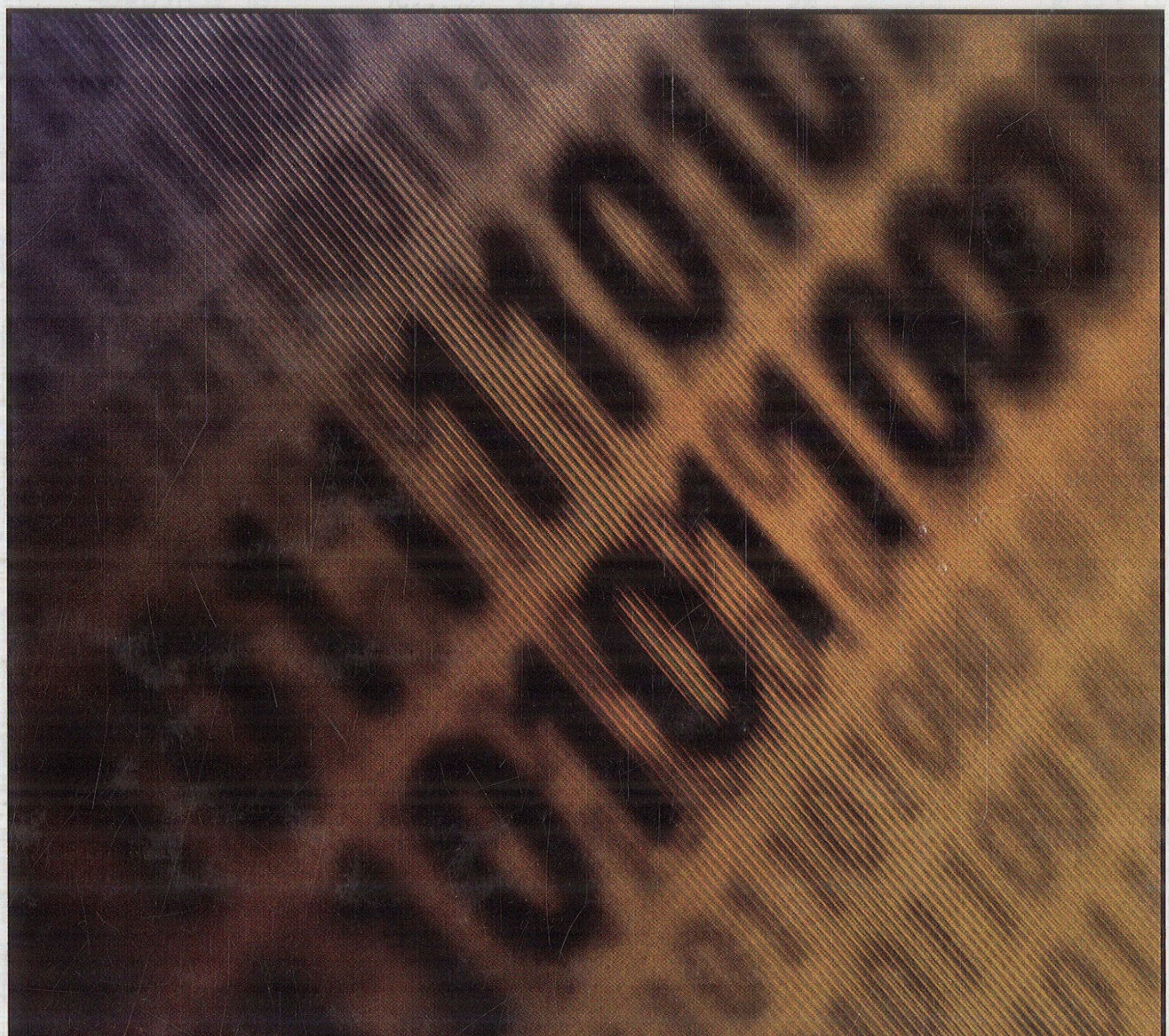
HANDBOOK OF RESEARCH ON

WEB LOG ANALYSIS



Bernard J. Jansen, Amanda Spink, & Isak Taksa

Detailed Table of Contents

Preface xix

Chapter I
Research and Methodological Foundations of Transaction Log Analysis 1
Bernard J. Jansen, Pennsylvania State University, USA
Isak Taksa, Baruch College, City University of New York. USA
Amanda Spink, Queensland University of Technology, Australia

This chapter outlines and discusses theoretical and methodological foundations for transaction log analysis. It first 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. From this research foundation, it then moves to the methodological aspects of transaction log analysis and examine the strengths and limitation of transaction logs as trace data. The chapter then 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

Chapter II
Historic Perspective of Log Analysis 18
W. David Penniman, Nylink, USA

This historical review of the birth and evolution of transaction log analysis applied to information retrieval systems provides two perspectives. First, a detailed discussion of the early work in this area, and second, how this work has migrated into the evaluation of World Wide Web usage. The chapter 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. The chapter concludes with the suggestion that a combination of transaction log analysis of the type used early in its

application along with additional more qualitative approaches will be essential for a deep understanding of user behavior (and needs) with respect to current and future retrieval systems and their design.

Chapter III
Surveys as a Complementary Method for Web Log Analysis 39
Lee Rainie, Pew Internet & American Life Project, USA
Bernard J. Jansen, Pennsylvania State University, USA

Every research methodology for data collection has both strengths and limitations, and this is certainly true for transaction log analysis. Therefore, researchers often need to use other data collection methods with transaction logs. This chapter discusses surveys as a viable alternate method for transaction log analysis. The chapter presents a brief review of survey research literature, with a focus on the use of surveys for Web-related research. We identify the steps in implementing survey research and designing a survey instrument. The chapter concludes 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 65
Sam Ladner, McMaster University, Canada

This chapter aims to improve the rigor and legitimacy of Web-traffic measurement as a social research method. The chapter compares two dominant forms of Web-traffic measurement and discusses the implicit and largely unexamined ontological and epistemological claims of both methods. 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 frame within which phenomena can be rendered intelligible. The chapter argues that Web-traffic measurement employs an ostensibly quantitative, positivistic ontology and epistemology in hopes of cementing the "scientific" legitimacy they engender. These claims to "scientific" method are unsubstantiated, thereby limiting the efficacy and adoption rates of log-file analysis in general. The chapter offers recommendations for improving these measurement tools, including more reflexivity and an explicit rejection of truth claims based on positivistic science.

Chapter V
Privacy Concerns for Web Logging Data 80
Kirstie Hawkey, University of British Columbia, Canada

This chapter examines two aspects of privacy concerns that must be considered when conducting studies that include the collection of Web logging data. After providing background about privacy concerns, the chapter first addresses the standard privacy issues when dealing with participant data. These include privacy implications of releasing data, methods of safeguarding data, and issues encountered with re-use of data. Second, the impact of data collection techniques on a researcher's 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. The chapter aim is that understanding the privacy issues associated with the logging of user actions on the Web will assist researchers as they

evaluate the tradeoffs inherent between the type of logging conducted, the richness of the data gathered, and the naturalness of captured user behavior.

Section II
Methodology and Metrics

Chapter VI
The Methodology of Search Log Analysis 100
Bernard J. Jansen, Pennsylvania State University, USA

Exploiting the data stored in search logs of Web search engines, Intranets, and Websites can provide important insights into understanding the information searching tactics of online searchers. This understanding can inform information system design, interface development, and information architecture construction for content collections. This chapter 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 124
Anthony Ferrini, Acquiremarketing.com, USA
Jakki J. Mohr, University of Montana, USA

As the Web's popularity continues to grow and as new uses of the Web are developed, the importance of measuring the performance of a given Website as accurately as possible also increases. This chapter discusses the various uses of Web analytics (how Web log files are used to measure a Website's performance), as well as the limitations of these analytics. We discuss 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. After reading this chapter, readers should have a nuanced understanding of the "how-to's" of Web analytics.

Chapter VIII
A Review of Methodologies for Analyzing Websites 143
Danielle Booth, Pennsylvania State University, USA
Bernard J. Jansen, Pennsylvania State University, USA

This chapter is 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 through the use of 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 165
Gi Woong Yun, Bowling Green State University, USA

This chapter discusses validity of units of analysis of Web log data. First, Web log units are compared to the unit of analysis of television to understand the conceptual issues of media use unit of analysis. Second, the validity of both Client-side and Server-side Web log data are examined along with benefits and shortcomings of each Web log data. Each method has implications on cost, privacy, cache memory, session, attention, and many other areas of concerns. The challenges were not only theoretical but, also, methodological. In the end, Server-side Web log data turns out to have more potentials than it is originally speculated. Nonetheless, researchers should decide the best research method for their research and they should carefully design research to claim the validity of their data. This chapter provides some valuable recommendations for both Client-side and Server-side Web log researchers.

Chapter X
Recommendations for Reporting Web Usage Studies 181
Kirstie Hawkey, University of British Columbia, Canada
Melanie Kellar, Google, USA

This chapter 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

Chapter XI
From Analysis to Estimation of User Behavior 206
Seda Ozmutlu, Uludag University, Turkey
Huseyin C. Ozmutlu, Uludag University, Turkey
Amanda Spink, Queensland University of Technology, Australia

This chapter 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 227
Gheorghe Muresan, Microsoft Corporation, USA

This chapter describes and discusses a methodological framework that integrates analysis of interaction logs with the conceptual design of the user interaction. It is based on (i) formalizing the functionality that is supported by an interactive system and the valid interactions that can take place; (ii) deriving schemas for capturing the interactions in activity logs; (iii) deriving log parsers that reveal the system states and the state transitions that took place during the interaction; and (iv) 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. We present the details of the methodology, and exemplify its use in a mediated retrieval experiment, in which the focus of the study is on studying the information-seeking process and on finding interaction patterns.

Chapter XIII
Tips for Tracking Web Information Seeking Behavior 256
Brian Detlor, McMaster University, Canada
Maureen Hupfer, McMaster University, Canada
Umar Ruhi, University of Ottawa, Canada

This chapter 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 of 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: i) the need to account for both task and individual differences in any Web information seeking behavior analysis; ii) how to collect Web metrics through deployment of a unique ID that links individual differences, task, and Web tracking data together; iii) the types of Web log metrics to collect; iv) how to go about collecting and making sense of such metrics; and v) 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 284
Sandro José Rigo, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil
José Palazzo M. de Oliveira, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil
Leandro Krug Wives, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

Adaptive Hypermedia is an effective approach to automatic personalization that overcomes the difficulties and deficiencies of traditional Web systems in delivering the appropriate content to users. One important issue regarding Adaptive Hypermedia systems is the construction and maintenance of the user profile. Another important concern is the use of Semantic Web resources to describe Web applications and to implement adaptation mechanisms. Web Usage Mining, in this context, allows the generation of Websites access patterns. This chapter describes the possibilities of integration of these usage patterns with semantic knowledge obtained from domain ontologies. 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 307
Brian K. Smith, Pennsylvania State University, USA
Priya Sharma, Pennsylvania State University, USA
Kyu Yon Lim, Pennsylvania State University, USA
Goknur Kaplan Akilli, Pennsylvania State University, USA
KyoungNa Kim, Pennsylvania State University, USA
Toru Fujimoto, Pennsylvania State University, USA
Paula Hooper, TERC, USA

Computers and networking technologies have led to increases in the development and sustenance of online communities, and much research has focused on examining the formation of and interactions within these virtual communities. The methods for collecting data and analyzing virtual online communities, especially very large-scale online discussion forums can be varied and complex. This chapter describes two analytical methods—qualitative data analysis and Social Network Analysis (SNA)–that we used to examine conversations within ESPN's Fast Break community, which focuses on fantasy basketball sports games. Two different levels of analyses—the individual and community level—allowed us to examine individual reflection on game strategy and decision-making as well as characteristics of the community and patterns of interactions between participants within community. The description of our use of these two analytical methods can help researchers and designers who may be attempting to analyze and characterize other large-scale virtual communities.

Section IV
Query Log Analysis

Chapter XVI
Machine Learning Approach to Search Query Classification 329
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

Search query classification is a necessary step for a number of information retrieval tasks. This chapter 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 345
Seda Ozmutlu, Uludag University, Turkey
Huseyin C. Ozmutlu, Uludag University, Turkey
Amanda Spink, Queensland University of Technology, Australia

This chapter 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 359
Elmer V. Bernstam, University of Texas Health Science Center at Houston, USA
Jorge R. Herskovic, University of Texas Health Science Center at Houston, USA
William R. Hersh, Oregon Health & Science University, USA

Clinicians, researchers and members of the general public are increasingly using information technology to cope with the explosion in biomedical knowledge. This chapter 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. This 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 378
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

More non-English contents are 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 we discusses some methods and techniques that can be used to analyze search queries in Chinese. We also show an example of applying our methods on a Chinese Web search engine. Some interesting findings are reported.

Chapter XX
Query Log Analysis for Adaptive Dialogue-Driven Search 389
Udo Kruschwitz, University of Essex, UK
Nick Webb, SUNY Albany, USA
Richard Sutcliffe, University of Limerick, Ireland

The theme of this chapter is the improvement of Information Retrieval and Question Answering systems by the analysis of query logs. Two case studies are discussed. The first describes an intranet search engine working on a university campus which can present sophisticated query modifications to the user. It does this via a hierarchical domain model built using multi-word term co-occurrence data. The usage log was analysed 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 results can be used to validate refinements in the domain model, and to suggest replacements such as domain-dependent spelling corrections. The second case study describes a dialogue-based question answering system working over a closed document collection largely derived from the Web. Logs here 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. Future versions of the system will encourage these forms of interaction. The conclusions of this chapter are firstly that there is a growing literature on query log analysis, much of it reviewed here, secondly that logs provide many forms of useful information for improving a system, and thirdly that mutual information measures taken with automatic term recognition algorithms and hierarchy construction techniques comprise one approach for enhancing system performance.

Section V
Contextual and Specialized Analysis

Chapter XXI
Using Action-Object Pairs as a Conceptual Framework for Transaction Log Analysis 416
Mimi Zhang, Pennsylvania State University, USA
Bernard J. Jansen, Pennsylvania State University, USA

This chapter presents the action-object pair approach as a conceptual framework for conducting transaction log analysis. We 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 expression of the user. An object is a self-contained information object, the recipient 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 chapter proposes that this approach can benefit system design by providing the organizing principle for implicit feedback and other interactions 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 our theoretical and conceptual understanding of transaction log analysis as a research method.

Chapter XXII
Analysis and Evaluation of the Connector Website 436
Paul DiPerna, The Blau Exchange Project, USA

This chapter proposes a new theoretical construct for evaluating websites that facilitate online social networks. The suggested model considers previous academic work related to social networks and online communities. This chapter'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 or not particular connector websites can be mechanisms for social change, and to quantify the nature of the observed social change. The chapter's aim is to introduce new applications for Web log analysis by evaluating connector websites and their organizations.

Chapter XXIII
Information Extraction from Blogs 469
Marie-Francine Moens, Katholieke Universiteit Leuven, Belgium

This chapter 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 488
Adriana Andrade Braga, Pontifícia Universidade Católica do Rio de Janeiro, Brazil

This chapter explores the possibilities and limitations of nethnography, an ethnographic approach applied to the study of online interactions, particularly computer-mediated communication. In this chapter, a brief history of ethnography, including its relation to anthropological theories and its key methodological assumptions is addressed. Next, one of the most frequent methodologies applied to Internet settings, that is to treat logfiles as the only or main source of data, is explored, and its consequences are analyzed. In addition, some strategies related to a naturalistic perspective for data analysis are examined. Finally, an example of an ethnographic study that involves participants of a Weblog is presented to illustrate the potential for nethnography to enhance the study of computer-mediated communication.

Chapter XXV
Web Log Analysis: Diversity of Research Methodologies 506
Isak Taksa, Baruch College, City University of New York, USA
Amanda Spink, Queensland University of Technology, Australia
Bernard J. Jansen, Pennsylvania State University, USA

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.

Glossary 523

Compilation of References 538

About the Contributors 593

Index 601