

# Introduction to Special Issue on Query Log Analysis: Technology and Ethics

EINAT AMITAY

IBM Research

and

ANDREI BRODER

Yahoo! Research

---

It has been ten years since the first published analysis of a Web search engine query log. The year was 1998, and the results were based on a sample of the AltaVista search query log which did not include any user or session information but sufficed for understanding query caching performance. Since then, there have been many studies surveying the potential use of query logs for improving search engines' efficacy and usability. Such studies mined the logs to improve numerous search engine capabilities such as query refinement and expansion, spell checking, ranking, targeted advertising, etc. As search engines have become the principal gateways to the Web, their query logs now capture almost the entire Web experience, tracking what users queried for, what results they chose to follow, and what subsequent queries they submitted after viewing the initial results. The value of the logs for research into anything Web related is beyond doubt, but the sheer amount of information captured in the logs makes protecting the user's privacy difficult.

Two years ago, in August 2006, AOL released a sample of a query log for research purposes. The release included the trace of queries submitted to AOL by 500,000 users over a period of three months and included anonymous user IDs. Unfortunately, the anonymization used by AOL could easily be defeated for some users. This fact caught the attention of the media and surfaced many ethical and practical issues not previously debated outside the Web Information Retrieval community. Privacy advocates called for search engine regulation and for increasing users' awareness and control over the systematic analysis of query data. The negative publicity raised by the AOL incident resulted in the restriction and even the cessation of research collaborations between industry and academia.

In editing this special issue, we wanted to present three considerations that should inform the query log use debate. The first and foremost consideration is the privacy concern raised by the use of query logs: The first article appearing in this issue was written by Alice Cooper from the Center for Democracy and Technology (CDT). Cooper surveys the threats to search engine users posed by keeping and analyzing query logs. She studies the currently available

technologies for obscuring user identity in query logs and the possible methodologies for seeking user consent for keeping private user information.

The second consideration is that the analysis of query logs is essential for the improvement of core search engine performance and that without such analyses search engines will be less efficient and less usable. We chose to show this through an article about caching that manages to significantly improve search performance. Baeza-Yates et al., in an article that stemmed from collaboration between Yahoo! Research and the Information Science and Technology Institute (ISTI) of the Italian National Research Council (CNR), reduce the size of the search results' cache based on correlations between the frequency of terms in the collection, the query log, and the documents retrieved. This article demonstrates that query logs are valuable even when all user-specific data is deleted; the only features required for improving the performance of the underlying search engine were the terms and the timestamps of the queries.

The third consideration is that query logs might reveal valuable information unrelated to search engines and unavailable from other sources. Matt Richardson from Microsoft Research aligns the terms found in queries over time periods that are much longer than the traditional search session. He shows that the analysis of query logs kept over longer time spans reveals nontrivial implicit correlations between events that otherwise seem unrelated. Such knowledge might become impossible to mine if query logs were to be deleted after a short interval or broken into smaller time spans. This article demonstrates both the cost of putting limits on the information stored and the obvious benefit of analyzing queries for other purposes such as improving medicine based on new facts mined from the logs.

Although we accepted only three articles for this special issue on query log analysis, there were many more papers submitted and reviewed. We would like to thank all the authors of submitted papers and encourage them to continue their research in this area. We would also like to thank the dozens of reviewers who read and commented on all the submissions and helped us choose the ones appearing in this issue.

In closing, we would also like to encourage the Web research community, the privacy community, and the commercial search engines to amplify their dialogue, understand each other concerns and goals, and find creative solutions that will allow richer datasets to be made available for research without endangering user privacy.

Have an interesting read.