Chapter IX
The Unit of Analysis and the Validity of Web Log Data

Gi Woong Yun
Bowling Green State University, USA

ABSTRACT

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.

INTRODUCTION

One of the main motivations of Internet content providers in expanding the availability of multimedia is the perception that the Internet provides unparalleled access to accurate usage data. It is generally felt that the Web log traces left by individual Internet users provide unprecedented quantity and quality of information to researchers and to those who would study consumer and market behavior. However, there were some warning signs about the validity of Web log data (Goldberg, 2001). This chapter will discuss most of the validity problems, but it should be noticed that many studies (e.g., Davis, 2004; Eveland & Dunwoody, 1998a; Eveland & Dunwoody, 1998b; Jansen & Resnick, 2005; Phippen, 2004) paid only minor attention to the validity of Web log data during the analysis. This might be because it is expected that the Internet use data collected from computers will provide precise and detailed information about users’ Internet use behavior.
The Unit of Analysis and the Validity of Web Log Data

(Eveland & Dunwoody, 1998a). Indeed, it is a reasonable assumption that Internet use behavior tracked by computer software will be more valid than previous media use tracking methods. This high expectation of validity is due to the pinpoint accuracy of the client computers’ or server computers’ data collection software.

Some researchers suspected the usefulness of the transaction log data (Peters, Kurth, Flaherty, Sandore, & Kaske, 1993; Kurth, 1993; Larson, 1991). Others argued that data structures and a complex collection algorithm should be explored for the meaningful analysis of the data, as this contributes greatly to the data quality and quantity (Phippen, 2004). For instance, a unit of analysis of the data needs more attention before scientific analysis of the Internet use data. Deciding a proper unit of analysis is difficult and it will influence predicting and including analysis units ahead of data collection.

The unit of analysis of Web site use can differ depending on researchers and the research topics. Hence, the unit of analysis of Web site use can be examined with various levels of analysis. Any research will need to choose a level or levels of analysis when they want to use Web log data to analyze user’s navigation patterns or content access habits. The researcher’s research concept will be a major factor determining the level of analysis. However, technical specifications of the Web log data sometimes limit what researchers can select as a unit of analysis for their research. Although many people have expectations of accuracy in Web log data, typical Web log data, both Server-side and Client-side data, have limitations and strengths.

The validity of Web logs cannot be taken for granted and there is much to learn about how to collect and accurately interpret online activity. This chapter will propose criteria in defining units of analysis of the Web site use with a media research paradigm after examining some theoretical frameworks of media use measurement.

A UNIT OF ANALYSIS

Many researchers already utilized Internet log data to understand individual patterns of knowledge seeking via the Internet. They created variables to track which Web pages users have visited (e.g., Eveland & Dunwoody, 1998a; Eveland & Dunwoody, 1998b; Phippen, Sheppard & Furnell, 2004), what users have queried (e.g., Jansen & Spink, 2005; Jansen, Spink & Pederson, 2005; Jones, Cunningham, McNab & Boddie, 2000; Sandore, 1993; Taha, 2004), what they wrote while they were using a computer, who they communicated with, what they communicated, or how they communicated (e.g., McTavish, Pingree, Hawkins, & Gustafson, 2003; Phippen, 2004). These units of analysis of Web site use have been operationalized based on the availability of Web log data.

Measurement Units

Internet use is different from watching a network TV program where millions of television viewers share a limited number of variations of channel surfing patterns. Each Internet user uniquely engages in non-linearly structured cyber space. Therefore, it is not an easy task to record and analyze all users’ navigation behavior. However, some measurement units within a computerized recording system can be traced. The analysis units can be the amount of time spent during the navigation or the number of computer files accessed.

Time

One of the most frequently measured units in media research is time. The sheer volume of time exposure has been investigated since the beginning of the media research field. Survey respondents are asked to answer questions like ‘how many hours did you spend reading a newspaper per week?’, ‘how many hours did you watch television last week?’, or ‘how many hours