Inf Retrieval (2010) 13:188–191
DOI 10.1007/s10791-009-9114-z

BOOK REVIEW

# Bernard J. Jansen, Amanda Spink, Isak Taksa: Handbook of research on web log analysis

## IGI Global, 2009, 628 pp, $265, ISBN: 978-1-59904-974-8

**Georg Buscher**

Published online: 10 September 2009


The modern world without the internet is hardly imaginable. Not only the size of the Web and the diversity of services provided there are constantly increasing, but also the number of people using the Web is shooting up. According to Internet World Stats (2009), the number of Web users worldwide increased by over 300% during the past 8 years, reaching over 1.5 billion internet users in 2009. And while surfing, every single one of those users is leaving a trace! Those traces, captured in logs of various kinds, can be extremely useful to service and content providers and researchers, e.g., for search engine improvement, for predicting Web traffic patterns and growth, and for understanding user intents and needs, etc.

The *Handbook of Research on Web Log Analysis* is a 25-chapter compilation of research papers from a variety of different authors and covers a broad range of issues concerning Web log analysis. It is composed of 5 sections: (1) a higher-level view on Web log analysis such as its development over time and critiques concerning privacy issues; (2) an overview of methodologies and metrics for analyzing log files; (3) approaches and tips for behavior analysis of single users or groups of users; (4) a section about the analysis of search query logs in specific; and (5) a final section about a variety of specialized analysis approaches.

Before going through some examples of contributed chapters, it is important to notice that the book intends to include research about the dynamically evolving field of Web log analysis in a great variety. It includes some chapters addressing researchers in the field of information retrieval who might be interested in query logs and browsing behavior logs as sources for relevance feedback and personalization. Yet, the book also comprises chapters addressing Website owners who want to analyze user traffic on their server, who like to know how their Web pages are used, and who might be interested in optimizing revenue through advertisements in a commercial setting. Some chapters address a third type of audience who is interested in analyzing user-generated content on highly dynamic Web pages like blogs, forums, and social networking sites. Especially the chapters about analyzing user-generated content on dynamic Web pages do not deal with analyzing client- or

G. Buscher (✉)
German Research Center for Artificial Intelligence (DFKI GmbH),
Knowledge Management Lab, Kaiserslautern, Germany
e-mail: georg.buscher@dfki.de

server-side log files containing information about user–computer interaction. They rather view the Web pages itself as log files of user–user interaction (e.g., a thread in a discussion forum) and analyze those. Finally, more higher-level chapters, e.g., about the history of logging user interaction on the Web and about privacy concerns are or should be of general interest for all researchers working on analyzing log files.

In the following, a few example chapters from each section of the book are shortly highlighted. As summarized in the end of this review, they all have different characteristics concerning level of detail, purpose, addressed audience, etc., and should help the reader of this review to get a better feeling of the structure and some characteristics of the book.

Section I of the book, "Web Log Analysis: Perspective, Issues, and Directions" contains four chapters providing higher-level views and thoughts about Web log analysis. As an example, a chapter about a "Historic Perspective of Log Analysis" by Penniman reviews the birth and evolution of transaction log analysis. For me as a researcher who joined information retrieval research only after the Internet was well established, this chapter provided interesting facts that explain why logging on the Web is how it is today. It is clear that analyzing log files is not an ideal and complete way of understanding user intents, which is one of the use cases for them today. Rather, log files were initially only used to describe and predict Web traffic on servers in order to make decisions with respect to network infrastructure. But from the beginning on, they also raised serious privacy concerns.

Section II about "Methodology and Metrics" consists of five chapters describing methodological approaches of interpreting log files. A chapter about the "Methodology of Search Log Analysis" by Jansen provides basic step-by-step instructions for simple search log analysis (data cleaning, parsing, normalization, analysis at different levels). The provided overview is very helpful for researchers starting in that field, because it conveys practical information and tips especially concerning the absolutely necessary data preparation process. However, with respect to analyzing the data, it does not go very deep.

A different chapter in this section deals with "Uses, Limitations, and Trends in Web Analytics" (by Ferrini and Mohr). It basically gives an overview of state-of-the-art Website traffic analysis tools such as Google Analytics for example. Additionally, it points out problems that arise with highly dynamic Web2.0 applications where interactions are difficult to track. This chapter is certainly more interesting for Website owners than for information retrieval researchers.

As a final example for this section, chapter "Recommendations for Reporting Web Usage Studies" by Hawkey and Kellar makes a very important point concerning reporting research about Web usage in general. The authors argue that the Web is highly dynamic and that the user population drastically changes over time. For that reason, conclusions from Web usage research from, e.g., 10 years ago might not be valid for today's Web users anymore. To help future researchers interpreting results from the past, Hawkey and Kellar give very useful recommendations how to report results from Web user studies in context. I recommend every researcher publishing in the field to be aware of the issues raised in this chapter.

Section III entitled "Behavior Analysis" comprises five chapters about analyzing and tracking user behavior. The chapters are very heterogeneous concerning their approaches to tracking behavior and also concerning their understandings of user behavior. For example, "From Analysis to Estimation of User Behavior" by Ozmutlu, Ozmutlu, and Spink views user behavior as the change of topics in Web search logs over time. Most of the chapter elaborates on a variety of different methods that could be used for topic detection and tracking based on queries (e.g., Dempster-Shafer theory, support vector machines, Markov models, etc.). Yet, the chapter stays mostly at a very detailed level and does not provide a more understandable "bigger picture".

Chapter "Tips for Tracking Web Information Behavior" by Detlor, Hupfer, and Ruhi provide a much broader conceptual view with respect to behavior analysis. To truly understand user behavior and intention, one eventually has to understand cognitive processes of the user. Of course, this is hardly ever to reach, but in order to go in that direction, the authors provide very useful tips for collecting as much information about Web interactions of users as possible. For example, they show how to combine client- and server-side logging each of which provides complementary information. This is especially inspirational for researchers planning future Web usage studies.

Section IV deals with "Query Log Analysis" and contains 5 chapters mostly addressing researchers in the field of information retrieval. The first chapter in this section entitled "Machine Learning Approach to Search Query Classification" by Taksa, Zelikovitz, and Spink has the unusual aim of classifying the age of the searcher issuing a query. They try to achieve this by detecting the topic of the query and then by assigning those topics of interest to certain age groups. This is an inspirational and inventive idea. Yet, the chapter has more the character of a very fresh research paper and the approach does not seem to work too well (which is understandable given the great variety of topics and queries searchers usually use (Spink and Jansen 2004). Furthermore, except that the analyzed collection of queries was extracted from query logs, the chapter has relatively little to do with analyzing query logs.

As a different example of a chapter in this section, "Query Log Analysis in Biomedicine" by Bernstam, Herskovic, and Hersh is a very inspiring paper describing how semantic technologies can be used for query log analysis in the medical domain. For example, they discuss how to detect session boundaries by estimating semantic distances between queries, how to determine broadening or narrowing information needs for a user within a search session, or how to detect query intent (navigational vs. informational). The paper points out very important directions for future research with respect to including semantic technologies in query log analysis and search and was one of the book's highlights for me personally.

The last Section V entitled "Contextual and Specialized Analysis" contains 5 chapters spanning a variety of different topics. As one example from this section, chapter "Information Extraction from Blogs" by Moens provides a comprehensive overview of how information extraction algorithms work and focuses on Blogs in specific. Blogs are viewed as log files of user–user interaction which can be analyzed in order to detect popular topics, detect changes in topics, mine opinions, etc. This chapter is exemplary for most of the chapters in this section in that it does not deal with server-side or client-side log files about user–computer interaction, but rather with user-generated content directly accessible on Websites.

Overall, the chapters contained in the book are of greatly varying level of detail. Some of them provide high-level and abstract overviews and discussions; some comprise concrete and practical overviews and collections of tips and tricks; some are very detailed and seem to be reports about late-breaking and very specific research that you would rather expect to see at a research conference. Furthermore, there is a considerable amount of redundancy across the different chapters. Almost all of the chapters are self-contained, explaining the basics methods and pros and cons of different techniques repeatedly. There is also no specific order in which a reader should go through the book. Noticeably, the self-containedness makes it easier and convenient to read just some single chapters that are of personal interest. Taking this into account together with the fact that different audiences are addressed by the different chapters, it is very probable that the book contains some interesting chapters for everyone working in the field (i.e., information retrieval

researchers, Website owners, and researchers of ethnography and social network analysis). However, as a consequence, most of the other chapters might not be too interesting for non-interdisciplinary readers.

According to the editors, the book is "an essential holding for library reference collections" and "will be of value to faculty seeking an advanced textbook in the field of log analysis, and researchers and practitioners looking for answers to consistently evolving theoretical and practical challenges". Given the structure of the book, it will indeed be of value to a great variety of mostly researchers, but also practitioners, as a resource that spans a very broad range of topics around Web log analysis. Especially for students in their early stages of research it might be very inspirational to go through the chapters of the book, reading about the variety of used methods, concepts, ideas, and techniques around Web log analysis.

## References

Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the web*. Berlin: Springer.
Miniwatts Marketing Group. (2009). *World internet usage and population statistics*. Internet World Stats at http://www.internetworldstats.com/stats.htm, August 14, 2009.