Chapter XXV
Web Log Analysis:
Diversity Of Research Methodologies

Isak Taksa
City University of New York, USA

Amanda Spink
Queensland University of Technology, Australia

Bernard J. Jansen
Pennsylvania State University, USA

ABSTRACT

Web log analysis is an innovative and unique field constantly formed and changed by the convergence of various emerging Web technologies. Due to its interdisciplinary character, the diversity of issues it addresses, and the variety and number of Web applications, it is the subject of many distinctive and diverse research methodologies. 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.

INTRODUCTION

Introduced less than twenty years ago, the Web has become the environment where people of all ages, languages and cultures conduct their daily digital lives. Working or entertaining, learning or socializing, home or on the road, individually or as a group, Web users are ubiquitously surrounded by an infrastructure of devices, networks and applications. This infrastructure combined with the perpetually growing amount of every imaginable type of information supports the user's intellectual or physical activity. Whether searching, using or creating and disseminating the information, users leave behind a great deal of data revealing their information needs, attitudes, personal and environmental facts. Web designers collect these artifacts in a variety of Web logs for subsequent analysis.

The Handbook of Web Log Analysis reflects on the multifaceted themes of Web use and demonstrates an equally diverse range of research methodologies. The next section briefly reviews research methodologies applied by contributing authors. Subsequent sections report research results obtained using these methodologies, and propose directions for future research in the field of Web log analysis.

RESEARCH METHODOLOGIES

What are the research methodologies frequently applied in Web-based research? Some researchers focus on collection and preparation of information for data analysis (Jansen, 2006), while others concentrate on elicitation; reduction and visualization for user-profiling (Romano et al., 2003). Researchers also benefit from a new, aggressively growing source of personal communication – blogs (Jing, 2006; Rossler, 2002).

In a different direction, there are a number of studies that focus on analysis of research methodologies. Powel (1999) uses a comprehensive classification developed by Kim (1996) to review, define and discuss quantitatively and qualitatively-driven methodologies. Another publication (Palvia et al., 2007) provided a slightly different but equally comprehensive classification of research methodologies. Using these three sources, we identified the following methodologies used by this handbook's authors:

- **Conceptual Framework / Inquiry:** Concepts are introduced and defined, and subsequently used to construct conceptual frameworks that provide study directions.
- **Phenomenology / Ethnomethodology:** An interpretive methodology that examines users' behavior. Ethnomethodology, an extension of phenomenology, examines individual and group interactions within a social structure.
- **Content Analysis:** A methodical and replicable methodology used to determine, quantify, and analyze the presence of research objects within a large data set.
- **Ethnography:** A qualitative study in which the researcher observes members of a chosen group in a natural environment over a long period of time.
- **Historical Method:** Collects and examines facts about events, people and the environment of the past.
- **Discourse Analysis:** A scientific argument evaluation method.
- **Case Study:** A comprehensive study of a single subject, influenced by a proper selection of unit of analysis.

CONCEPTUAL FRAMEWORK / INQUIRY

Many research studies clearly specify and explain the methodologies used to describe or explain the subject under study. These studies usually introduce a set of concepts related to an existing (or future systems), or to a set of objects, or to behavior aspects of participants. Concepts are then used to construct conceptual frameworks, which provide the plan, purpose and direction for the study. Depending on the goals, data and technology, the conceptual frameworks offer a choice of methodologies: surveys, data analysis, literature review or many others. The conceptual frameworks methodology is widely used in many Web studies including information retrieval (Jansen 2006; Jansen et al, 2000), Web log analysis in e-commerce (Meersman et al., 2003), education and library studies (Nicholson 2004; Vrana, 2002).

Transaction Log Analysis

Transaction log analysis is a broad category of methods used for macro and micro analysis of transaction logs - electronic records of interactions that have occurred between a system and users of that system. Among others, these methods include Web log analysis, (i.e., analysis of Web system logs), blog analysis (i.e., analysis of Web blogs) and search log analysis (i.e., analysis of search engine logs).

Chapter I "Research and Methodological Foundations of Transaction Log Analysis" introduces, outlines and discusses the 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. The chapter 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 both 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. The chapter concludes with some essential ethical and legal questions: use of the logs for research, ownership, and user consent and access control.

Complementing the Web Log Analysis Methodology

Whether to validate or to substantiate existing research results, or to gain a new perspective on an existing issue, researchers often need complementary sources and methodologies to collect subject data.

Chapter III "Surveys as a Complementary Method to Web Log Analysis" examines surveys as a viable complementary method for transaction log analysis. The chapter presents a brief overview of survey research literature, with a focus on the use of surveys for Web-related research. It continues with a comprehensive overview of a 10 - step process to plan and conduct a survey and a comprehensive guide to designing a survey instrument. To illustrate the benefits of a survey in conjunction with transaction logs, a case study (including data analysis) of a large electronic survey is presented. The chapter concludes by stressing complementary capabilities of a survey specifically in the areas of understanding the underlying motivations, affective characteristics, cognitive factors, and contextual aspects that influence user behavior.

Search Logs Analysis

The data stored in search logs of Web search engines, Intranets, and Websites provides important insights into understanding the information searching tactics of online searchers. This understanding can assist information system designers and interface developers.

Chapter VI "The Methodology of Search Log Analysis" presents a review of, and foundation for, conducting Web search transaction log analysis. A search log analysis methodology is outlined consisting of three stages: collection (the process of collecting the interaction data for a given period in a transaction log including *User Identification, Date, Time, and Search query terms*), preparation (the process of preparing the transaction log data for analysis including *cleaning, parsing and normalizing*), and analysis (the process of analyzing the prepared data including *Term, Query, and Session level analysis*). The chapter continues with possible venues for analysis and concludes with suggestions for further research consisting of unobtrusive data collection to preserve the unaltered behavior of searchers, use of cookies to identify individual sessions, and use of survey data to get reasonable estimations of needed demographic data.

Website Analytics

Operational Website management necessitates a way to track and measure visitors' traffic, visitors' behavior and even more importantly how this behavior compares to the expected behavior.

Chapter VII "Uses, Limitations, and Trends in Web Analytics" focuses on measuring the performance of a Website. The measuring includes tracking the traffic (number of visitors), and visitors' activity and behavior while visiting the site. The chapter discusses current methodologies to log data and evaluate Website performance, stressing the limitations of log file analysis (e.g. lack of personal information, missing duration of the visit); clarify new techniques (e.g. site overlay, Geo-mapping) in Web analytics that supplement traditional log file analysis; and analyze trends in Web analytics as related to Web 2.0 technologies (social networking, tagging, blogging). As part of the Web 2.0 discussion the authors touch on the issue of "long tail". The chapter is concluded with suggestions to improve the accuracy of existing metrics (by using cookies and page tagging), and identify the need for a new set of metrics and analytics for the Web 2.0.

Website Key Performance Indicators

Web analytics studies visitor behavior on a Website. By collecting various Web analytics metrics one can develop key performance indicators (KPIs) – a versatile analytic model that measures visitor trends.

Chapter VIII "A Review of Methodologies for Analyzing Websites" provides an overview of the process of Web analytics. The chapter outlines how basic visitor information such as number of visits, number of visitors and visit duration can be collected through the use of log files and page tagging. This basic information is then combined to create important 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. First, the authors discuss the metrics that can be collected from the Website visitor, its types and potential uses. Then they analyze the two primary methods for gathering visitor information - log files and page tagging, detailing advantages and disadvantages of each method, and enumerating types of support information, and examples of data format. Once the data is collected, the selection and construction of KPIs is discussed and followed up by a description of the entire process with advice for Web analytics integration. The chapter is concluded with suggestions on what to look for when choosing analytics tools, as well as a comparison of several specific tools.

Action-Object Pairs

There are two basic components in the interaction between the user and the system that are recorded in a transaction log, they are action and object. An action is a specific expression of the user. An object is a self-contained information object. These two components form one interaction set or an action-object pair. A series of action-object pairs represents the interaction session.

Chapter XXI "Using Action-Object Pairs as a Conceptual Framework for Transaction Log Analysis" presents the action-object pair approach as a conceptual framework for three major steps of log analysis: the i) collection, ii) analysis, and iii) understanding of data from transaction logs. The authors present the scientific foundation and provide a detailed description of the proposed concept, and use the above three steps to illustrate the concept's applicability. The chapter is concluded with several case studies using the action-object pair approach. The studies illustrate the benefits of the approach and also how it facilitated the system performance. The authors suggest ways of using this approach to answer many questions still facing researchers involved in transaction log analysis.

PHENOMENOLOGY / ETHNOMETHODOLOGY

The design and acceptance of information systems is usually determined by a dichotomy between technology and behavior. While some approaches stress technological advance, others focus on users' behavior (Verbeek & Slob, 2006). Phenomenology is an interpretive methodology that examines users' behavior. It examines events and actions by which individual users give meaning to, and make sense of, interactions with technology (Budd, 2005). Phenomenology uses an interdisciplinary approach to investigate the reason, purpose and analysis of users' actions while searching for and deciding on the relevance of search results (Nicolas et al., 2007); it relies on technological mediation to explain amplification and reduction – "an increased capacity to engage with the world in a particular way, accompanied by a reduced capacity to engage with it in other ways" (Arnold, 2003; p. 240). It aims at supporting and improving the quality of interaction in an action-centric environment (Fernaeus, 2008). Ethnomethodology, an extension of phenomenology, examines individual and group interactions in transitory social structures created by on-demand connectivity (Westbrook, 2004).

Estimating 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.

Chapter XI "From Analysis to Estimation of User Behavior" details the studies performed on analysis and estimation of search engine user behavior, and surveys analytical methods that have been used to accomplish this task. The first part of the chapter is devoted to a review of existing search engine user behavior studies including multimedia, multitasking and e-commerce searches followed by detailed explanation of methodologies used for Web log and user behavior analysis including correlation and test of independence, Markov models, and Poisson sampling.

The second part follows the same process. First, the authors provide a detailed overview of studies estimating search engine user behavior including topic and session identification, and topic estimation, followed by detailed explanation of methodologies used for user behavior estimation including probabilistic and statistical methods, Monte-Carlo simulation and artificial intelligence methods. The chapter is concluded with specific ideas for further research such as the use of multivariate techniques to cluster user queries, the analysis of time-based behavior of users (seasons, holidays. etc.), and use of artificial intelligence and statistical learning methods for studying the content-based behavior.

Interaction Design for Studying User Behavior

A good understanding of people – what they are like, why they use a certain piece of software, and how they might interact with it – is essential for successful design of interactive systems, which help people achieve their goals.

Chapter XII "An Integrated Approach to Interaction Design and Log Analysis" describes a methodological framework that integrates analysis of interaction logs with the conceptual design of the user interaction. This approach is particularly useful for studying user behavior when using highly interactive systems. The author proposes a formal procedure that integrates the modeling of the interaction, the logging and the analysis of logged data. The procedure allows for capture of the functionality, states and user action in each state. When applied to a particular kind of interaction (such as interactive information retrieval), the proposed procedure can be used to investigate user behavior or to test the usability of a user interface. To demonstrate the capability of this procedure the author uses a case study of a MIR (Mediated Information Retrieval) project. The chapter is concluded with a comprehensive plan for future research, including studying patterns of behavior by building Hidden Markov Models (HMM) based on the analysis of state transitions recorded in the logs, granularity of hierarchical structure of states and visualization of user behavior.

Tips for Tracking Web User Behavior

Developing and employing Web tracking to better understand end-user experiences with the Web portal seems to be a simple process. However, setting up, collecting, and analyzing Web tracking data is surprisingly more difficult than originally expected.

Chapter XIII "Tips for Tracking Web Information Seeking Behavior" provides various tips for practitioners and researchers who wish to track end-user Web information seeking behavior. These tips are derived from the authors' own experience in collecting and analyzing individual differences, tasks, and Web tracking data to investigate people's online information seeking behaviors at a specific municipal community portal site (myhamilton.ca). The tips proposed and discussed in this chapter include: i) the need to account for both task and individual (learning and cognitive) differences in any Web information seeking behavior analysis; ii) how to collect Web metrics through deployment of a unique ID (a key strength of this research project) that links individual differences, task, and Web tracking data together; iii) the types of Web log metrics to collect (including raw metrics and composite analytics); iv) how to go about collecting and making sense of such metrics (visitor footprints, navigation tracks and information seeking trails); and v) the importance of addressing privacy concerns (including location? privacy legislation requirements and privacy impact assessment) at the start of any collection of Web tracking information. The chapter is complemented with an extensive questionnaire to assist portal developers in tracking users' behavior.

User Profiling for Dynamic Page Customization

Adaptive Hypermedia is an effective approach to establishing better user experience and delivering user relevant content.

Chapter XIV "Identifying Users Stereotypes for Dynamic Web Pages Customization" 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 major tasks regarding Adaptive Hypermedia systems:

i. The construction and maintenance of the user profile, which is achieved through integration of semantic information obtained from the Website domain ontology with usage information obtained from the data gathered from user sessions.
ii. The use of Semantic Web resources to describe Web applications (Universal Resource Identifier, Resource Description Framework, Ontology Web Language).
iii. Implementation of adaptation mechanisms (e.g. education, information retrieval and tourism). 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. To illustrate their approach the authors use an open source Web Content Management system to implement the Web usage data acquisition and to generate the structure adaptations.

Social Networks

The latest developments in Web design introduced a connector Website – a Website that facilitates social (or business) interactions among participants.

Chapter XXII "Analysis and Evaluation of the Connector Website" proposes a new theoretical model for evaluating Websites that facilitate online social networks. The chapter reviews previous academic work related to social networks and online communities, defines a new kind of social institution called a connector Website, and provides a brief history of several generations of the connector Websites. To conduct the analysis the author collected monthly Website traffic data from thirteen connector Websites and applied several statistical approaches to gauge Website-level growth, trend lines, and volatility. One interesting finding that the author is trying to explain is worth mentioning here "six connectors have produced rather unexpected social epidemics in terms of huge gains (or loss) in user traffic". The chapter is concluded with some lessons learned and their implications for future connector Websites. Important directions for further research are indicated, specifically, social values and tradeoffs, and differentiation and specialization of existing connector Websites.

CONTENT ANALYSIS

Content analysis is defined as a methodical and replicable methodology (Stemmer, 2001) used to i) determine presence of research objects within a large data set; ii) quantify and categorize the presence of the objects; iii) analyze the validity (Rourke & Anderson, 2004), reliability (Lombard et al.,2002) and significance of the obtained results for meaning and relationships of objects; and conjecture the demographics (age, gender), time and location of an activity that creates the object, and behavior patterns (needs, intent, attitude) of owners or creators of these objects (Krippendorf, 2004). Objects are usually defined as prearranged or derived terms, phrases or topics (in any language), and images. There are two basic types of content analysis: conceptual (looking to quantify and categorize objects) and semantic (looking to find and predict meaning) in the set of objects (Murphy & Ciszewska-Carr, 2005).

Query Classification

Usually search queries are very short (2 - 3 search terms), display little class specific information per single query and are therefore a weak source for machine learning an established tool for classification tasks.

Chapter XVI "Machine Learning Approach to Search Query Classification" presents a novel method of non-hierarchical classification of search queries that focuses on two specific areas of machine learning: short text classification and limited manual labeling. To improve the effectiveness of the proposed method the chapter introduces background knowledge discovery by using information retrieval techniques. The uniqueness of this method is that instead of actually incorporating the newly retrieved background knowledge into the learning algorithm, it is used for the purpose of finding previously unknown class related terms. By iteratively applying this process, a large number of classification terms was developed and successfully applied to a task of age classification of a corpus of queries from a commercial search engine.

Since query classifications are done on earlier recorded logs it is interesting to see how calendar dates (e.g. "back to school" season, New Year, etc.) and current events (e.g. Elections) impinge on the effectiveness of the process. Another issue that possibly affects the classification process is the use of current Web collections (significantly larger collections, language of new Websites reflecting social and cultural shifts) to classify older logs. Another promising direction is the creation and expansion of a definitive set of age related terms and phrases.

Topic Analysis

Topic analysis and identification of queries is an important task related to the discipline of information retrieval that is a key element for the development of successful personalized search engines. The problem is more difficult for search engine user queries due to real-time requirements and the limited number of terms in the user queries. Topic identification of search engine queries relates to many studies, ranging from term analysis of search engine queries, to topic estimation, automatic new topic identification, session identification and query clustering, and then to the broader concept of text categorization and natural language processing.

Chapter XVII "Topic Analysis and Identification of Queries" includes a) a detailed literature review on topic analysis and identification, with an emphasis on search engine user queries, b) a survey of the analytical methods that have been and can be used, and c) outlines the challenges and research opportunities related to topic analysis and identification. A comprehensive review covers domain specific search queries (medical, e-commerce, sexual), generic session/topic identification and query clustering approaches and concludes with text classification and categorization models.

The second part of the chapter is devoted to the overview of methodologies used for topic identification including statistical learning methods (regression and Support Vector Machines), artificial intelligence methods (neural networks), statistical and stochastic methods (Markov chains, Dempster-Shafer theory) and methods based on conditional probabilities. The chapter is concluded with suggestions for future research, which include reducing the dimension of text categorization to search engine queries and improving the computational complexity of the topic identification algorithms.

Domain Specific Log Analysis

Clinicians, researchers and members of the general public are increasingly using information technology to cope with the explosion in biomedical knowledge. The vast amount of this knowledge in many areas of biomedicine, and science in general, far exceeds the cognitive capacity of any human. Fortunately for the search engine user, the biomedical domain is associated with many (relatively) well-developed controlled terminologies and vocabularies (ontologies) that make the search process easier and more structured.

Chapter XVIII "Query Log Analysis in Biomedicine" discusses these features of the biomedical domain. 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. While biomedical query log analysis is similar to other domains in its limitations it also exhibits a major advantage – query logs can be complemented by other analyses such as field studies and instrumented user panels. Additionally, mapping the layman's query to controlled taxonomies allows for semantic analysis to understand the meaning of queries.

While assessing the success of the human genome project, the authors predict the explosion of biomedical information and foresee an information challenge facing health care providers and health care consumers. Both groups have sometimes conflicting views, "paternalistic" vs.

participative, and may need different interfaces while searching for identical/similar information. The chapter is concluded with ideas for new tools and user interfaces for biomedical information retrieval.

Language Specific Log Analysis

More and more non-English content is now available on the World Wide Web and the number of non-English users on the Web is increasing. 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.

Chapter XIX "Processing and Analysis of Search Query Logs in Chinese" discusses various methods and techniques that can be used to analyze search queries in Chinese. Stressing the one most notable feature of the Chinese language (an ideographical, character-based language) vs. the English language (an alphabetical, word-based language), the authors explain the difficulty of using traditional log analysis methods for Chinese query logs. For character-based languages, most of the meaningful words are built up by combining single characters, and an individual character may deliver different meanings in different words.

Moreover, in Chinese the syntax of words is quite different from that in English. There is no space between terms in Chinese, making it difficult to correctly perform segmentation, whereas in English every word is basically delimited by space. This specific characteristic of Chinese would result in many apparently different searching behaviors. The discussion is complemented by an example of log analysis based on the Timway search engine which indexes and searches the collection in both languages: Chinese and English. The chapter is concluded with an observation that not all Asian languages follow the analysis pattern and therefore need a different infrastructure for data collection and log analysis.

Goal Specific Query Analysis

Information retrieval and question answering systems often operate in much wider domains for which appropriate corpora are not available. As a result, query logs are an extremely valuable resource for increasing our understanding of the complex interactions involved and hence in developing more sophisticated systems. Logs contain a huge amount of information but effective methods for extracting it are only now being developed.

Chapter XX "Query Log Analysis for Adaptive Dialogue-Driven Search" analyses two case studies, both aimed at improving Information Retrieval and Question Answering systems. The first describes an intranet search engine (UKSearch) 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 (HITIQA - High Quality Interactive Question Answering) 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 and therefore can be used to improve the underlying domain model and the model of interaction, and hence the quality of interaction in a system. Authors conclude with critique of the large body of log analysis literature (extensively reviewed in the chapter) that usually concentrates on determining general trends of usage instead of trying to improve the system under study. Several improvements of domain-dependent spelling correction are suggested as well as a need for generic models of analytic interaction over data.

ETHNOGRAPHY

Ethnography is a qualitative study in which the researcher observes members of a chosen group in a natural environment over a long period of time. It relies on the researcher gaining and maintaining entry into an active group, a challenging (Gorman & Clayton, 2005) and time consuming (Labaree, 2002) process. Whether as an active face-to-face member or as an unobtrusive observer, the researcher plays a variety of roles in order to monitor and register group members' actions and behavior. Nethnography, a portmanteau of net and ethnography, is a recent phenomenon born with the advent of computer-mediated communication (CMC). Analyzing the content of asynchronous or synchronous CMC allows for the discovery and understanding of conventions and types of human interaction, finding meaning and comprehension of the context, discerning topics and distinguishing multiple discussion threads (Hewitt, 2003). The following three chapters demonstrate several methods and frameworks for analysis of three distinct types of Web logs.

Nethnography

While its predecessor – ethnography – relies on active, face-to-face participation of the observer in the study process, nethnography relies on Web logs as the sole source of data.

Chapter XXIV "Nethnography: a Naturalistic Approach towards Online Interaction" explores the potential 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. To illustrate the potential of nethnography to enhance the study of CMC, the author presents an example of an ethnographic study. The chapter is concluded with suggestions on how the nethnography methodology can be adapted to analyze other social networking sites, once care is taken to account for local differences in contexts, topics of interest or activities carried on by participants.

The Blogs

A blog (short for Web log) is part of the network of social media designed and popularized by participants to exchange information, expressed opinions or discuss just about any topic under the sun. According to blog search engine Technorati there are over 100 million blogs.

Chapter XXIII "Information Extraction from Blogs" introduces information extraction from textual blogs. The author argues that the classic techniques for information extraction that are commonly used for mining well-formed texts lose some of their validity in the context of blogs. With the addition of Web 2.0 applications (e.g. tagging) the blog "language" became less structured, more ambiguous and difficult to understand. These findings are demonstrated by considering each step in the information extraction process and by illustrating these problems in different blog applications such as topic and thread detection, opinion mining, and argumentation mining. In order to tackle the problem of mining content from blogs, the author suggests ideas for future research including combining different sources of evidence found in blog texts, their tags, comments and links. The author also suggests some novel applications such as a translator for blogs using community languages and an anti-spammer with the ability to detect and ignore irrelevant content added to mislead filtering and monitoring software.

Finding Meaning in Online Discussions

Proliferation of networking technologies, applications and Web based services allowed for the increase in the number of virtual communities where members join to share common ideas, interests or just desire to express themselves. Much research has been focused on examining the formation of and interactions within these virtual communities. However, the methods for collecting and analyzing data in very large-scale online discussion forums can be varied and complex.

Chapter XV "Methods to Find Meaning in Online Discussion" provides an understanding of how participants come together to form large scale virtual communities and how knowledge flows between participants over time. In this chapter, two analytical methods are described: qualitative data analysis and Social Network Analysis (SNA). Both are used to examine conversations within ESPN's *Fast Break* community, which focuses on fantasy basketball sports games. The first method of analysis, qualitative data analysis, examines threads and collections of messages related by topic and offers insights into the major conversational themes. Individual messages related to these themes are categorized and analyzed to discover the major discussion topics. This method also reflects the individual's game strategy and decision-making.

On the other hand, Social Network Analysis is not focused on the subject matter of the discussion; rather it is concerned with recurring communications between occasional and frequent participants, identifying the primary contributors in the social network and explaining the spread of knowledge in the discussion forum. The chapter is concluded with interesting directions for future research including the need for algorithms and technology to deal with the large number of messages, design of collaborative tools for data collection and analysis, and use of ethnography and SNA to answer specific questions about the online community of interest.

HISTORICAL METHOD

The historical method (historiography) collects and examines facts about events, people and the environment of the past. It attempts to narrate, understand, and interpret the historical data (Godfrey, 2006). It analyzes historical facts and recreates participants' behavior and environment in time and space (Barab & Squire, 2004).

Historic Perspective

What started as transaction log analysis evolved in name and in practice into analysis of Web logs in general and informational retrieval logs in particular.

Chapter II "Historic Perspective of Log Analysis" provides an historical review of the birth and progress of transaction log analysis applied to information retrieval systems. It offers a detailed discussion of the early work (mid-1960's to the late 1970's) evaluating systems performance; explains how this work has migrated (late 1970's through the mid-1980's) to Online Public Access Catalogues evaluation with emphasis on both system use and user behavior; followed by a decade (mid-1980's through the mid-1990's) of data base evaluation; and finally into the evaluation of World Wide Web usage (mid-1990's and on) in countless research directions limited only by imaginations and technology constraints. A discussion of privacy issues with a framework for addressing the same is presented. The chapter is concluded with ideas for research directions (including merging transaction logs with demographic data), new domains (marketing and e-commerce), and suggestions for hybrid research designs that combine the highly quantitative approach of stochastic modeling with the more qualitative approaches available.

DISCOURSE ANALYSIS

Discourse analysis (DA) challenges the conventional thinking along interdisciplinary lines (Weiss & Weiss, 2003) and it shapes the construction and approval of information systems (Vasconcelos, 2007). DA is also used as a scientific argument evaluation method (Acuff, 2007) sometimes understating the underlying scientific or technological concepts. Introducing ontology and epistemology (Hirschheim et al., 1995) into discourse analysis enhances the thoroughness and authenticity of this research method.

Web-Traffic Measurement

Web-traffic measurement is the analysis of data between client and server computers. It provides insight into how people use computers and is commonly used in research.

Chapter V "Watching the Web: An Ontological and Epistemological Critique of Web-Traffic Measurement" provides a brief history of the topic, presents and compares two dominant forms of Web-traffic measurement - log file analysis and ASP-based tools, and critically evaluates the implicit and largely unexamined ontological and epistemological claims of both methods. This evaluation suggests that like all research methods, Web-traffic measurement has implicit ontological and epistemological assumptions embedded within it, albeit to a limited degree. To remedy "ontological and epistemological difficulties" the author suggests several improvements: the Web-traffic measurement must include a systematic method of reflexivity and Web traffic researchers ought explicitly to adopt a more interpretivist stance, based on qualitative approaches. On the applied side, the author also suggests periodically scheduled "reliability tests" whereby Web-traffic researchers can review shifts in traffic patterns.

CASE STUDY

A case study is a comprehensive study of a single subject. Selecting the appropriate unit of analysis to investigate a single subject or a single hypothesis significantly impacts the compilation and explanation of experimental data (Henning et al., 2004). Frequently, the selection and identification of unit of analysis is a complex process that requires additional and sometimes intensive studies (Dubé & Paré, 2003). These additional studies often shift the choice of unit of analysis from larger units to much smaller units, affecting the validity of final results. Lack of consensus among researchers conducting investigation in the same field, as to what the unit of analysis should be, leads to disparate results in longitudinal studies (Yin, 2000).

Unit of Analysis and Validity of Web Log Data

It is a common belief, and a reasonable assumption, that the Web log traces left by the individual Website visitor and collected by the server provide a tremendous amount of quality data. Two issues which concern researchers are what data to measure and how accurate (valid) is the data.

Chapter IX "The Unit of Analysis and the Validity of Web Log Data" examines these issues and explains limitations of the data collection and interpretation processes, as well as sources of such data. First the authors define the measurement units to trace (interaction time, frequency of logins, active/passive involvement, page requests), then they discuss two types of log files (client and server) and explain methodological challenges such as caching, user recognition, and session's length calculation that result in questioning the validity of collected data. The authors suggest guidelines for selecting units of analysis and insuring the validity of log data such as: examine the content structure, consider site specifications, realize and compensate for time inaccuracy at the server level.

DIVERSE RESEARCH METHODOLOGIES, COMMON ISSUES

There are two key issues that influence the design and administration of a research study: privacy of the user and reporting study results. There is growing awareness and concern in the scientific community that careful consideration should be given to the protection of privacy and confidentiality of on-line users (Akram, 2006). Unfamiliar with, or uncertain about Web site privacy and security policy, users/visitors tend to reveal personal demographic and environmental data (Gates & Whalen, 2006; Ward et al., 2005). Storing, sharing and protecting users' information is a frequent topic in scientific research (Patil et al., 2006; Karat et al., 2005; Ngai & Wat, 2002). While self-regulation is an accepted policy for privacy protection in the United States, the European Community favors a legislative approach and tight government supervision (O'Connor, 2006). Another point of agreement among the majority of researchers is the need for improved structure and composition of scientific reporting. Researchers in the medical field responded to this need with several "checklist" standards such as CONSORT (Moher et al., 2001) and CHERRIES (Gunther, 2004). On the other hand, general science researchers use less comprehensive "checklists" such as IMRAD (Sollaci & Pereira, 2004) for reporting research results in addition to a set of writing guidelines popular among researchers (Holliday, 2001). Further needs were also identified for reporting longitudinal research (Tooth et al., 2005).

Privacy and Web Logging

Privacy is a significant concern when planning research that examines human behavior. There are two aspects of privacy that play an important role in conducting such research: strict compliance with existing regulations and alleviation of users' uneasiness with being observed during the experiments.

Chapter IV "Privacy Issues Associated with Web Logging Data" examines these two aspects of privacy. The chapter briefly examines the first aspect as it applies to the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA) and organizational regulations, such as a university's local research ethics board (REB), and then devotes the major part of the chapter to the second aspect – privacy enhancing mechanisms and assurances to encourage natural Web browsing behavior. The author offers an expansive literature overview of general privacy theory while addressing privacy concerns and challenges associated with Web browsing data. The chapter is concluded with numerous recommendations for increasing understanding and trust during observational data collection and suggestions for future analysis of privacy issues impacting collaborative and context browsing.

Recommendations for Reporting Web Usage Studies

Since the advent of Internet, the Web, and search engines, studies of users' use of systems are a hot topic of research and generate a wide variety of studies and reports.

Chapter X "Recommendations for Reporting Web Usage Studies**"** 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.

CONCLUSION

Web logs are increasingly being used, by academic and industry researchers, to study, understand and improve the interaction between the user and Web services. The Handbook of Web Log Analysis focuses on complex issues and answers many hard questions. The handbook tackles issues of privacy, social interaction and community building. It focuses on analysis of the user's behavior during Web activities, and also investigates current methodologies and metrics for Web log analysis. This chapter reviewed various quantitative and qualitative research methodologies used by contributing authors. It summarized results reported in individual chapters and presented new research directions and novel applications of existing knowledge.

REFERENCES

Acuff, J. M. (2007). *Ontological and Epistemological Pluralism in the Evaluation of Scientific Arguments*. Presented at 103rd Annual Meeting of the American Political Science Association, Chicago, IL, August 30-September 2. http://www.allacademic.com/meta/p211986_index.html

Akram, A. A., (2006). Electronic Privacy: Patient Concerns, *Communications of the IIMA,* Vol. 6 (1), 67-82.

Arnold, M. (2003). On the phenomenology of technology: the "Janus-faces" of mobile phones, *Information and Organization*, Vol. 13 (4), 231–256.

Barab, S., & Squire, K. (2004). Design-Based Research: Putting a Stake in the Ground, *The Journal of the learning sciences*, Vol. 13 (1), 1-14

Budd, J. M. (2005). Phenomenology and information studies, *Journal of Documentation,* Vol. 61 (1), 44-59.

Dick, P. (2004), Discourse analysis, in Cassell, C., Symon, G. (Eds), *Essential Guide to Qualitative Methods in Organizational Research*, Sage, London, 203-213.

Dubé, L., & Paré, G. (2003). Rigor in Information Systems Positivist Case Research: Current Practices, Trends, and Recommendations, *MIS Quarterly,* Vol. 27 (4), 597-636.

Fernaeus, Y, Tholander, J., & Jonsson, M. (2008). Towards a New set of Ideals: Consequences of the Practice Turn in Tangible Interaction, *Proceedings of the Second International Conference on Tangible and Embedded Interaction (TEI'08),* 223-230.

Gates, C. & Whalen, T. (2006) Personal Information on the Web: Methodological Challenges and Approaches. *CHI 2006 Workshop on Privacy and HCI: Methodologies for Studying Privacy Issues.*

Godfrey, D. G. (2006). *Methods of Historical Analysis in Electronic Media*. Mahwah, NJ: Erlbaum, 2006

Gorman, G. E., & Clayton, P. (2005). *Qualitative research for the information professional.* London: Facet.

Guther, E. (2004). Improving the quality of Web surveys: The Checklist for Reporting Results of Internet E-Surveys (CHERRIES). *Journal of Medical Internet Research*, Vol. 6 (3) e34. Retrieved March 23, 2008, from http://www.jmir.org/2004/3/e34/.

Henning, E. ,Van Rensburg, W., & Smit, B. (2004). *Finding your way in qualitative inquiry*. Pretoria: Van Schaik.

Hewitt, J. (2003). How Habitual Online Practices Affect the Development of Asynchronous Discussion Threads, *Journal of Educational Computing Research*, Vol. 28 (1), 31-45.

Hirschheim, R. A., H.-K. Klein, H. K. & Lyytinen K. (1995) *Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations*. Cambridge; New York: Cambridge University Press.

Holliday, A. (2001). *Doing and Writing Qualitative Research,* London: Sage.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it, *Library & Information Science Research*, Vol. 28 (3), 407–432.

Jansen, B. J. & Spink, A. (2006).How are we searching the world wide Web?: a comparison of nine search engine transaction logs, *Information Processing and Management,* Vol. 42 (1), 248-263.

Jansen, B. J., Spink, A. & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management*, Vol. 36 (2), 207-227.

Jing, T. Y. (2006). Supporting Research with Weblogs: A Study on Web-Based Research Support Systems, *Proceedings of Web Intelligence and Intelligent Agent Technology Workshops, IEEE/WIC/ACM International Conference*, 161-164.

Karat, C-M., Karat, J., & Brodie, C. (2005). Why HCI research in privacy and security is critical now, *International Journal of Man-Machine Studies*, Vol. 63 (1-2), 1-4

Kim, M. (1996). Research Record, *Journal of Education for Library and Information Science*, Vol. 37, 378-380

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology* (2nd), Thousand Oaks, CA: Sage.

Labaree, R. V. (2002). The risk of "going observationalist": Negotiating the hidden dilemmas of being an insider participant observer. *Qualitative Research,* Vol. 2 (1), 97-122.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, Vol. 28 (4), 587-604.

Meersman, R.; Aberer, K.; & Dillon, T. (2003). Semantic Issues in e-Commerce Systems. Series: *IFIP International Federation for Information Processing*, Vol. 111.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, 357, 1191-1194.

Murphy, E., & Ciszewska-Carr, J. (2005). Contrasting syntactic and semantic units in the analysis of online discussions. *Australasian Journal of Educational Technology,* Vol. 21 (4), 546-566.

Naidu, S. & Järvelä, S. (2006). Analyzing CMC content for what? *Computers & Education,* Vol. 46 (1), 96-103.

Ngai , E. W. T. & Wat, F. K. T. (2002). A literature review and classification of electronic commerce research, *Information & Management*, Vol. 39 (5), 415–429.

Nicholas, D., Huntington, P., & Jamali, H. R. (2007). Diversity in the Information Seeking Behaviour of the Virtual Scholar: Institutional

Comparisons, *Journal of Academic Librarianship*, Vol. 33 (6), 629-638.

Nicholson, S. (2004). A conceptual framework for the holistic measurement and cumulative evaluation of library services, *Journal of Documentation,* Vol. 60 (2), 164-182.

O'Connor, P. (2006). An International Comparison of Approaches to Online Privacy Protection: Implications for the Hotel Sector, *Journal of Services Research*, Vol. 6, 7-26.

Palvia, P., Pinjani, P., & Sibley, E. H. (2007). A profile of information systems research. *Information & Management, Information & Management*, Vol. 44 (1), 1-11.

Patil, S., Romero, N. A. & Karat, J. (2006). Privacy and HCI: methodologies for studying privacy issues, *CHI Extended Abstracts*, 1719-1722.

Powell R.R., (1999). Recent Trends in Research: A Methodological Essay. *Library and Information Science Research,* Vol. 21 (1) 91-119.

Romano, N. C., Donovan, C., Chen, H. C., & Nunamaker, J. F. (2003). A methodology for analyzing Web-based qualitative data. *Journal of Management Information Systems*, Vol. 19 (4), 213-246.

Rossler, P. (2002). Content analysis in online communication: A challenge for traditional methodology. In Batinic, B., Reips, U. D., & Bosnjak, M. (Eds.), *Online social sciences,* Seattle, WA: Hogrefe & Huber.

Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development, 52* (1), 5-18.

Sollaci, L. B. & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey, *Journal of the Medical Library Association*, Vol. 92 (3), 364-371

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17), Retrieved March 15, 2008 from http://pareonline.net/getvn.asp?v=7&n=17

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

Strijbos, J-W., Martens, R. L. Prins, F.J. & Jochems, W. M. G. (2006), Content analysis: What are they talking about? *Computers & Education*, Vol. 46 (1), 29-48.

Tooth, L., Ware, R., Bain, C., Purdie, D. M. & Dobson, A. (2005). Quality of reporting of observational longitudinal research. *American Journal of Epidemiology*, Vol. 161 *(3), 280-288.*

Vasconcelos, A. C. (2007). The role of professional discourses in the organizational adaptation of information systems, *International Journal of Information Management*, Vol. 27 (4), 279-293.

Verbeek, P. P. & Slob, A. (2006), *User Behavior and Technology Development – Shaping Sustainable Relations between Consumers and Technologies*, Dordrecht: Springer.

Vrana, R. (2002). *Digital Libraries - Creating Information Space Excellence: Is It Already Time for Benchmarking?* Paper presented at the 2002 CARNet Users Conference.

Ward, S., Bridges, K., & Chitty, B. (2005). Do incentives matter? An examination of On-line privacy concerns and willingness to provide personal and financial information, *Journal of Marketing Communications*, Vol. 11 (1), 21-40.

Weiss, G. & Weiss. R. (2003). Critical Discourse Analysis. Theory and Interdisciplinarity, New York: Palgrave.

Westbrook, J.I., Braithwaite, J., Iedema, R.A., & Coiera, E.W. (2004). Evaluating the impact of information communication technologies on complex organizational systems: a multi-disciplinary, multi-method framework, *Proceedings of* *the 11th World Congress on Medical Informatics*, Fieschi M, Coiera E & Yu-Chan J, (Eds.), IOS, Washington, USA, 1323 -1327.

Yin, R. K. (2000). Rival explanations as an alternative to reforms as "experiments". In Bickman, L., ed. *Validity and Social Experimentation. Donald Campbell's Legacy*, vol. 1. Thousand Oaks, CA/London/New Delhi: Sage, 239-266.

KEY TERMS

Conceptual Framework/Inquiry: Methodology to build and use conceptual framework as a plan and direction for research.

Content Analysis: Methodical and replicable methodology to determine, quantify and analyze presence of research objects within large data sets.

Discourse Analysis: Scientific argument evaluation method.

Ethnography: A qualitative study in which the researcher observes members of a chosen group in a natural environment over long period of time.

Historical Method: Collects and examines, and interprets facts about events, people and environment of the past.

Phenomenology: An interpretive methodology that examines users' behavior.

Research Methods: Specific approaches employed in research that are typically derived from the research questions or aims.