Chapter I
Research and Methodological Foundations of Transaction Log Analysis

Bernard J. Jansen
Pennsylvania State University, USA

Isak Taksa
City University of New York. USA

Amanda Spink
Queensland University of Technology, USA

ABSTRACT

This chapter outlines and discusses theoretical and methodological foundations for transaction log analysis. We first address 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, we move to the methodological aspects of transaction log analysis and examine the strengths and limitation of transaction logs as trace data. We then review the conceptualization of transaction log analysis as an unobtrusive approach to research, and present 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.

INTRODUCTION

Conducting research involves the use of both a set of theoretical constructs and methods for investigation. For empirical research, the results are linked conceptually to the data collection process. Quality research papers must contain a thorough methodology section. In order to understand em-

pirical research and the implications of the results, one must thoroughly understand the techniques by which the researcher collected and analyzed data. When conducting research concerning users and information systems, there is a variety of methods at ones disposal. These research methods are qualitative, quantitative, or mixed. The selection of an appropriate method is critically important if the research is to have effective outcomes and be efficient in execution. The method of data collection also involves a choice of methods. Transaction logs and transaction log analysis is one approach to data collection and a research method for both system performance and user behavior analysis that has been used since 1967 (Meister & Sullivan, 1967) and in peer reviewed research since 1975 (Penniman, 1975).

A transaction log is *an electronic record of interactions that have occurred between a system and users of that system*. These log files can come from a variety of computers and systems (Websites, OPAC, user computers, blogs, listserv, online newspapers, etc.), basically any application that can record the user – system – information interactions. Transaction log analysis is the methodological approach to studying online systems and users of these systems. Peters (1993) defines transaction log analysis as *the study of electronically recorded interactions between on-line information retrieval systems and the persons who search for information found in those systems*. Since the advent of the Internet, we have to modify Peter's (1993) definition, expanding it to include systems other than information retrieval systems.

Transaction log analysis is a broad categorization of methods that covers several sub-categorizations, including Web log analysis (i.e., analysis of Web system logs), blog analysis and search log analysis (analysis of search engine logs). Transaction log analysis enables macroanalysis of aggregate user data and patterns and microanalysis of individual search patterns. The results from the analyzed data develop systems and services based on user behavior or system performance.

From the user behavior side, transaction log analysis is one of a class of unobtrusive methods (a.k.a., non-reactive or low-constraint). Unobtrusive methods allow data collection without directly asking participants. The research literature specifically describes unobtrusive approaches as those that do not require a response from participants (c.f., McGrath, 1994; Page, 2000; Webb, Campbell, Schwarz, & Sechrest, 2000). This data can be observational or existing data. Unobtrusive methods are in contrast to obtrusive or reactive approaches such as questionnaires, tests, laboratory studies, and surveys (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). A laboratory experiment is an example of an extreme obtrusive method. Certainly, the line between unobtrusive and obtrusive methods is sometimes blurred. For example, conducting a survey to gauge the reaction of users to information systems is an obtrusive method. However, using the posted results from the survey is an unobtrusive method.

In this chapter, we address the research and methodological foundations of transaction log analysis. We first address the concept of transaction logs as a data collection technique from the perspective of behaviorism. We then review the conceptualization of transaction log analysis as trace data and an unobtrusive method. We present the strengths and shortcomings of the unobtrusive methodology approach, including benefits and shortcomings of transaction log analysis specifically from the perspective of an unobtrusive method. We end with a short summary and open questions of transaction logging as a data collection method.

The use of transaction logs for academic purposes certainly falls conceptually within the confines of the behaviorism paradigm of research. The behaviorism approach is the conceptual basis for the transaction log approach.

BEHAVIORISM

Behaviorism is a research approach that emphasizes the outward behavioral aspects of thought. Strictly speaking, behaviorism also dismisses the inward experiential and procedural aspects (Skinner, 1953; Watson, 1913); behaviorism has come under critical fire for this narrow viewpoint.

However, for transaction log analysis, we take a more open view of behaviorism. In this more encompassing view, behaviorism emphases the observed behaviors without discounting the inner aspects that may accompany these outward behaviors. This more open outlook of behaviorism supports the viewpoint that researchers can gain much from studying expressions (i.e., behaviors) of users where interacting with information systems. These expressed behaviors may reflect both aspects of the person's inner self but also contextual aspects of the environment within which the behavior occurs. These environmental aspects may influence behaviors that are also reflective of inner cognitive factors.

The underlying proposition of behaviorism is that all things that people do are behaviors. These behaviors include actions, thoughts, and feelings. With this underlying proposition, the behaviorism position is that all theories and models concerning people have observational correlates. The behaviors and any proposed theoretical constructs must be mutually complementary. Strict behaviorism would further state that there are no differences between the publicly observable behavioral processes (i.e., actions) and privately observable behavioral processes (i.e., thinking and feeling). We take the position that, due to contextual, situational, or environmental factors, there many times may be such disconnection between the cognitive and affective processed. Therefore, there are sources of behavior both internal (i.e., cognitive, affective, expertise) and external (i.e., environmental and situational). Behaviorism focuses primarily on only what an observer can see or manipulate.

We see the effects of behaviorism in many types of research and especially in transaction log analysis. Behaviorism is evident in any research where the observable evidence is critical to the research questions or methods. This is especially true in any experimental research where the operationalization of variables is required. A behaviorism approach at its core seeks to understand events in terms of behavioral criteria (Sellars, 1963, p. 22). Behaviorist research demands behavioral evidence. Within such a perspective, there is no knowable difference between two states unless there is a demonstrable difference in the behavior associated with each state.

Research grounded in behaviorism always involves *somebody* doing *something* in a *situation.* Therefore, all derived research questions focus on *who* (actors), *what* (behaviors), *when* (temporal), *where* (contexts), and *why* (cognitive). The actors in a behaviorism paradigm are people at whatever level of aggregation (e.g., individuals, groups, organizations, communities, nationalities, societies, etc.) whose behavior is studied. Such research must focus on behaviors, all aspects of what the actors do. These behaviors have a temporal element, when and how long these behaviors occur. The behaviors occur within some context, which are all the environmental and situational features in which these behaviors are embedded. The cognitive aspect to these behaviors is the rational and affective processes internal to the actors executing the behaviors.

From this research perspective, each of these (i.e., actor, behaviors, temporal, context, and cognitive) are behaviorist constructs. However, for transaction log analysis, one is primarily concerned with "what is a behavior?"

Behaviors

A variable in research is an entity representing a set of events where each event may have a different value. In log analysis, session duration or number of clicks may be variables that a research

is interested in. The particular variables that a research is interested in are derived from the research questions driving the study.

One can define variables by their use in a research study (e.g., independent, dependent, extraneous, controlled, constant, and confounding) and by their nature. Defined by their nature, there are three types of variables, which are environments (i.e., events of the situation, environment, or context), subject (i.e., events or aspects of the subject being studied), and behavioral (i.e., observable events of the subject of interest).

For transaction log analysis, behavior is the essential construct of the behaviorism paradigm. At its most basic, a behavior is an observable activity of a person, animal, team, organization, or system. Like many basic constructs, behavior is an overloaded term, as it also refers to the aggregate set of responses to both internal and external stimuli. Therefore, behaviors address a spectrum of actions. Because of the many associations with the term, it is difficult to characterize a term like behavior without specifying a context in which it takes place to provide meaning.

However, one can generally classify behaviors into four general categories, which are:

1. Behavior is something that one can detect and, therefore, record.
2. Behavior is an action or a specific goal-driven event with some purpose other than the specific action that is observable.
3. Behavior is some skill or skill set.
4. Behavior is a reactive response to environmental stimuli.

In some manner, the researcher must observe these behaviors. By observation, we mean studying and gathering information on a behavior concerning what the actor does. Classically, observation is visual, where the researcher uses his/her own eyes. However, observation is assisted with some recording device, such as a camera. We extend the concept of observation to include other recording devices, notably logging software. Transaction log analysis focuses on descriptive observation and logging the behaviors, as they would occur.

When studying behavioral patterns during transaction log analysis and other similar approaches, researchers use ethograms. An ethogram is an index of the behavioral patterns of a unit. An ethogram details the different forms of behavior that an actor displays. In most cases, it is desirable to create an ethogram in which the categories of behavior are objective, discrete, not overlapping with each other. The definitions of each behavior should be clear, detailed and distinguishable from each other. Ethograms can be as specific or general as the study or field warrants.

Spink and Jansen (2004), and Jansen and Pooch (2001) outline some of the key behaviors for search log analysis, a specific form of transaction log analysis. Hargittai (2004) and Jansen and McNeese (2005) present examples of detailed classifications of behaviors during Web searching. As an example, Table 1 presents an ethogram of user behaviors interacting with a Web browser during a searching session, with Table 2 (as an appendix) presenting the complete ethogram.

There are many way to observe behaviors. In transaction log analysis, we are primarily concerned with observing and recording these behaviors in a file. As such, one can view the recorded fields as trace data.

Trace Data

The researcher has several options to collect data for research, but there is no one single best method for collection. The decision about which approach or approaches to use depends upon the research questions (i.e., what needs to be investigated, how one needs to record the data, what resources are available, what is the timeframe available for data collection, how complex is the data, what is the

Table 1. Taxonomy of user-system interactions (Jansen & McNeese, 2005)

State	Description
View results	Interaction in which the user viewed or scrolled one or more pages from the results listing. If a results page was present and the user did not scroll, we counted this as a View Results Page.
With Scrolling	*User scrolled the results page.*
Without Scrolling	*User did not scroll the results page.*
but No Results in Window	*User was looking for results, but there were no results in the listing.*
Selection	Interaction in which the user makes a selection in the results listing.
Click URL (in results listing)	*Interaction in which the user clicked on a URL of one of the results in the results page.*
Next in Set of Results List	*User moved to the Next results page.*
Previous in Set of Results List	*User moved to the Previous results page.*
GoTo in Set of Results List	*User selected a specific results page.*
View document	Interaction in which the user viewed or scrolled a particular document in the results listings.
With Scrolling	*User scrolled the document.*
Without Scrolling	*User did not scroll the document.*
Execute	Interaction in which the user initiated an action in the interface.
Execute Query	*Interaction in which the user entered, modified, or submitted a query without visibly incorporating assistance from the system. This category includes submitting the original query, which was always the first interaction with system.*
Find Feature in Document	*Interaction in which the user used the FIND feature of the browser.*
Create Favorites Folder	*Interaction in which the user created a folder to store relevant URLs.*
Navigation	Interaction in which the user activated a navigation button on the browser, such as Back or Home.
Back	*User clicked the Back button.*
Home	*User clicked the Home button.*
Browser	Interaction in which the user opened, closed, or switched browsers.
Open new browser	*User opened a new browser.*
Switch /Close browser window	*User switched between two open browsers or closed a browser window.*
Relevance action	Interaction such as print, save, bookmark, or copy.
Bookmark	*User bookmarked a relevant document.*

frequency of data collection, and how the data is to be analyzed.).

For transaction log data collection, we are generally concerned with observations of behavior. The general objective of observation is to record the behavior, either in a natural state or in a laboratory study. In both settings, ideally, the researcher should not interfere with the behavior. However, when observing people, the knowledge that they are being observed is likely to alter participants' behavior. In laboratory studies, a researcher's instructions may change a participant's behavior. With logging software, the introduction of the application may change a user's behavior.

With these limitations of observational techniques in mind, when investigating user behaviors, the researcher must make a record of these behaviors to have access to this data for future analysis. The actor, a third party, or the researcher, can make the record of behaviors. Transaction logging is an indirect method of recording data about behaviors, and the actors themselves, with the help of logging software, via traces make these data records of behavior. Thus, transaction log records are a source of trace data.

The processes by which people conduct the activities of their daily lives many times create things, create marks, or reduce some existing material. Within the confines of research, these things, marks, and wear become data. Classically, trace data are the physical remains of interaction (Webb et al., 2000, p. 35 - 52). This creation can be intentional (i.e., notes in a diary) or accidental (i.e., footprints in the mud). However, trace data can also be through third party logging applications. In transaction log analysis, we are primarily interested in this data from third party logging. We refer to this data as trace data.

Researchers use physical or, as in the case of transaction log analysis, virtual traces as indicators of behavior. These behaviors are the facts or data that researchers use to describe or make inferences about events concerning the actors. Researchers (Webb et al., 2000) have classified trace data, into two general types. These two general types of trace measures are erosion and accretion. Erosion is the wearing away of material leaving a trace. Accretion is the build-up of material, making a trace. Both erosion and accretion have several subcategories. In transaction log analysis, we are primarily concerned with accretion trace data.

Trace data or measures offer a sharp contrast to directly collected data. The greatest strength of trace data is that it is unobtrusive. The collection of the data does not interfere with the natural flow of behavior and events in the given context. Since the data is not directly collected, there is no observer present in the situation where the behaviors occur to affect the participants' actions. Trace data is unique; as unobtrusive and nonreactive data it can make a very valuable research direction. In the past, trace data was often time consuming to gather and process, making such data costly. With the advent of transaction logging software, trace data for the studying of behaviors of users and systems has really taken off.

Interestingly, in the physical world, erosion data is what typically reveals usage patterns (i.e., trails worn in the woods, footprints in the snow, wear on a book cover). However, with transaction log analysis, logged accretion data provides us the usage patterns (i.e., access to a Website, submission of queries, Webpages viewed). Specifically, transaction logs are a form of controlled accretion data, where the researcher or some other entity alters the environment in order to create the accretion data (Webb et al., 2000, p. 35 - 52). With a variety of tracking applications, the Web is a natural environment for controlled accretion data collection.

Like all data collection methods, trace data for studying users and systems has strengths and limitations. Trace data are valuable for understanding behavior (i.e., trace actions) in naturalistic environments, offering insights into human activity obtainable in no other way. For example, data from transaction logs is on a scale available in few

other places. However, one must interpret trace data carefully and with a fair amount of caution, as trace data can be misleading. For example, with the data in transaction logs, the research can say a given number of search engine users only looked at the first result page. However, using trace data alone, the researcher could not conclude whether the users left because they found their information or because they were frustrated because they could not find it.

Trace data from transaction logs should be examined during analysis based on the same criteria as all research data. These criteria are credibility, validity, and reliability.

Credibility refers to how trustworthy or believable is the data collection method. The researcher must make the case that the data collection methodology records the data needed to address the underlying research questions.

Validity describes if the measurement actually measures what it is supposed to measure. There are three kinds of validity:

a. Face or internal validity addresses the extent to which the contents of the test or procedure the researcher is measuring looks like what they are supposed to measure.
b. Content or construct validity addresses the extent to which the content of the test or procedure adequately represents all that is required for validity.
c. External validity is the extent to which one can generalize the research results across populations, situations, environments, and contexts.

In inferential or predictive research, one must also be concerned with statistical validity (i.e., the degree of strength of the independent and dependent variable relationships),

Reliability is a term used to describe the stability of the measurement. Does the measurement measure the same thing, in the same way, in repeated tests.

How to address the issues of credibility, validity, reliability? Building on the work of (Holst, 1969), six questions must be addressed in every research project using trace data from transaction logs:

1. **Which data are analyzed?** The researcher must clearly articulate in a precise manner and format what trace data was recorded. With transaction log software, this is much easier than in other forms of trace data, as logging applications can be reverse engineered to clearly articulate exactly what behavioral data is recorded.
2. **How is this data defined?** The researcher must clearly define each trace measure in a manner that permits replication of the research on other systems and with other users. As transaction log analysis has proliferated in a variety of venues, more precise definitions of measures are developing (Park, Bae & Lee, 2005; Wang, Berry, & Yang, 2003; Wolfram, 1999).
3. **What is the population from which the researcher has drawn the data?** The researcher must be cognizant of the actors, both people and systems that created the trace data. With transaction logs on the Web, this is sometimes a difficult issue to address directly, unless the system requires some type of logon and these profiles are then available. In the absence of these profiles, the researcher must rely on demographic surveys, studies of the system's user population, or general Web demographics.
4. **What is the context in which the researcher analyzed the data?** It is important for the researcher to clearly articulate the environmental, situational, and contextual factors under which the trace data was recorded. With transaction log data, this refers to providing complete information about the temporal factors of the data collection (i.e., the time the data was recorded) and the

make up of the system at the time of the data recording, as system features undergo continual change. Transaction logs have the significant advantage of time sampling of trace data. In time sampling, the researcher can make the observations at predefined points of time (e.g., every five minutes), and then record the action that is taking place, using the classification of action defined in the ethogram.

5. **What are the boundaries of the analysis?** Research using trace data from transaction logs is tricky, and the researcher must be careful not to over reach with the research questions and findings. The implications of the research are confined by the data and the method of the data collected. For example, with transaction log data, one can rather clearly state whether or not a user clicked on a link. However, transaction log trace data itself will not inform the researcher why the user clicked on a clink.
6. **What is the target of the inferences?** The researcher must clearly articulate the relationship among the separate measures in the trace data to either inform descriptively or in order to make inferences. Trace data can be used for both descriptive research for understanding and predictive research in terms of making inferences. These descriptions and inferences can be at any level of granularity (i.e., individual, collection of individuals, organization, etc.). However, Hilber and Redmiles (1998) point out that transaction log data is best used for aggregate level analysis, based on their experiences.

Transaction logs are an excellent way to collect trace data on users of Web and other information systems. The researcher then examines this data using transaction log analysis. The use of trace data to understand behaviors makes the use of transaction logs and transaction logs analysis an unobtrusive research method.

UNOBTRUSIVE METHOD

Unobtrusive methods are research practices that do not require the researcher to intrude in the context of the actors. Unobtrusive methods do not involve direct elicitation of data from the research participants or actors. This approach is in contrast to obtrusive methods such as laboratory experiments and surveys requiring that the researchers physically interject themselves into the environment being studied. This intrusion can lead the actors to alter their behavior in order to look good in the eyes of the researcher or for other reasons. For example, a questionnaire is an interruption in the natural stream of behavior. Respondents can get tired of filling out a survey or resentful of the questions asked. Unobtrusive measurement presumably reduces the biases that result from the intrusion of the researcher or measurement instrument. However, unobtrusive measures reduce the degree of control that the researcher has over the type of data collected. For some constructs, there may simply not be any available unobtrusive measures.

Why is it important for the research not to intrude upon the environment? There are at least three justifications. First, is the uncertainty principle (a.k.a., the Heisenberg uncertainty principle). The Heisenberg uncertainty principle is from the field of quantum physics. In quantum physics, the outcome of a measurement of some system is not deterministic or perfect. Instead, a measurement is characterized by a probability distribution. The larger the associated standard deviation is for this distribution, the more "uncertain" are the characteristics measured for the system. The Heisenberg uncertainty principle is commonly stated as "One cannot accurately and simultaneously measure both the position and momentum of a mass." (http://en.wikipedia.org/wiki/Uncertainty_principle). In this analogy, when researchers are interjected into an environment, they become part of the system. Therefore, there just being there will affect measurements.

A common example in the information technology area is the interjection of a recording device into an existing information technology system just for the purposes of measuring may slow the response time of the system.

The second justification is the observer effect. The observer effect refers to the difference that is made to an activity or a person's behaviors by it being observed. People may not behave in their usual manner if they know that they are being watched or when being interviewed while carrying out an activity. In research, this observer effect specifically refers to changes that the act of observing will make on the phenomenon being observed. In information technology, the observer effect is the potential impact of the act of observing a process output while the process is running. A good example of the observer effect in transaction log analysis is pornographic searching behavior. Participants rarely search for porn in a laboratory study while studies employing trace data shows it is a common searching topic (Jansen & Spink, 2005).

The third justification is observer bias. Observer bias is error that the researcher introduces into measurement when observers overemphasize behavior they expect to find and fail to notice behavior they do not expect. Many fields have common procedures to address this, although seldom used in information and computer science. For example, the observer bias is why medical trials are normally double-blind rather than single-blind. Observer bias is introduced because researchers see a behavior and interpret it according to what it means to them, whereas it may mean something else to the person showing the behavior. Trace data helps in overcoming the observer bias in the data collection. However, as with other methods, it has no effect on the observer bias in interpretation of the results from data analysis.

We discuss three types of unobtrusive measurement that are applicable to transaction log analysis research, which are indirect analysis, context analysis, and second analysis. Transaction logs analysis is an indirect analysis method. The researcher is able to collect the data without introducing any formal measurement procedure. In this regard, transaction log analysis typically focuses in the interaction behaviors occurring among the users, system, and information. There are several examples of utilizing transaction analysis as an indirect approach (Abdulla, Liu & Fox, 1998; Beitzel, Jensen, Chowdhury, Grossman & Frieder, 2004; Cothey, 2002; Hölscher & Strube, 2000).

Content analysis is the analysis of text documents. The analysis can be quantitative, qualitative or a mixed methods approach. Typically, the major purpose of content analysis is to identify patterns in text. Content analysis has the advantage of being unobtrusive and depending on whether automated methods exist can be a relatively rapid method for analyzing large amounts of text. In transaction log analysis, content analysis typically focuses on search queries or analysis of retrieved results. There is a variety of examples in this area of transaction log research (Baeza-Yates, Caldeón-Benavides & Gonźalez, 2006; Beitzel, Jensen, Lewis, Chowdhury & Frieder, 2007; Hargittai, 2002; Wang et al., 2003; Wolfram, 1999).

Secondary data analysis, like content analysis, makes use of already existing sources of data. However, secondary analysis typically refers to the re-analysis of quantitative data rather than text. Secondary data analysis is the analysis of preexisting data in a different way or to address different research questions than originally intended during data collection. Secondary data analysis utilizes the data that was collected by someone else. Transaction log data is commonly collected by Websites for system performance analysis. However, researchers can also use this data to address other questions. Several transaction log studies have focused on this aspect of research (Nico Brooks, 2004; N. Brooks, 2004; Choo, Betlor, & Turnbull, 1998; Chowdhury & Soboroff 2002; Croft, Cook, & Wilder, 1995; Joachims, Granka, Pan, Hembrooke, & Gay,

2005; Montgomery & Faloutsos, 2001; Rose & Levinson, 2004).

As a secondary analysis method, transaction log analysis has several advantages. First, it is efficient in that it makes use of data collected by a Website application. Second, it often allows the researcher to extend the scope of the study considerably by providing access to a potentially large sample of users over a significant duration (Kay & Thomas, 1995). Third, since the data is already collected, the cost of existing transaction log data is cheaper than collecting primary data.

However, the use of secondary analysis is not without difficulties. First, secondary data is frequently not trivial to prepare, clean, and analyze, especially large transaction logs. Second, researchers must often make assumptions about how the data was collected as the logging applications were developed by third parties. Third, there is the ethics of using transaction logs as secondary data. By definition, the researcher is using the data in a manner that may violate the privacy of the system users. In fact, some point out a growing distaste for unobtrusive methods due to increased sensitivity toward the ethics involved in such research (Page, 2000).

Transaction Log Analysis as Unobtrusive Method

Transaction logs analysis has significant advantages as a methodology approach for the study and investigation of behaviors. These factors include:

- *Scale*: Transaction log applications can collect data to a degree that overcomes the critical limiting factor in laboratory user studies. User studies in laboratories are typically restricted in terms of sample size, location, scope, and duration.
- *Power*: The sample size of transaction log data can be quite large, so inference testing can highlight statistically significant relationships. Interestingly, sometimes the amount of data in transaction logs from the Web is so large, that nearly every relation is significantly correlated due to the large power.
- *Scope*: Since transaction log data is collected in natural context, the researchers can investigate the entire range of user – system interactions or system functionality in a multi-variable context.
- *Location*: Transaction log data can be collected in a naturalistic, distributed environment. Therefore, the users do not have to be in an artificial laboratory setting.
- *Duration*: Since there is no need for specific participants recruited for a user study, transaction log data can be collected over an extended period.

All methods of data collection have both strengths not available with other methods, but they also have inherent limitations. Transactions logs have several shortcomings. First, transaction log data is not nearly as versatile relative to primary data as the data may not have been collected with the particular research questions in mind. Second, transaction log data is not as rich as some other data collection methods and therefore not available for investigating the range of concepts some researchers may want to study. Third, the fields that the transaction log application records are many times only loosely linked to the concepts they are alleged to measure. Fourth, with transaction logs, the users may be aware that they are being recorded and may alter their actions. Therefore, the user behaviors may not be altogether natural.

Given the inherent limitations in the method of data collection, transaction log analysis also suffers from shortcomings deriving from the characteristics of the data collection. Hilbert and Redmiles (2000) maintain that all research methods suffer from some combination of abstraction, selection, reduction, context, and evolution prob-

lems that limit scalability and quality of results. Transaction log analysis suffers from these same five shortcomings:

- *Abstraction problem*: How does one relate low-level data to higher-level concepts?
- *Selection problem*: How does one separate the necessary from unnecessary data prior to reporting and analysis?
- *Reduction problem*: How does one reduce the complexity and size of the data set prior to reporting and analysis?
- *Context problem*: How does one interpret the significance of events or states within state chains?
- *Evolution problem*: How can one alter data collection applications without impacting application deployment or use?

Because each method has its own combination of abstraction, selection, reduction, context, and evolution problems, this points to the need for complementary methods of data collection and analysis. This is similar to the conflict inherent in any overall research approach. Each research method for data collection tries to maximize three desirable criteria: *generalizability* (i.e., the degree to which the data applies to overall populations), *precision* (i.e., the degree of granularity of the measurement), and *realism* (i.e., the relation between the context in which evidence is gathered relative to the contexts to which the evidence is to be applied). Although the researcher always wants to maximize all three of these criteria simultaneously - it cannot be done. This is one fundamental dilemma of the research process. The very things that increase one of these three features will reduce one or both of the others.

CONCLUSION

Recordings of behaviors via transaction log applications on the Web opens a new era for researchers by making large amounts of trace data available for use. The online behaviors and interactions among users, systems and information create digital traces that permit analysis of this data. Logging applications provide data obtained through unobtrusive methods, massively larger than any data set obtained via surveys or laboratory studies, and collected in naturalistic settings with little to no impact by the observer. Researchers can use these digital traces to analyze a nearly endless array of behavior topics.

The use of transaction log analysis is a behaviorist research method, with a natural reliance on the expressions of interactions as behaviors. The transaction log application records these interactions, creating a type of trace data. Trace data in transaction logs are records of interactions as people use these systems to locate information, navigate Websites, and execute services. The data in transaction logs is a record of user – system, user – information, or system – information interactions. As such, transaction logs provide an unobtrusive manner of collecting these behaviors. Transaction logs provide a method of collecting data on a scale well beyond what one could collect in confined laboratory studies.

The massive increased availability of Web trace data has sparked concern over the ethical aspects of using unobtrusively obtained data from transaction logs. For example, who does the trace data belong to - the user, the Website that logged the data, or the public domain? How does (or should one) seek consent to use such data? If researchers do seek consent, from whom does the researcher seek it? Is it realistic to require informed consent for unobtrusively collected data? These are open questions.

REFERENCES

Abdulla, G., Liu, B., & Fox, E. (1998). Searching the World-Wide Web: implications from studying different user behavior. *Paper presented at*

the World Conference of the World Wide Web, Internet, and Intranet, Orlando, FL.

Baeza-Yates, R., Caldeŕon-Benavides, L., & Gonźalez, C. (2006, 11-13 October). The intention behind web queries. *Paper presented at the String Processing and Information Retrieval (SPIRE 2006), Glasgow, Scotland.*

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004, 25-29 July). Hourly analysis of a very large topically categorized web query log. *Paper presented at the 27th Annual International Conference on Research and Development in Information Retrieval, Sheffield, U.K.*

Beitzel, S. M., Jensen, E. C., Lewis, D. D., Chowdhury, A., & Frieder, O. (2007). Automatic classification of Web queries using very large unlabeled query logs. *ACM Transactions on Information Systems, 25*(2), Article No. 9.

Brooks, N. (2004, July). *The Atlas Rank Report I: How Search Engine Rank Impacts Traffic.* Retrieved 1 August, 2004, from http://www.atlasdmt.com/media/pdfs/insights/RankReport.pdf

Brooks, N. (2004, October). *The Atlas Rank Report II: How Search Engine Rank Impacts Conversions.* Retrieved 15 January, 2005, from http://www.atlasonepoint.com/pdf/AtlasRankReportPart2.pdf

Choo, C., Detlor, B., & Turnbull, D. (1998). A behavioral model of information seeking on the web: Preliminary results of a study of how managers and IT specialists use the web. *Paper presented at the 61st Annual Meeting of the American Society for Information Science, Pittsburgh, PA.*

Chowdhury, A., & Soboroff, I. (2002). Automatic evaluation of world wide web search services. *Paper presented at the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland.*

Cothey, V. (2002). A longitudinal study of World Wide Web users' information searching behavior. *Journal of the American Society for Information Science and Technology, 53*(2), 67-78.

Croft, W. B., Cook, R., & Wilder, D. (1995, 11-13 June). Providing government information on the internet: Experiences with THOMAS. *Paper presented at the Digital Libraries Conference, Austin, TX.*

Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's web use skills. *Journal of the American Society for Information Science and Technology, 53*(14), 1239-1244.

Hargittai, E. (2004). Classifying and coding online actions. *Social Science Computer Review, 22*(2), 210-227.

Hilbert, D., & Redmiles, D. (1998, 10-13 May). Agents for collecting application usage data over the internet. *Paper presented at the Second International Conference on Autonomous Agents (Agents '98), Minneapolis/St. Paul, MN.*

Hilbert, D. M., & Redmiles, D. F. (2000). Extracting usability information from user interface events. *ACM Computing Surveys 32*(4), 384-421.

Hölscher, C., & Strube, G. (2000). Web search behavior of internet experts and newbies. *International Journal of Computer and Telecommunications Networking, 33*(1-6), 337-346.

Holst, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, Massachusetts: Perseus Publishing.

Jansen, B. J., & McNeese, M. D. (2005). Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *Journal of the American Society for Information Science and Technology, 56*(14), 1480-1503.

Jansen, B. J., & Pooch, U. (2001). Web user studies: A review and framework for future work. *Journal*

of the American Society of Information Science and Technology, 52(3), 235-246.

Jansen, B. J., & Spink, A. (2005). How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 248-263.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005, 15-19 August). Accurately interpreting clickthrough data as implicit feedback. *Paper presented at the 28th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, Salvador, Brazil*.

Kay, J., & Thomas, R. C. (1995). Studying long-term system use. *Communications of the ACM, 38*(7), 61-69.

McGrath, J. E. (1994). Methodology matters: Doing research in the behavioral and social sciences. In R. Baecker & W. A. S. Buxton (Eds.), *Readings in Human-Computer Interaction: An Interdisciplinary Approach* (2nd ed., pp. 152-169). San Mateo, CA: Morgan Kaufman Publishers.

Meister, D., & Sullivan, D. J. (1967). *Evaluation of User Reactions to a Prototype On-line Information Retrieval System: Report* prepared under Contract No. NASw-1369 *by Bunker-Ramo Corporation. Report Number NASA CR-918*. Oak Brook, IL: Bunker-Ramo Corporationo. Document Number N67-40083).

Montgomery, A., & Faloutsos, C. (2001). Identifying web browsing trends and patterns. *IEEE Computer, 34*(7), 94-95.

Page, S. (2000). Community research: The lost art of unobtrusive methods. *Journal of Applied Social Psychology, 30*(10), 2126- 2136.

Park, S., Bae, H., & Lee, J. (2005). End user searching: A web log analysis of NAVER, a Korean web search engine. *Library & Information Science Research, 27*(2), 203-221.

Penniman, W. D. (1975, 26-30 October). A stochastic process analysis of online user behavior. *Paper presented at the Annual Meeting of the American Society for Information Science, Washington, DC*.

Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech, 42*(11), 41-66.

Rose, D. E., & Levinson, D. (2004, 17–22 May). Understanding user goals in web search. *Paper presented at the World Wide Web Conference (WWW 2004), New York, NY, USA*.

Sellars, W. (1963). Philosophy and the scientific image of man. In *Science, Perception, and Reality* (pp. 1 - 40). New York: Ridgeview Publishing Company.

Skinner, B. F. (1953). *Science and Human Behavior*. New York: Free Press.

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

Wang, P., Berry, M., & Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 743-758.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*, 158-177.

Webb, E. J., Campbell, D. T., Schwartz, R. D. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive Measures in the Social Sciences* (2nd ed.). Boston, MA: Houghton Mifflin.

Webb, E. J., Campbell, D. T., Schwarz, R. D., & Sechrest, L. (2000). *Unobtrusive Measures (Revised Edition)*. Thousand Oaks, California: Sage.

Wolfram, D. (1999). Term co-occurrence in internet search engine queries: An analysis of the Excite data set. *Canadian Journal of Information and Library Science, 24*(2/3), 12-33.

KEY TERMS

Behaviorism: A research approach that emphasizes the outward behavioral aspects of thought. For transaction log analysis, we take a more open view of behaviorism. In this more encompassing view, behaviorism emphasizes the observed behaviors without discounting the inner aspects that may accompany these outward behaviors.

Ethogram: An index of the behavioral patterns of a unit. An ethogram details the different forms of behavior that an actor displays. In most cases, it is desirable to create an ethogram in which the categories of behavior are objective, discrete, not overlapping with each other. The definitions of each behavior should be clear, detailed and distinguishable from each other. Ethograms can be as specific or general as the study or field warrants.

Trace Data (or measures): Offer a sharp contrast to directly collected data. The greatest strength of trace data is that it is unobtrusive. The collection of the data does not interfere with the natural flow of behavior and events in the given context. Since the data is not directly collected, there is no observer present in the situation where the behaviors occur to affect the participants' actions. Trace data is unique; as unobtrusive and nonreactive data, it can make a very valuable research course of action. In the past, trace data was often time consuming to gather and process, making such data costly. With the advent of transaction logging software, trace data for the studying of behaviors of users and systems has really taken off.

Transaction Log: *An electronic record of interactions that have occurred between a system and users of that system*. These log files can come from a variety of computers and systems (Websites, OPAC, user computers, blogs, listserv, online newspapers, etc.), basically any application that can record the user – system – information interactions. For transaction log analysis, behavior is the essential construct of the behaviorism paradigm. At its most basic, a behavior is an observable activity of a person, animal, team, organization, or system. Like many basic constructs, behavior is an overloaded term, as it also refers to the aggregate set of responses to both internal and external stimuli. Therefore, behaviors address a spectrum of actions. Because of the many associations with the term, it is difficult to characterize a term like behavior without specifying a context in which it takes place to provide meaning.

Transaction Log Analysis: A broad categorization of methods that covers several sub-categorizations, including Web log analysis (i.e., analysis of Web system logs), blog analysis and search log analysis (analysis of search engine logs).

Unobtrusive Methods: Research practices that do not require the researcher to intrude in the context of the actors. Unobtrusive methods do not involve direct elicitation of data from the research participants or actors. This approach is in contrast to obtrusive methods such as laboratory experiments and surveys requiring that the researchers physically interject themselves into the environment being studied.

APPENDIX

Table 2. Taxonomy of user-system interactions (Jansen & McNeese, 2005)

State	Description
View results	Interaction in which the user viewed or scrolled one or more pages from the results listing. If a results page was present and the user did not scroll, we counted this as a View Results Page.
View results: With Scrolling	*User scrolled the results page.*
View results: Without Scrolling	*User did not scroll the results page.*
View results: but No Results in Window	*User was looking for results, but there were no results in the listing.*
Selection	Interaction in which the user made some selection in the results listing.
Click URL(in results listing)	*Interaction in which the user clicked on a URL of one of the results in the results page.*
Next in Set of Results List	*User moved to the Next results page.*
GoTo in Set of Results List	*User selected a specific results page.*
Previous in Set of Results List	*User moved to the Previous results page.*
View document	Interaction in which the user viewed or scrolled a particular document in the results listings.
View document: With Scrolling	*User scrolled the document.*
View document: Without Scrolling	*User did not scroll the document.*
Execute	Interaction in which the user initiated an action in the interface.
Execute Query	*Interaction in which the user entered, modified, or submitted a query without visibly incorporating assistance from the system. This category includes submitting the original query, which was always the first interaction with system.*
Find Feature in Document	*Interaction in which the user used the FIND feature of the browser.*
Create Favorites Folder	*Interaction in which the user created a folder to store relevant URLs.*
Navigation	Interaction in which the user activated a navigation button on the browser, such as Back or Home.
Navigation: Back	*User clicked the Back button.*
Navigation: Home	*User clicked the Home button.*
Browser	Interaction in which the user opened, closed, or switched browsers.
Open new browser	*User opened a new browser.*
Switch /Close browser window	*User switched between two open browsers or closed a browser window.*
Relevance action	Interaction such as print, save, bookmark, or copy.
Relevance Action: Bookmark	*User bookmarked a relevant document.*
Relevance Action: Copy Paste	*User copy-pasted all of, a portion of, or the URL to a relevant document.*
Relevance Action: Print	*User printed a relevant document.*
Relevance Action: Save	*User saved a relevant document.*
View assistance	Interaction in which the user viewed the assistance offered by the application.
Implement Assistance	*Interaction in which the user entered, modified, or submitted a query, utilizing assistance offered by the application.*
Implement Assistance: PHRASE	*User implemented the PHRASE assistance.*

continued on following page

Table 2. (continued)

State	Description
Implement Assistance: Spelling	*User implemented the SPELLING assistance.*
Implement Assistance: Previous Queries	*User implemented the PREVIOUS QUERIES assistance.*
Implement Assistance: Synonyms	*User implemented the SYNONYMS assistance.*
Implement Assistance: Relevance Feedback	*User implemented the RELEVANCE FEEDBACK assistance.*
Implement Assistance: AND	*User implemented the AND assistance.*
Implement Assistance: OR	*User implemented the OR assistance.*