Preface

Web use has become a ubiquitous online activity for people of all ages, cultures and pursuits. Whether searching, shopping, or socializing users leave behind a great deal of data revealing their information needs, mindset, and approaches used. Web designers collect these artifacts in a variety of Web logs for subsequent analysis. *The Handbook of Research on Web Log Analysis* reflects on the multifaceted themes of Web use and presents various approaches to log analysis. The handbook looks at the history of Web log analysis and examines new trends including the issues of privacy, social interaction and community building. It focuses on analysis of the user's behavior during the Web activities, and investigates current methodologies and metrics for Web log analysis. The handbook proposes new research directions and novel applications of existing knowledge. The handbook includes 25 chapters in five sections, contributed by a great variety of researchers and practitioners in the field of Web log analysis.

Chapter I "Research and Methodological Foundations of Transaction Log Analysis" by Bernard J. Jansen (Pennsylvania State University, USA), Isak Taksa (Baruch College, City University of New York, USA), Amanda Spink (Queensland University of Technology, Australia), introduces, outlines and discusses theoretical and methodological foundations for transaction log analysis. The chapter addresses the fundamentals of transaction log analysis from a research viewpoint and the concept of transaction logs as a data collection technique from the perspective of behaviorism. It continues with the methodological aspects of transaction log analysis and examines the strengths and limitations of transaction logs as trace data. It reviews the conceptualization of transaction log analysis as an unobtrusive approach to research, and presents the power and deficiency of the unobtrusive methodological concept, including benefits and risks of transaction log analysis specifically from the perspective of an unobtrusive method. Some of the ethical questions concerning the collection of data via transaction log application are discussed.

Section I, *Web Log Analysis: Perspectives, Issues, and Directions* consists of four chapters presenting a historic perspective of web log analysis, examining surveys as a complementary method for transaction log analysis, and investigating issues of privacy and traffic measurement.

Chapter II "Historic Perspective of Log Analysis" by W. David Penniman (Nylink, USA), provides a historical review of the birth and evolution of transaction log analysis applied to information retrieval systems. It offers a detailed discussion of the early work in this area and explains how this work has migrated into the evaluation of Web usage. The author describes the techniques and studies in the early years and makes suggestions for how that knowledge can be applied to current and future studies. A discussion of privacy issues with a framework for addressing the same is presented, as well as an overview of the historical "eras" of transaction log analysis.

Chapter III "Surveys as a Complementary Method for Web Log Analysis" by Lee Rainie (Pew Internet & American Life Project, USA), Bernard J. Jansen (Pennsylvania State University, USA) examines surveys as a viable complementary method for transaction log analysis. It presents a brief overview of survey research literature, with a focus on the use of surveys for Web-related research. The authors

identify the steps in implementing survey research and designing a survey instrument. They conclude with a case study of a large electronic survey to illustrate what surveys in conjunction with transaction logs can bring to a research study.

Chapter IV "Watching the Web: An Ontological and Epistemological Critique of Web-Traffic Measurement" by Sam Ladner (York University, Canada), compares two dominant forms of Web-traffic measurement and discusses the implicit and largely unexamined ontological and epistemological claims of both methods. It suggests that like all research methods, Web-traffic measurement has implicit ontological and epistemological assumptions embedded within it. An ontology determines what a researcher is able to discover, irrespective of method, because it provides a framework within which phenomena can be rendered intelligible.

Chapter V "Privacy Concerns for Web Logging Data" by Kirstie Hawkey (University of British Columbia, Canada) examines two aspects of privacy that must be considered when conducting studies of user behavior that includes the collection of web logging data. First considered are the standard privacy concerns when dealing with participant data. These include privacy implications of releasing the data, methods of safeguarding the data, and issues encountered with re-use of data. Second, the impact of data collection techniques on the researchers' ability to capture natural user behaviors is discussed. Key recommendations are offered about how to enhance participant privacy when collecting Web logging data to encourage these natural behaviors.

Section II, *Methodology and Metrics*, consists of five chapters reviewing the foundations, trends and limitations of available and prospective methodologies, examining granularity and validity of log data, and recommending context for future log studies.

Chapter VI "The Methodology of Search Log Analysis" by Bernard J. Jansen (Pennsylvania State University, USA) presents a review of and foundation for conducting Web search transaction log analysis. A search log analysis methodology is outlined consisting of three stages (i.e., collection, preparation, and analysis). The three stages of the methodology are presented in detail with discussions of the goals, metrics, and processes at each stage. The critical terms in transaction log analysis for Web searching are defined. Suggestions are provided on ways to leverage the strengths and addressing the limitations of transaction log analysis for Web searching research.

Chapter VII "Uses, Limitations, and Trends in Web Analytics" by Tony Ferrini (Acquiremarketing.com, USA), Jakki J. Mohr (University of Montana, USA), emphasizes the importance of measuring the performance of a Website. The measuring includes tracking the traffic (number of visitors), visitors' activity and behavior while visiting the site. The authors examine various uses of Web Metrics (how to collect Web log files) and Web analytics (how Web log files are used to measure a Website's performance), as well as the limitations of these analytics. The authors also propose options for overcoming these limitations, new trends in Web analytics, including the integration of technology and marketing techniques, and challenges posed by new Web 2.0 technologies.

Chapter VIII "A Review of Methodologies for Analyzing Websites" by Danielle Booth (Pennsylvania State University, USA), Bernard J. Jansen, (Pennsylvania State University, USA) provides an overview of the process of Web analytics for Websites. It outlines how basic visitor information such as number of visitors and visit duration can be collected using log files and page tagging. This basic information is then combined to create meaningful key performance indicators that are tailored not only to the business goals of the company running the Website, but also to the goals and content of the Website. Finally, this chapter presents several analytic tools and explains how to choose the right tool for the needs of the Website. The ultimate goal of this chapter is to provide methods for increasing revenue and customer satisfaction through careful analysis of visitor interaction with a Website.

Chapter IX "The Unit of Analysis and the Validity of Web Log Data" by Gi Woong Yun (Bowling Green State University, USA), discusses challenges and limitations in defining units of analysis of Web

site use. The author maintains that unit of analysis depends on the research topic and level of analysis, and therefore is complicated to predict ahead of data collection. Additionally, technical specifications of the Web log data sometimes limit what researchers can select as a unit of analysis for their research. The author also examines the validity of data collection and interpretation processes as well as sources of such data. The chapter concludes with proposed criteria for defining units of analysis of a Web site and measures for improving and authenticating validity of web log data.

Chapter X "Recommendations for Reporting Web Usage Studies" by Kirstie Hawkey (University of British Columbia, Canada), Melanie Kellar (Google Inc., USA), presents recommendations for reporting context in studies of Web usage including Web browsing behavior. These recommendations consist of eight categories of contextual information crucial to the reporting of results: user characteristics, temporal information, Web browsing environment, nature of the Web browsing task, data collection methods, descriptive data reporting, statistical analysis, and results in the context of prior work. This chapter argues that the Web and its user population are constantly growing and evolving. This changing temporal context can make it difficult for researchers to evaluate previous work in the proper context, particularly when detailed information about the user population, experimental methodology, and results is not presented. The adoption of these recommendations will allow researchers in the area of Web browsing behavior to more easily replicate previous work, make comparisons between their current work and previous work, and build upon previous work to advance the field.

Section III, *Behavior Analysis*, consists of five chapters summarizing research in user behavior analysis during various web activities and suggesting directions for identifying, finding meaning and tracking user behavior.

Chapter XI "From Analysis to Estimation of User Behavior" by Seda Ozmutlu (Uludag University, Turkey), Huseyin C. Ozmutlu (Uludag University, Turkey), Amanda Spink (Queensland University of Technology, Australia), summarizes the progress of search engine user behavior analysis from search engine transaction log analysis to estimation of user behavior. Correct estimation of user information searching behavior paves the way to more successful and even personalized search engines. However, estimation of user behavior is not a simple task. It closely relates to natural language processing and human computer interaction, and requires preliminary analysis of user behavior and careful user profiling. This chapter details the studies performed on analysis and estimation of search engine user behavior, and surveys analytical methods that have been and can be used, and the challenges and research opportunities related to search engine user behavior or transaction log query analysis and estimation.

Chapter XII "An Integrated Approach to Interaction Design and Log Analysis" by Gheorghe Muresan (Microsoft Corporation, USA), describes and discusses a methodological framework that integrates analysis of interaction logs with the conceptual design of the user interaction. It is based on (1) formalizing the functionality that is supported by an interactive system and the valid interactions that can take place; (2) deriving schemas for capturing the interactions in activity logs; (3) deriving log parsers that reveal the system states and the state transitions that took place during the interaction; and (4) analyzing the user activities and the system's state transitions in order to describe the user interaction or to test some research hypotheses. This approach is particularly useful for studying user behavior when using highly interactive systems. Details of the methodology and examples of use in a mediated retrieval experiment are presented.

Chapter XIII "Tips for Tracking Web Information Seeking Behavior" by Brian Detlor (McMaster University, Canada), Maureen Hupfer (McMaster University, Canada), Umar Ruhi (University of Ottawa, Canada), provides various tips for practitioners and researchers who wish to track end-user Web information seeking behavior. These tips are derived in large part from the authors' own experience in collecting and analyzing individual differences, task, and Web tracking data to investigate people's online information seeking behaviors at a specific municipal community portal site (myhamilton.ca). The

tips discussed in this chapter include: (2) the need to account for both task and individual differences in any Web information seeking behavior analysis; (2) how to collect Web metrics through deployment of a unique ID that links individual differences, task, and Web tracking data together; (3) the types of Web log metrics to collect; (4) how to go about collecting and making sense of such metrics; and (5) the importance of addressing privacy concerns at the start of any collection of Web tracking information.

Chapter XIV "Identifying Users Stereotypes for Dynamic Web Pages Customization" by Sandro José Rigo, José Palazzo M. de Oliveira, Leandro Krug Wives, (Instituto de Informática, Universidade Federal do Rio Grande do Sul, Brazil), explores Adaptive Hypermedia as an effective approach to automatic personalization that overcomes the complexities and deficiencies of traditional Web systems in delivering user-relevant content. The chapter focuses on three important issues regarding Adaptive Hypermedia systems: the construction and maintenance of the user profile, the use of Semantic Web resources to describe Web applications, and implementation of adaptation mechanisms. Web Usage Mining, in this context, allows the discovery of Website access patterns. The chapter describes the possibilities of integration of these usage patterns with semantic knowledge obtained from domain ontology. Thus, it is possible to identify users' stereotypes for dynamic Web pages customization. This integration of semantic knowledge can provide personalization systems with better adaptation strategies.

Chapter XV "Finding Meaning in Online, Very-Large Scale Conversations" by Brian K. Smith, Priya Sharma, Kyu Yon Lim, Goknur Kaplan Akilli, KyoungNa Kim, Toru Fujimoto (Pennsylvania State University, USA), Paula Hooper (TERC, USA), provides understanding of how people come together to form virtual communities and how knowledge flows between participants over time. It examines ways to collect data and describes two methods–qualitative data analysis and Social Network Analysis (SNA)–which were used to analyze conversations within ESPN's *Fast Break* virtual community, which focuses on fantasy basketball sports games. Furthermore, the authors utilize the individual and community level analysis to examine individual reflection on game strategy and decision-making, as well as patterns of interactions between participants within the community.

Section IV, *Query Log Analysis*, consists of five chapters examining query classification and topic identification in search engines, analyzing queries in the biomedical domain and Chinese Information Retrieval, and presenting a comprehensive review of the research publications on query log analysis.

Chapter XVI "Machine Learning Approach to Search Query Classification" by Isak Taksa (Baruch College, City University of New York, USA), Sarah Zelikovitz (The College of Staten Island, City University of New York, USA), Amanda Spink (Queensland University of Technology, Australia), presents an approach to non-hierarchical classification of search queries that focuses on two specific areas of machine learning: short text classification and limited manual labeling. Typically, search queries are short, display little class specific information per single query and are therefore a weak source for traditional machine learning. To improve the effectiveness of the classification process the chapter introduces background knowledge discovery by using information retrieval techniques. The proposed approach is applied to a task of age classification of a corpus of queries from a commercial search engine. In the process, various classification scenarios are generated and executed, providing insight into choice, significance and range of tuning parameters.

Chapter XVII "Topic Analysis and Identification of Queries" by Seda Ozmutlu (Uludag University, Turkey), Huseyin C. Ozmutlu (Uludag University, Turkey), Amanda Spink (Queensland University of Technology, Australia), emphasizes topic analysis and identification of search engine user queries. Topic analysis and identification of queries is an important task related to the discipline of information retrieval, which is a key element for the development of successful personalized search engines. Topic identification of text is also no simple task, and a problem yet unsolved. The problem is even harder for search engine user queries due to real-time requirements and the limited number of terms in the user

queries. The chapter includes a detailed literature review on topic analysis and identification, with an emphasis on search engine user queries, a survey of the analytical methods that have been and can be used, and the challenges and research opportunities related to topic analysis and identification.

Chapter XVIII "Query Log Analysis in Biomedicine" by Elmer V. Bernstam (UT-Houston, USA), Jorge R. Herskovic (UT-Houston, USA), William R. Hersh (Oregon Health & Science University, USA), describes the purpose of query log analysis in the biomedical domain as well as features of the biomedical domain such as controlled vocabularies (ontologies) and existing infrastructure useful for query log analysis. The chapter focuses specifically on MEDLINE, which is the most comprehensive bibliographic database of the world's biomedical literature, the PubMed interface to MEDLINE, the Medical Subject Headings vocabulary and the Unified Medical Language System. However, the approaches discussed here can also be applied to other query logs. The chapter concludes with a look toward the future of biomedical query log analysis.

Chapter XIX "Processing and Analysis of Search Query Logs in Chinese", by Michael Chau (The University of Hong Kong, Hong Kong), Yan Lu (The University of Hong Kong, Hong Kong), Xiao Fang (The University of Toledo, USA), Christopher C. Yang (Drexel University, USA), argues that more non-English content is now available on the World Wide Web and the number of non-English users on the Web is increasing. While it is important to understand the Web searching behavior of these non-English users, many previous studies on Web query logs have focused on analyzing English search logs and their results may not be directly applied to other languages. This chapter discusses some methods and techniques that can be used to analyze search queries in Chinese language. The authors show an example of applying these methods to a Chinese Web search engine.

Chapter XX "Query Log Analysis for Adaptive Dialogue-Driven Search" by Udo Kruschwitz (University of Essex, UK), Nick Webb (SUNY Albany, USA), Richard Sutcliffe (University of Limerick, Ireland), presents an extensive review of the research publications on query log analysis and analyses two case studies, both aimed at improving Information Retrieval and Question Answering systems. The first describes an intranet search engine that offers sophisticated query modifications to the user. It does this via a hierarchical domain model that was built using multi-word term co-occurrence data. The usage log is analyzed using mutual information scores between a query and its refinement, between a query and its replacement, and between two queries occurring in the same session. The second case study describes a dialogue-based Question Answering system working over a closed document collection largely derived from the Web. Logs are based around explicit sessions in which an analyst interacts with the system. Analysis of the logs has shown that certain types of interaction lead to increased precision of the results.

Section V, *Contextual and Specialized Analysis*, consists of four chapters presenting a conceptual framework for transaction log analysis, proposing a new theoretical model for evaluating connector websites that facilitate online social networks, introducing information extraction from blog texts, and exploring the use of nethnography in the study of computer-mediated communication (CMC).

Chapter XXI "Using Action-Object Pairs as a Conceptual Framework for Transaction Log Analysis" by Mimi Zhang (Pennsylvania State University, USA), Bernard J. Jansen (Pennsylvania State University, USA), presents the action-object pair approach as a conceptual framework for transaction log analysis. The authors argue that there are two basic components in the interaction between the user and the system recorded in a transaction log, which are action and object. An action is a specific utterance of the user. An object is a self-contained information object, the receipt of the action. These two components form one interaction set or an action-object pair. A series of action-object pairs represents the interaction session. The action-object pair approach provides a conceptual framework for the collection, analysis, and understanding of data from transaction logs. The authors suggest that this approach can benefit system design by providing the implicit feedback concerning the user and delivering, for example, personalized service

to the user based on this feedback. Action–object pairs also provide a worthwhile approach to advance the theoretical and conceptual understanding of transaction log analysis as a research method.

Chapter XXII "Analysis and Evaluation of the Connector Website" by Paul DiPerna (The Blau Exchange Project, USA), proposes a new theoretical model for evaluating websites that facilitate online social networks. The suggested model considers previous academic work related to social networks and online communities. This study's main purpose is to define a new kind of social institution, called a "connector website", and provide a means for objectively analyzing web-based organizations that empower users to form online social networks. Several statistical approaches are used to gauge website-level growth, trend lines, and volatility. This project sets out to determine whether particular connector websites can be mechanisms for social change, and to quantify the nature of the observed social change. The author hopes this chapter introduces new applications for Web log analysis by evaluating connector websites and their organizations.

Chapter XXIII "Information Extraction from Blogs" by Marie-Francine Moens (Katholieke Universiteit Leuven, Belgium), introduces information extraction from blog texts. It argues that the classical techniques for information extraction that are commonly used for mining well-formed texts lose some of their validity in the context of blogs. This finding is demonstrated by considering each step in the information extraction process and by illustrating this problem in different applications. In order to tackle the problem of mining content from blogs, algorithms are developed that combine different sources of evidence in the most flexible way. The chapter concludes with ideas for future research.

Chapter XXIV "Nethnography: A Naturalistic Approach Towards Online Interaction" by Adriana Andrade Braga (Pontifícia Universidade Católica do Rio de Janeiro), explores the possibilities and limitations of nethnography, an ethnographic approach applied to the study of online interactions, particularly computer-mediated communication (CMC). The chapter presents a brief history of ethnography, including its relation to anthropological theories and its key methodological assumptions. The presentation focuses on common methodologies that treat log files as the only or main source of data and discusses results of such an approach. In addition, it examines some strategies related to a naturalistic perspective of data analysis. Finally, to illustrate the potential for nethnography to enhance the study of CMC, the authors present an example of an ethnographic study.

Finally, Chapter XXV "Web Log Analysis: Diversity of Research Methodologies" by Isak Taksa (Baruch College, City University of New York, USA), Amanda Spink (Queensland University of Technology, Australia), and Bernard J. Jansen (Pennsylvania State University) focuses on the innovative character of Web log analysis and the emergence of its new applications. Web log analysis is the subject of many distinctive and diverse research methodologies due to its interdisciplinary nature and the diversity of issues it addresses. This chapter examines research methodologies used by contributing authors in preparing the individual chapters for this handbook, summarizes research results, and proposes new directions for future research in this area.

The Handbook of Research on Web Log Analysis with its full spectrum of topics, styles of presentation and depth of coverage 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.

Bernard J. Jansen, Amanda Spink, and Isak Taksa
Editors