Query Log Analysis: Social and Technological Challenges

G. Craig Murray
College of Information Studies
University of Maryland
gcraigm@umd.edu

Jaime Teevan
Microsoft Research
teevan@microsoft.com

Abstract

Analysis of search engine query logs is an important tool for developers and researchers. However, the potentially personal content of query logs raises a number of questions about the use of that data. Privacy advocates are concerned about potential misuse of personal data; search engine providers are interested in protecting their users while maintaining a competitive edge; and academic researchers are frustrated by barriers to shared learning though shared data analysis. This paper reports on a workshop held at the WWW 2007 Conference to foster dialogue on the social and technical challenges that are posed by the content of query logs and the analysis of that content.

1 Introduction

The past few years have seen an increase in research that uses personal search histories and search systems' query logs. Such research is enabling users to find more of what they are looking for, quickly and easily. However, it comes at a social cost. Query logs capture explicit descriptions of users' information needs. Logs of interactions that follow a user's query (e.g., click-though and navigation patterns) capture derivative traces that further characterize the user and their interests. The data is rich with personal detail, creating opportunities and risk. The social and technological challenges of working with such data have important implications for query log analysis research.

Within the research community we need an open dialogue on query log analysis. Collecting query log data and tapping into the collaborative knowledge that can be found in query logs is challenging. Sharing information without compromising user privacy is a major hurdle. This line of investigation has important implications. Internet search has become a multi-billion dollar industry, and the stakeholders are highly motivated to improve search products. Internet search has also become an important facet of daily life. Although query logs give only one view into search behavior, researchers investigating information seeking behaviors cannot look at the whole picture without including an analysis of query logs.

Several recent events have increased the public awareness of how much information is stored within large search engines query logs, and how this information can be used to profile a single user without

their knowledge. As a result, public advocacy groups have pressured search engine companies to stop recording any user data, or to purge the data more frequently. The public scrutiny has impacted query log research efforts and academic data sharing has been greatly curtailed. The current state of affairs limits the potential for innovation to silos of industry research. This poses a challenging problem to the research community: how can we share knowledge, verify each others claims and build on innovations?

From Internet search to enterprise search to personalized information management, researchers are turning more and more to the vast stream of log data that users generate while looking for information. The information retrieval community must find a safe balance between the gains to be had from studying users' query and browsing histories and the risks of collecting and analyzing personal use data. The Query Log Analysis workshop held at WWW 2007 provided an interdisciplinary venue for collective thinking about query log analysis from multiple angles. The workshop brought together individuals from multiple backgrounds, including academic institutions, Internet search engines, industry research, as well as experts in information and technology policy. The workshop program focused on expanding our understanding of the problems and concerns. We focused on where the state of the art is in learning from web search query logs, what the risks and benefits are for search engine users, where the challenges lie in exchanging information and sharing knowledge, and what policies and practices are needed to ensure a bright future. This report on the workshop begins with an overview of the workshop's structure, followed by a more detailed discussion of the specific presentations. It concludes with a summary of some high level take away messages from the workshop.

2 Structure of the workshop

The workshop was organized into three foci of discussion: *advancing technology*, *broader social issues*, and *possible solutions*. The morning sessions (*advancing technology*) were primarily technical in nature and focused on the benefits and limitations of query log analysis. Six research presentations covered various aspects of query log analysis, including how and why it is used and the kinds of improvement to technology that can be gained. The afternoon included two panel discussions. The first (*broader social issues*) focused on broad issues of data availability, including social concerns. The second (*possible solutions*) focused on technical solutions and policy solutions for preserving user privacy in query logs. The day was concluded with an open group discussion on future plans and ways of moving forward.

2.1 Workshop participants

The workshop attracted participants from academia and industry. Six full papers and five short papers were accepted to the proceedings. Panel discussions at the workshop also included contributions from two policy experts, John Morris from Center for Democracy and Technology, and Daniel Weitzner, director of the W3C's Technology and Society activities.

2.2 Overview of discussion

A primary objective of the workshop was to offer participants opposing views on the value and cost of query log analysis, and an opportunity to see multiple perspectives. Industry researchers had an opportunity to demonstrate why query log analysis is so important to the industry. Academic researchers had an opportunity to voice concerns over the possible negative effects of not have access

to query log data. Policy analysts had an opportunity to point out parallels in other technologies and to describe relevant legal precedents.

Three sets of questions were proposed at the beginning of the day as a starting point for group discussion:

1. Can query log data be safely collected and analyzed? Should it?
2. Can query log data be anonymized and shared? If so, how should it be done?
3. Can we establish standards of practice for query log analysis?

Group discussion was encouraged throughout the day, in particular during the panel and open discussion sessions. The main topics of discussion included:

- Advancing Technology
  - Sufficiency of query log data for user behavior analysis
  - Use of query log data for machine learning tasks
  - Use of query log data for system evaluation and tuning
- Broader Social Issues
  - Ethics of query log data collection and use
  - Legal issues in query log data collection and use
  - Policies of practice and implications of user specific data
  - Public awareness about query logs
- Possible Solutions
  - Anonymization of query log data
  - Standards of practice for producing and sharing logs

3 Presentations and panel discussions

Below is an overview of presentations given and panel discussions that were held throughout the workshop. Details on any particular paper or talk can be found in the WWW conference proceedings and to the workshop website (http://querylogs2007.webir.org). Here we merely attempt to capture the main highlights and general spirit of the discussions.

3.1 Technical presentations on advancing technology

As mentioned earlier, the first part of the day focused on the technical advances made possible through query log analysis. The purpose was to give voice to research projects that gain from the data available in query logs and to underscore why query logs are an asset worth developing. The six presentations on the benefits and limitations of query log analysis were:

- *Query Logs Alone are not Enough*
  Carrie Grimes, Diane Tang and Daniel M. Russell (Google)
- *Comparing Click Logs and Editorial Labels for Training Query Rewriting*
  Vivian Zhang & Rosie Jones (Yahoo!)
- *Can We Find Common Rules of Browsing Behavior?*
  Ganesan Velayathan and Seiji Yamada (National Institute of Informatics, Tokyo, Japan)
- *Functional Faceted Web Query Analysis*
  Viet Bang Nguyen and Min-Yen Kan (National University of Singapore)
- *Web Search Engine Evaluation using Clickthrough Data and a User Model*
  Georges Dupret, Vanessa Murdock and Benjamin Piwowarski (Yahoo!)

- *A Study of Mobile Search Queries in Japan*
  Ricardo Baeza-Yates, Georges Dupret and Javier Velasco (Yahoo!)

Each of the presentations highlighted a different technology gains derived from query log analysis. Diane Tang's presentation, however, underscored the fact that query logs only give a small view of what a user is actually doing or looking for. Logs can play a part in describing users' information behaviors but have a limited capacity on their own for providing useful detail on an individual. Click logs are often used in conjunction with query logs for predicting users' intent and for analyzing system enhancements. Zhang and Jones reported on their use of clicks in conjunction with query logs for query rewriting. They showed promising preliminary results. George Dupret presented two other research projects also from Yahoo! in which query logs were instrumental. In all of the presentations it is clear that some amount of machine learning and data analysis performed on quality sets of user query logs can have high payoff both for the user and the search provider.

Faceted classification schemes have been very successful in information classification. Bang Ngueyen presented a four-facet scheme for classification of users' queries. Their scheme presents actionable differences between queries. That is, based on a classification within these four facets, search engines could use different retrieval algorithms. Using an available query log and human judgments of queries over the four facets, they evaluated machine learning techniques for automating query classification. The implication of their work is that using a large enough set of tagged queries, search engines can react to different kinds of requests and improve service to their users.

Although the term "query log" is typically used to refer to a large centralized store of (server side) data, Ganesan Velayathan presented research conducted with Seiji Yamada that utilized client side logging for machine learning of user behavior patterns. Their work explored connections between user interest and user behavior, and offered an alternative method for evaluating web pages by incorporating client side logs. This approach might circumvent some of the issues presented by centralized query logs, but must come at a cost of accessibility of the data. For example, learning algorithms trained on client side logs have no opportunity from learning about general patterns of behavior in the broader population of users.

Three central themes emerged from these of presentations: (a) that we can model a lot of important aspects of information seeking by looking at logs of information seekers query and click-through behavior (b) that the data on its own is not enough, and (c) that serious advances in technology are possible if we can overcome problems associated with query log data.

3.2 Panel discussion on broader social issues

The first afternoon panel discussion focused on several broader issues associated with query log data collection, retention and dissemination. The purpose of the panel was to give voice to research projects that suffer from lack of access to query log data and to pose some of the difficult questions about broader risks of query log collection, retention and dissemination.

Access to Query Logs - An Academic Researcher's Point of View – Judit Bar-Ilan (Bar-Ilan University) presented a position paper on how industry can benefit from sharing data with researchers. She pointed out that many academic researchers are keen to triangulate findings from other data sources against findings from query log data, but that only a sparse set of very old query

logs are generally available to researchers outside of the search companies. She also pointed out that academic institutions will have difficulty training researchers in this field without access to data. In order to foster collaboration, Bar-Illan suggested that clearer guidelines are needed. Search engine companies each have their own public privacy policy guidelines, but in some cases it is not transparent to the public what is done with user data internally. She suggested setting up review boards and clear guidelines. Challenges that lie ahead for such guidelines include interpreting common rules for the protection of human subjects in behavioral and social science research as they apply to online environments. Important within the existing guidelines is a distinction between *privacy* and *confidentiality*. Privacy refers to individuals and their interest in controlling access of others to themselves, whereas confidentiality refers to exiting data and agreements regarding restrictions on access to the data.

Preserving the Collective Expressions of the Human Consciousness – Bernard J. Jansen (Penn State University) presented a position paper on why society in the long term might benefit from retaining query data. Query logs, he suggested, are a reflection of the character, values, fears, hopes and desires of the people who issue them. He pointed out that we know what the first words spoken on a telephone were, but we have no idea what were the first words typed into a search engine. Individually, we tend not to preserve our digital communications. Few of us can say what the first email message we sent was, or what was the first thing we searched for on the Web. Internet search has become a major component in our modern lifestyles, but years from now when the technology has changed, social scientists may have very little evidence of what we were searching for. Jansen proposed that we collect and preserve query logs as a snapshot of the zeitgeist of our times. He suggested that a cooperative partnership might be set up to architect long term storage of such data and control access as needed.

John Morris (Center for Democracy & Technology) gave a careful analysis of why retaining and sharing data are ethically and legally challenging. Library catalog search presents a close analogy to Web search. Libraries are held to strict legal guidelines regarding what data they can retain and under what circumstances they are allowed, or even compelled, to disclose that data. The American Library Association also has strong guidelines limiting data collection and retention as a means to protecting patrons' privacy. As a result, libraries follow carefully constructed policies for minimizing what data they collect and purging data frequently, including Web browser caches. Morris highlighted four important components to privacy policies: data minimization, clear notice, informed consent, and the ability to refuse consent. He also contrasted the reasons for retaining data that contains personal information against the threats raised by data retention. Data that is retained is potentially data that can be disclosed and should not, whether that disclosure is accidental, malicious, or legally compelled.

The panel discussion yielded a number of difficult open questions. Many were recurrent themes throughout the day and are distilled further in Section 4 below. Chief among them were questions of the value of data to users and to society. Expectations of privacy may be shifting as we balance individual concerns against societal threats. (Consider the current debate over warrantless surveillance.) Is privacy going away or are we just learning to manage our privacy differently? Issues of data privacy vs. data use call into question our understanding of user specific data in broader contexts. Beyond machine learning and improved service, user specific data has implications as a representation of an individual within society. Should we just throw away all this data? What

data is appropriate to collect, and what is acceptable use of that data? What are the classes of research questions are we trying to answer with query log data, and what data are truly needed to answer those questions? Finding answers to these questions should be a high priority within the research community.

3.3 Panel discussion on possible solutions

Eytan Adar (University of Washington) gave a presentation from his paper *User 4XXXXX9: Anonymizing query logs.* He demonstrated why certain approaches to anonymizing log data fail and why many approaches will only partially protect personally identifiable information. Ultimately, there is a trade off between usefulness of the data and identity protection for the individual. There are some simple steps that can be taken to reduce the probability of identifying an individual from that individual's queries. Each technique comes with a different impact on the usefulness of the data for a given task. For example, queries that are highly specific to an individual occur very infrequently. Setting a threshold on minimal number of occurrences for inclusion in a query log (e.g., removing singletons) would greatly reduce risk of exposure, but would also create problems for identifying new queries. Splitting an individual's queries into blocks, either by time or by topic, can reduce the amount of information that can be aggregated about that individual. But splitting makes certain objectives, such as personalization or query recommendation, more difficult. Cryptographic solutions come with similar (possibly stronger) benefits at similar costs to the search provider. The objective of preventing queries from being mapped to an individual must come at a cost of service to the individual, abilities of the service provider, or both.

Li Xiong (Emory University) gave a related presentation extending from the paper *Towards Privacy-Preserving Query Log Publishing*. Like Adar, Xiong showed that techniques for de-identifying user data—such as removing named entities and user ids, or conflating search session—all involve a trade-off between maximizing utility of the data and minimizing potential privacy breaches. Xiong also pointed to guidelines governing other types of individual data—such as HIPPA for medical records—as a potential model for better policies governing retention and use of query log data.

Daniel Weitzner (W3C, MIT-CSAIL) spoke to how policy may be used to solve some of the privacy challenges. He presented a number of relevant legal precedents, mostly from the realm of technology, that illustrated ways in which our understanding of privacy has shifted. As technology has enabled more remote access to personal conversations, the definition of privacy has shifted away from one about physical circumstance (i.e., location, use of specific technologies, etc.) and toward an understanding of intent. This means that there are certain activities a person can do in their own home without being entitled to an expectation of privacy, while there are other activities for which that same person may be entitled to privacy protections in public spaces. Weitzner cautioned that we should not think about query logs in isolation, as query logs can be linked with other public data. Cultural norms also have significant privacy implications. In as much as the Web is global, it is not clear what norms are appropriate for establishing laws or policies governing Web search query logs or other Web-related personal activities. This echoed earlier points made by John Morris that different nations have very different notions about what protections an individual deserves. One poignant solution proposed by Weitzner was that policies for query log analysis might follow a model like the governances imposed on credit records. A tremendous amount of personal data is collected in an individual's credit history, but in order to use the data credit agencies must limit themselves to preapproved purposes and individuals must be allowed to audit that information. Among the many

important questions raised during this panel discussion—echoing the earlier panel—was the question, "What do we think are acceptable uses of the data?"

TrackMeNot – During breaks between sessions a short video presented a demonstration of TrackMeNot[1], a browser plug-in that gives covering traffic as a means of protecting user privacy. This software solution sends a large quantity of pseudo-random queries from a user's browser to mask that user's actually query. The result is that for each legitimate query from that user, the search engine's query logs will contain hundreds of queries illegitimate queries. This highlights the possibility of an escalation between users seeking additional privacy protections and search companies seeking more information about their users. As search engine companies seek more information about their users, wary users may respond with antagonistic solutions. This begs the question, is query log anonymization an inherently adversarial model?

Following the panel presentations, a number of possible solutions to protecting data privacy were discussed. Informed consent mechanisms are clearly not strong enough in many of the data collection environments. One suggested solution was to establish a client side equivalent to the robots.txt file, which is used to signal to Web crawlers what part(s) of a website should and should not be indexed. By way of analogy, a user.txt file could allow users to declare what aspects of their behavior should or should not be logged. An obvious problem with this solution is that it is dependant on compliance by the data collector. Other technical solutions discussed included merging search engine logs from multiple search engines, or building search engines specifically for collecting research data. These have clear advantages for protecting search providers from risk while yielding useful data for research, but they leave several issues open with respect to protecting individuals' expectation of privacy.

In all of the presentations it was clear that technology alone cannot solve the problems associated with privacy. One important motivation for the establishment of better privacy protections in query log analysis is the risk of Congressional legislation. If industry does not self-regulate, a dissatisfied public may pressure the government to intervene. Possible legislative solutions include mandatory deletion of logs and/or required access to data by the users themselves.

3.4 Group discussion

At the end of each of the panel discussions, and in a separate session at the end of the day, workshop participants were encouraged to engage in a dialogue about the issues at hand. What became clear in these discussions is that there are no clear answers. The workshop discussion touch on some of the most pressing questions and some proposed solutions to the issues that have been raised.

Many of the solutions proposed involve some level of anonymization applied to the data in query logs. One extreme alternative solution is to dispose of query logs all together. The benefits to industry and to the consumer seem to make this an unlikely solution. Therefore, an important question arises as to what data we really need to be collecting and retaining. Logging technology has enabled the collection of a tremendous amount of data. It would be advisable for researchers to focus on a few important questions and aggregate data suited to those purposes. Built-in time delays for the use of data are also advisable. Using less immediate data has lower potential downside for

[1] http://mrl.nyu.edu/~dhowe/trackmenot/

researchers and can go a long way toward addressing privacy concerns. However, it is not clear what an appropriate period of time would be.

Indeed, it is not clear what is appropriate on a number of points regarding query log analysis. One suggestion that came from the group discussion is that the community may want to establish a "best practices" working group to investigate issues related to query logs. Similarly it was suggested that institutions may benefit greatly from setting up a review board to govern the collection and use of user data logs. This could address many of the concerns of privacy advocates and academic researchers, but it could be insufficient or too constrained for industry.

The scope of the conversations that concluded the day was broad. A concise report of each and every point is not possible within the limits of this paper. That said, we hope to have hit some of the most important notes and to have conveyed the highlights accurately.

4 Conclusion

In public spaces and private ones, many institutions are capturing information about our preferences and needs—sometimes deliberately, sometimes inadvertently, but always with a degree of risk to both the individual and the institution. Retailers offer deep discounts to customers who allow their purchasing habits to be tracked (e.g., via store branded "bonus" cards). Meanwhile other institutions like public libraries make concerted efforts to purge their data caches frequently and to effectively clean away traces of patron behaviors. Search engines must figure out an appropriate balance within this space. Query logs and query log analysis pose a number of unanswered questions regarding technology and policy. The recorded search behaviors of individuals can compromise their identity, and institutions holding such data must be careful what they record and with whom they share it.

The workshop on Query Log Analysis: Social and Technological Challenges contributed to opening a dialogue within the research community on how to advance query log research without compromising ethical integrity. There were a number of important take-away messages. A central theme in all of these was the trade-off between costs (both social and economic) and benefit. Chief among there were:

- Benefits from collecting and analyzing query log data:
  - Improve search for the general public
  - Improve search for the individual (personalization)
  - Record of human history
  - Competitive advantage for holder of the data
  - Useful for law enforcement

- Costs associated with collection of query log data:
  - Data could be misused by the collecting institution
  - Data could be leaked to individuals with malicious intent
  - Data can be subpoenaed by the government

The workshop succeeded in its goal to represent multiple views and to foster discussion on difficult problems. Further research and development is needed not just in the technology but also in the associated policies. Many of the participants left with a greater understanding of the problems we

have to face together and the nature of the landscape that lies ahead. It is our hope that this workshop is a first step in arriving at an acceptable solution.

5 Acknowledgements

We would like to thank the W3C for hosting the workshop. We would also like to thank Einat Amitay for her hard work in helping organize the workshop, and Rosie Jones for invaluable input and support.

6 References

Proceedings of the Sixteenth International World Wide Web Conference (WWW2007).
General Chairs: Carey Williamson and Mary Ellen Zurko
Program Chairs: Peter Patel-Schneider and Prashant Shenoy
May 8-12, 2007. Banff, Alberta, Canada. ACM Press: New York.
(Available online at http://www2007.org/proceedings.html)