Internet Usage Transaction Log Studies: The Next Generation

Sponsored by SIG USE

Dietmar Wolfram, Moderator.

School of Information Studies, University of Wisconsin-Milwaukee Milwaukee, WI 53201 . dwolfram@uwm.edu

Bernard Jim Jansen.

School of Information Sciences and Technology Pennsylvania State University State College, PA 16803 . jjansen@ist.psu.edu

Soo Young Rieh.

School of Information, University of Michigan Ann Arbor, MI 48109-1092 . rieh@umich.edu

Amanda Spink.

School of Information Sciences, University of Pittsburgh Pittsburgh, PA 16260 . aspink@mail.sis.pitt.edu

Peiling Wang.

School of Information Sciences, University of Tennessee Knoxville, TN 37996-4330 . peilingw@utk.edu

Introduction

The availability of large transaction logs of Internet-based search tools has produced a wealth of research over the past ten years on search patterns for public search engines, vendor database products, and institutional search services. Much of the early research dealt with describing observed regularities in querying and browsing behaviors. Findings of these studies have revealed--with notable regularity across search tools--that users do not engage in lengthy search sessions, submit brief queries, and do not browse extensively. With many studies now having

uncovered the presence of search and browsing regularities, what are the next steps for Internet user research? Panel members will discuss developments on these fronts and research directions for the near future.

Standardization of variables studied– At present, different software collects different types of data and researchers define search process events differently (e.g., session, query term, query modification).
Bernard Jim Jansen.

In a critique of transaction log analysis (TLA), Kurth (1993) identifies three methodological issues: execution, conception, and communication. TLA can be difficult to execute due to collection, storage and analysis issues associated with the hefty volume and complexity of the data set (i.e., significant number of variables). These limitations will be discussed. With complex data sets, it is sometimes difficult to develop a methodology for analyzing dependent variables. Communication problems occur when researchers do not define terms and metrics in sufficient detail to allow other researchers to interpret and verify their results; however, this is an issue with many methodologies. The focus on variables and metrics addresses the second and third (i.e., conception and communication) of Kurth's three methodological issues.

Additionally, panelist Jansen will address transaction log software limitations. There are two shortcomings in this area. First, transaction logs are primarily a server-side data collection method; therefore, some interactions events are masked from these logging mechanisms, such as when the user clicks on the back or print button on the browser software, or cuts or pastes information from one window to another on a client computer. Second, transaction logs do not record the underlying situational, cognitive, or affective elements of the searching process.

Integration of information needs and seeking theory - Transaction log analysis studies are now moving beyond descriptive analysis and are integrating theories of information needs and seeking for interactive IR. How well do existing theoretical models fit in this environment?
Soo Young Rieh, Peiling Wang.

Panelist Wang will discuss how findings of longitudinal Web log studies have challenged current IR interactions developed in the pre-Internet environment. A new approach to effective interactions should focus on (1) system's ability to understand user needs at a conceptual level, and (2) interactive Web design to incorporate changes.

For a conceptual match between a searcher's needs and relevant Web information objects, effective tools must be developed to facilitate representations of knowledge and cognitive structures. If a system is capable of understanding user needs at a conceptual level, not merely user queries at a symbolic level, the same query from different searchers may be processed quite differently based on their cognitive structures. Interactive Web design means incorporating research results into system improvements and redesign.

Mining longitudinal query data can provide a better understanding of user needs and changes in search behavior. (Wang, Barry, & Yang, 2003). Such analysis should be an integral part of Web search engines. Although today's search engines capture behavioral data such as queries and clicks, analytical and visual tools are not available. This represents one type of gap between theoretical research and real-world practice. In a well-defined context, user needs and behavioral changes are situational, thus often predictable to certain extent. As an example, the investigator’s university changed its registration system from several optional methods (phone, mail-in, etc.) to a Web-based online system. Queries related to the registration system have soared to become among the most frequently occurring queries. The difficulties for students to find the registration Webpage should be anticipated and the queries treated efficiently to produce a high precision search result.

Panelist Rieh will discuss how Interactive IR models provide a useful theoretical framework for analyzing transaction logs and interpreting findings, as they offer insights into which logs are considered to be the product of user interaction with the system. To better understand query reformulation behaviors, for instance, diverse patterns can be identified by examining search logs in terms of interactive and iterative processes rather than merely calculating frequency of reformulation within a search session. In addition to providing simple usage data, transaction logs can offer the most naturally occurring behavioral data that can be used as the basis for evaluating the success of IR systems.

Two case studies of the Excite search engine will be presented in which the findings of log analysis were directly translated into developing a new search tool or redesigning a search interface. The first case study investigated the patterns of multiple query reformulations, focusing on the reformulation sequences (Rieh & Xie, in press). The results of this study demonstrated

that innovative search tools were needed to support dynamic and complex patterns of query reformulation. The second case study was conducted in order to decide whether to redesign Excite's Advanced Search page. The log analysis centered around addressing three research questions: (1) At what point do users decide to go to the Advanced Search (AS) page?; (2) To what extent is each AS feature used?; and, (3) after using the AS page, what subsequent actions are taken by users? The findings identified the weaknesses of the existing AS design and provided directions for the redesign.

Integration of study findings into search practice and system design – How can findings best be translated into practice, whether for system design or service development and how can researchers and practitioners work with developers to ultimately benefit users?
Amanda Spink

Panelists Spink and Jansen have been collaborating with various Web search engine companies since 1997, including Excite, AlltheWeb.com, Alta Vista, Ask Jeeves, Vivisimo.com and InfoSpace, Inc. This ongoing collaborative research has produced a large dataset of millions of Web search sessions from 1997 to 2005, longitudinal, trend and usability studies in over 80 publications, and the book Web Search: Public Searching of the Web (Spink & Jansen, 2004). The goals of the ongoing collaborative research include: 1) further longitudinal and trends analysis of transaction log analysis, 2) usability studies, 3) development of theories and user models, and 4) integration of study findings into insights and recommendations for the Web industry.

Several aspects of academic researcher collaboration with Web industry companies will be discussed. Web companies have diverse approaches to research; some have in-house research, some collaborate with academic researchers, and some conduct limited or no research. There is a growing interest by Web companies in understanding their users and conducting user-based studies. Therefore, an opportunity exists for academic researchers to collaborate with the Web industry. However, such collaborations are grounded in the academic researcher's ability to conduct studies that provide valuable insights and recommendations for Web companies to develop their competitive advantage. Academic-Web industry collaboration is often a difficult and challenging process due to Web companies' competitive and privacy concerns, and diverse academic and business goals.

Generalization of observed behaviors – What do observed regularities and power laws tell us about user audiences?
Dietmar Wolfram

A number of authors in recent years have pointed out that observed patterns of electronic content and use follow inverse power laws, exemplified by Zipf's Law and Lotka's Law. Most usage characteristics when observed on a large scale do exhibit inverse patterns, where many instances of an observed characteristic occur only one time, and few instances of the characteristic occur many times (e.g., the number of queries per session, the number of pages viewed per query, the frequency with which queries are submitted across users, resource visitation by given identifiers). There are some exceptions to this such as the number of words used in a query, where the most frequent values are two or more terms per query. These observations indicate that users do not put much effort into their searching; but is this truly the case? As outlined in 1) above, transaction logs capture objective search processes, but on their own do not tell us things such as user motivation. These data may allow researchers to identify large-scale groups of users based on search behavior, for which different search mechanisms or interfaces may be supported.

For describing overall behaviors, methods with little rigor such as visual inspection have been used to confirm conformance of data sets to power laws. Departures between observed and expected results are dismissed as data anomalies. For descriptive purposes, visual inspection may be sufficient, but for modeling purposes for development of user simulations, for example, the use of more rigorous forms of data fitting techniques beyond visual inspection is needed. Panelist Wolfram will discuss how fitting of large observed data sets can be particularly challenging given the shortcomings of some model fitting techniques, the more elaborate functions that may be needed for optimal modeling (Wolfram, 2003), and the possibility that simple models based on power laws may not be the most appropriate for applications where high precision in modeling is needed.

References

Kurth, M. (1993). The limits and limitations of transaction log analysis. *Library Hi Tech, 11*(2), 98-104.

Rieh, S. Y. & Xie, H. (in press). Analysis of Multiple Query Reformulations on the Web: The Interactive Information Retrieval Context. *Information Processing & Management*.

Spink, A., & Jansen, B. J. (2004). *Web Search: Public Searching of the Web*. Dordrecht: Kluwer.

Wang, P. Berry, M. W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and

patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 743-758.

Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, CT: Libraries Unlimited.