by the user
4. The user found the answer to her question in the snippets and there was no need to visit a specific result
5. The user clicked on a sponsored result (it is not clear how these case are recorded in the log)
6. The user had a quick look at the results and decided that they were totally irrelevant and decided to try a different phrasing (without relying on information in the snippets)
7. The user was frustrated with the results and abandoned the search altogether

Table 1: Queries without click-through

	Total queries	Empty queries	% empty queries
Part1	3459420	1598882	46.2%
Part2	3515952	1611723	45.8%
Part3	3577420	1636010	45.7%
Part4	3556331	1635002	46.0%
Part5	3695071	1684517	45.6%
Part6	3464977	1617285	46.7%
Part7	3561753	1632454	45.8%
Part8	3521425	1613277	45.8%
Part9	3524110	1611978	45.7%
Part10	3512732	1614165	46.0%
Total	**35389191**	**16255293**	**45.9%**

The above-mentioned reasons are just speculations based mostly on my personal experience with searching. Some of these points can be resolved through a more thorough examination of the query logs, but for the AOL query log it is not clear whether whole user sessions were sampled. The results raise interesting questions regarding "abandoned queries" and are definitely worth further investigation. In order to understand this behavior, multiple methods should be used. The point I am trying to make is that interesting results can be obtained from query logs without jeopardizing the privacy of the users.

3. ACCESS FOR ACADEMIC RESEARCHERS

The major question is whether it possible to come up with a framework that would allow researchers in the academia access to query logs?

Research in medical sciences, but in social and behavioral sciences as well is routinely reviewed by Institutional Review Boards [30]. These Review Boards follow guidelines (like the Helsinki Declaration for medical research [7]). The NSF has a special page on the protection of human subjects for behavioral and social science research, where a section relates to issues of privacy and confidentiality [21]. The issues discussed there are relevant to query log analysis as well, since the major risk is the invasion of privacy [2]. The rules set up for medical and behavioral sciences allow researchers to "advance science", while at the same time respecting the rights of the patients. It seems to me that the Web research community should follow this example and set up rules for the proper conduct of research. This would allow academics to participate more actively in Web IR. Another question is what incentives the search engines have to hand over their data to academic researchers? One possible answer is that students would get a more realistic basic training in Web IR and thus the search engines would be able to recruit better qualified researchers and engineers.

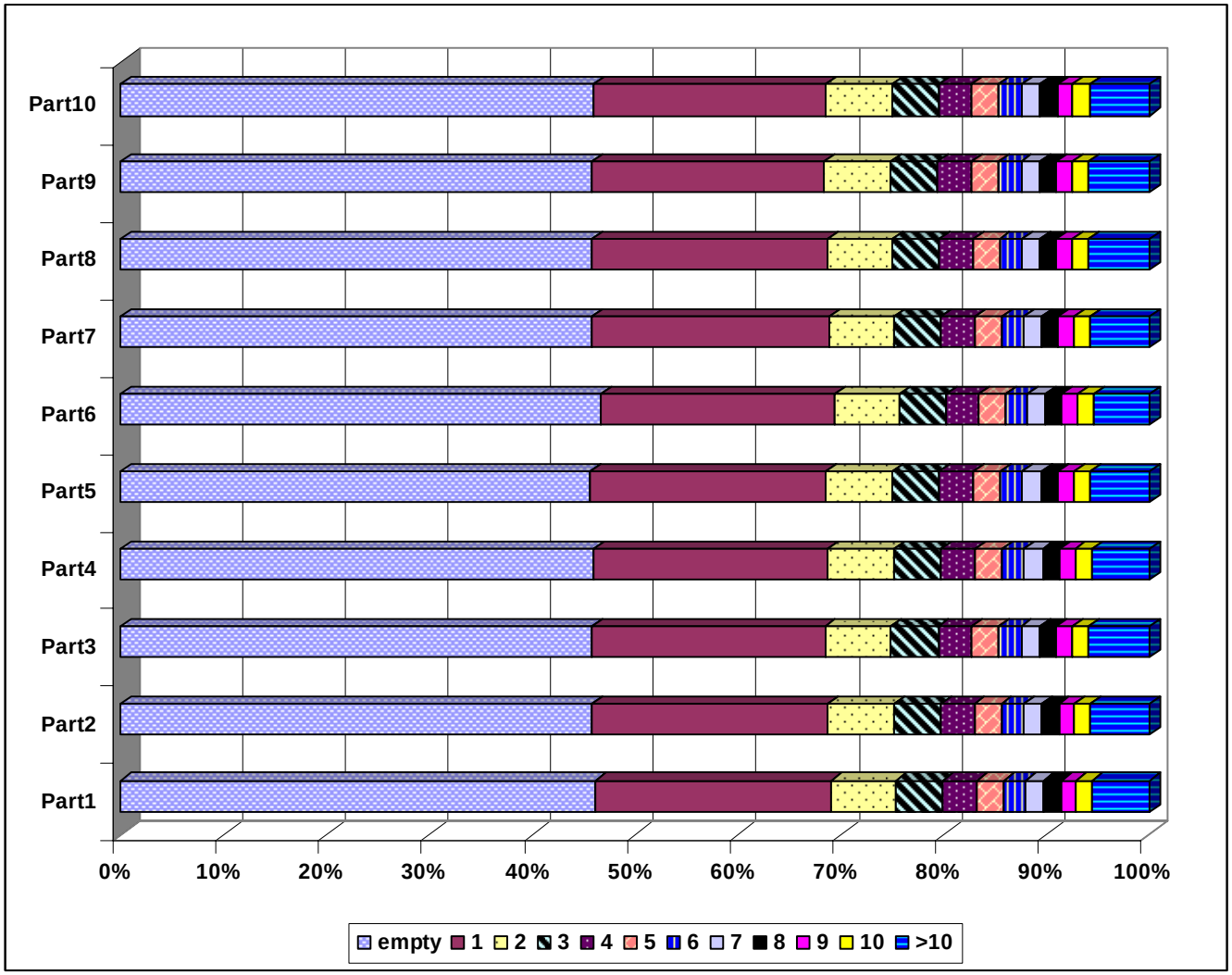

Figure 1: Distribution of the rank of the clicked search results

4. REFERENCES

[1] AOL. AOL 500k User Session Collection 2006. http://imdc.datcat.org/collection/1-003M-5=AOL+500k+User+Session+Collection

[2] Barbard, M. and Zeller, T. A Face Is Exposed for AOL Searcher No. 4417749. The New York Times. August 9, 2006. http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000&en=996f61c946da4d34&ei=5088&partner=rssnyt&emc=rss

[3] Bar-Ilan, J. False Web memories: A case study on finding information about Andrei Broder. First Monday, 11 (2006) (9). http://www.firstmonday.org/issues/issue11_9/barilan/

[4] Beitzel, S.M., Jensen, E. C, Chowdhury, A., Frieder, O., and Grossman, D. Temporal analysis of a very large topically categorized web query log. Journal of the American Society for Information Science and Technology, 58 (2007), 166-178.

[5] M. Chau, X. Fang and O. R. L. Sheng. Analysis of the query logs of a Web site search engine. Journal of the American Society for Information Science and Technology 56 (2005), 1363-1376

[6] Cleverdon, C. W., Mills, J., and Keen, M. Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 1, Design; Part 1, Text. http://hdl.handle.net/1826/861

[7] Declaration of Helsinki. http://www.cirp.org/library/ethics/helsinki/

[8] Enquiro. Did-It, Enquiro and Eyetools uncover search's golden triangle, 2005 http://www.enquiro.com/eye-tracking-pr.asp

[9] Fallows, D. Search engine users. Pew Internet and American Life Project, 2005. http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf

[10] Fox, S. Online health search 2006. Pew Internet and American Life Project, 2006. http://www.pewinternet.org/pdfs/PIP_Online_Health_2006.pdf

[11] Google. Privacy policy highlights. 2005. http://www.google.com/privacy.html

[12] Hafner, K. Researchers yearn to use AOL logs but they hesitiate. 2006. http://www.nytimes.com/2006/08/23/technology/23search.html?ex=1313985600&en=cc878412ed34dad0&ei=5088&partner=rssnyt&emc=rss

[13] Henzinger, M. R., Motwani, R., and Silverstein, C. Challenges in Web search engines. SIGIR Forum 36 (2002), (2), 11-22.

[14] Hjorland, B. Cranfield experiments. 2006. http://www.db.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/cranfield_experiments.htm

[15] Jansen, B. J. Search log analysis: What is it; what's been done; how to do it. Library and Information Science Research, 28 (2006) (3), 407-432.

[16] Kawamoto, D., and Mills, E. AOL apologizes for release of user search data. 2006. http://news.com.com/2100-1030_3-6102793.html

[17] Koshman, S., Spink, A., and Jansen, B. J. Web searching on the Vivisimo search engine. Journal of the American Society for Information Science and Technology, 57 (2006), 1875-1887.

[18] Microsoft. Online privacy statement. 2006 http://privacy.microsoft.com/en-us/fullnotice.aspx

[19] Microsoft. External research & programs. 2006. http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx

[20] Nielsen/NetRatings. November U.S. search share rankings. http://www.netratings.com/pr/pr_061219.pdf

[21] NSF. Interpreting the Common Rule for the protection of human subjects for behavioral and social science research. Confidentiality-privacy. 2006. http://www.nsf.gov/bfa/dias/policy/hsfaqs.jsp#cp

[22] Pass, G. Chowdhury, A., and Torgeson, C.A Picture of Search. The First International Conference on Scalable Information Systems, Hong Kong, June, 2006

[23] Ravid, G., Bar-Ilan, J., Rafaeli, S., and Baruchson-Arbib, S. Popularity and findability through log analysis of search terms and queries. Journal of Information Science, to appear.

[24] Rieh, S. Y. On the Web at home: Information seeking and Web searching in the home environment. Journal of the American Society for Information Science and Technology, 55 (2004) 743-753.

[25] Shen, X. Chronicle of AOL Search Query Log Release Incident. 2006. http://sifaka.cs.uiuc.edu/xshen/aol_querylog.html

[26] Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33 (1999) (1). http://www.acm.org/sigir/forum/F99/Silverstein.pdf

[27] Spink, A., and Jansen, B. J. Web search: Public searching of the Web. Kluwer Academic Publishers, Dordrecht, Netherlands, 2004.

[28] Yahoo! Privacy policy. 2006. http://info.yahoo.com/privacy/us/yahoo/details.html

[29] P. Wang, M. W. Berry and Y. Yang. Mining longitudinal Web queries: Trends and patterns. Journal of the American Society for Information Science and Technology 54 (2003), 743-758

[30] Wikipedia contributors. Institutional Review Board, Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Institutional_Review_Board&oldid=102528161