Preserving the Collective Expressions of the Human Consciousness

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania 16802
jjansen@acm.org

ABSTRACT
Web search engines use transaction log files to record a copious number of interactions that occur between the user, the Web search engine, and Web content. Search engine companies use these records of interactions to improve system design and online marketing. In order to address privacy concerns, some question whether it is wise for search engine companies to preserve these query logs. However, not preserving the query logs from Web search engines would be (and is) a critical loss of a temporal record of the expression of the collective human consciousness. In this paper, an outline of an action plan to preserve these records is proposed to generate discussion of such a course of action.

Categories and Subject Descriptors
E.5 [**Files**]: Organization/structure

General Terms
Human Factors, Legal Aspects

Keywords
Search Logs, Search Engines, Public Records

1. INTRODUCTION
The Web has become an essential facet in the daily lives of many people with impact in nearly every area of human endeavor. Its influence will continue to grow as the Web increases its penetration to all areas of the globe. Search engines are the main portal to the Web. With nearly 70% of Web searchers using a search engine as their point of entry, the major Web search engines service millions of queries per day and present billions of results per week [3]. Search engines are "the app" that many people use on a daily basis for accessing information, navigating to Internet sites, and locating transactional services on the Web.

These search engines record in transaction logs the interactions among users, the search engine, and Websites. A transaction log is *an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine* [2]. Transaction logs (also referred to as search logs or query logs) are an unobtrusive method of collecting significant amounts of searching data on a sizable number of system users.

The use of data stored in transaction logs of Web search engines, Intranets, and Web sites can provide valuable insight into understanding the information-searching process of online Web searchers. Such an understanding can inform information system design, assist interface development, and aid in the devising of information architecture for content collections.

Web search engine companies use queries logs to research searching trends and effect system or interface improvements. Search engine companies also provide aggregate searching trends to the general public (c.f., Google Zeitgeistat http://www.google.com/press/zeitgeist.html, Yahoo Buzz Index at http://buzz.yahoo.com/buzz_log/?fr=fp-buzz-morebuzz, AOL Hot Searches at http://hot.aol.com/, DogpileSearchSpy http://www.dogpile.com/info.dogpl/searchspy/ and MSN Search Insider at http://www.imagine-msn.com/insider/).

However, so far the search engine companies have received most of the benefits from these query logs. The public has access to only the most general and popular of searching statistics. In other words, these transaction logs are a private resource of the search engine companies. Conversely, one can view these transaction logs as a public resource. These query logs can and should be viewed as public records that must be preserved as expressions of the collective human consciousness.

In the following sections, we outline the concerns of public access to query logs of Web search engines. Then, the costs of not preserving these logs are discussed. Finally, an outline of a proposal to preserve these logs is presented in order to stimulate a discussion of the necessity of preserving these logs.

2. BACKGROUND
Humans appear to have always found was to express themselves. There are cave drawings that are tens of thousands of years old in many parts of the world. There are records from the original Chinese dynasty. There are rope records from the Inca Empire. There are papyrus scrolls from ancient Egypt. We have copies of Christian writings that are nearly 2,000 years old. We have original copies of the Gutenberg Bible. Thanks to copious lab notes, we have a record of the first telephone call (i.e., *Mr. Watson -- come here -- I want to see you.*). We have many such records of human expression, each articulated in the medium of the time.

What is interesting is that this record of human expression has not carried over into the era of Web search engines. We do not know the first search query, even though search engines entered the scene in only the early 1990s.

In 1993, Matthew Gray created what most consider the first Web robot, called World Wide Web Wanderer. Initially used for counting Web servers, World Wide Web Wanderer later obtained

WWW 2007, May 8--12, 2007, Banff, Canada.

uniform resource locators (URLs), forming the first database of Web sites called Wandex. Also in 1993, Martijn Koster created ALIWEB (Archie-Like Indexing of the Web). ALIWEB allowed users to submit their own pages for indexing. There are several good reviews of this pre-Web history, including several good Web sites such as http://en.wikipedia.org/wiki/Search_engine and http://www.search-marketing.info/search-engine-history/.

From these initial efforts, the Web search engine landscape really took off. Excite was introduced in 1993, was incorporated, and went online in December 1995. Jerry Yang and David Filo created Yahoo! in 1994. Later in 1994, WebCrawler was introduced as the first full-text search engine on the Internet, where the entire text of each page was indexed for the first time. AltaVista began in 1995 being the first search engine to allow natural language inquires and advanced searching techniques. It also provided a multimedia search for photos, music, and videos. Google was launched in 1997 by Sergey Brin and Larry Page. In 1998, MSN Search and the Open Directory were started. The search engine market (in the US) is current dominated by a handful of major search engines (Ask.com, Google, Yahoo!, MSN LiveSearch), although there are numerous others (e.g., Dogpile, AltaVista, AlltheWeb.com, Lycos). There were many other search engines along the way that had an impact, including Infoseek, NorthernLight, and Inktomi.

Over more than a decade, millions of searchers have submitted billions of queries and visited multiple billions of Web pages. Unfortunately, we have very little records of these expressions from searchers to search engines. We have lost a wealth of vital and important records of human expressions. Battelle [1] calls this record the "Database of Intentions, defined in his blog as *"a massive database of desires, needs, wants, and likes that can be discovered, subpoenaed, archived, tracked, and exploited to all sorts of ends*." It is certainly this, but it much more. At a global level, it is the collective expressions of the human consciousness. As such, it deserves preservation.

Human consciousness is the perceived relationship between oneself and one's environment at some temporal point. These queries, as expressions, are unique in capturing the "gaps" or incidents that trigger one to go to the Web to seek information or services. Query logs provide a unique longitudinal view of these incidents on a scale that exist no where else.

Existing in electronic form, this expression of humanity in Web search engine queries is being lost. It continues a trend of a lost of electronic records. The lost of early radio and television shows is well documented. We are not sure who send the first email, but it was probably Ray Tomlinson. However, we do not have the contents of first email message. We do not have the first Web page (http://info.cern.ch/hypertext/WWW/TheProject.html was its' URL though). Jerry Yang and David Filo do not have the first Yahoo! Webpage [1]; the earliest version available is from 1996.

This loss of records also occurs at the individual level. Do you, the reader, have your first email message? Did you record your first search query? These expressions could tell you a lot about yourself at that period of your life.

Preservation of query logs has similarities to other efforts. Preserving our electronic heritage was the prime motivator for the Internet Archive Project (a.k.a. the WayBackMachine http://www.archive.org/about/about.php), which provides access to snapshots of Web pages from about 1996 to the present. At the individual level, there is Gordon Bell's MyLifeBits (http://research.microsoft.com/~gbell/). There should be a similar project for preserving query logs and making these available for researchers, businesses, governments, and the general public.

3. PROPOSAL OUTLINE

The following steps are proposed to start the process of preserving query logs from Web search engines.

1. Begin the collection and preserving of query logs that are currently publicly available.
2. Commence a cooperative partnership among academia, Web search engine companies, and governmental agencies, both national and globally, to establish a more systematic recording and preservation of query logs.
3. Investigate methods for de-identifying log data (e.g., removing transactional information about the user from content information).
4. Develop the architecture for long-term storage, management, and retrieval of content contained from these query logs.

4. DISCUSSION

Certainly preserving query logs has benefits, risks, and costs. Benefits include preserved records of human expression in an increasing important area of people's daily lives. One can use these records for a variety of social, commercial, and historical purposes. Certainly, there are risks with preserving these expressions. We need to do a careful analysis of who is exposed to risk by preserving these records and what are the mitigation approaches, including possibly "shelving" logs for a period of time. Finally, there are costs involved to the owners of the logs. We have to find ways to overcome these costs or provide incentives for preservation.

5. CONCLUSION

In this position paper, we propose that the query logs from Web search engines are an important cultural artifact that should be preserved. These logs will become even more valuable as historical archives as the Web expands into more geographically areas and economically diverse segments of the population. There are commerce and privacy consideration that must be addressed; however, there are significant possible benefits in having access to these historical records. We hope to begin the discussion of achieving the preservation of the collective expressions of the human consciousness.

6. REFERENCES

[1] Battelle, J. (2005). The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed our Culture. New York, Penguin Group.

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

[3] Sullivan, D. 2006. Nielsen / NetRatings Search Engine Ratings. *2006,* 1 June.