ANALYSIS OF WEB SITE USAGE DATA: HOW MUCH CAN WE LEARN ABOUT THE CONSUMER FROM WEB LOGFILES?

Balaji Padmanabhan
Shahana Sen
Alexander Tuzhilin
Norman H. White
Roger Stein

IS-96-18

ANALYSIS OF WEB SITE USAGE DATA: HOW MUCH CAN WE LEARN ABOUT THE CONSUMER FROM WEB LOGFILES?

Balaji Padmanabhan
Doctoral Student
Dept. of Information Systems, Leonard N. Stern School of Business, NYU
44 West 4th Street, New York, NY 10012-1126
Tel: (212) 998-0812, Fax: (212) 995-4228
E-mail: bpadmana@stern.nyu.edu

Shahana Sen
Assistant Professor
Dept. of Marketing, Graduate School of Business Administration
Fordham University
113 West 60th Street, New York, NY 10023
Tel: (212) 636-6133

Alexander Tuzhilin
Associate Professor
Dept. of Information Systems, Leonard N. Stern School of Business, NYU
44 West 4th Street, New York, NY 10012-1126
Tel: (212) 998-0832, Fax: (212) 995-4228
E-mail: atuzhili@stern.nyu.edu

Norman White
Associate Professor
Dept. of Information Systems, Leonard N. Stern School of Business, NYU
44 West 4th Street, New York, NY 10012-1126
Tel: (212) 998-0842, Fax: (212) 995-4228
E-mail: nwhite@stern.nyu.edu

Roger Stein
Vice President, Moody's Investors Service and
Doctoral Student
Dept. of Information Systems, Leonard N. Stern School of Business, NYU
44 West 4th Street, New York, NY 10012-1126
Tel: (212) 998-0800, Fax: (212) 995-4228
E-mail: rstein@stern.nyu.edu

December 1996

Working Paper Series
Stern #IS-96-18

Analysis of Web Site Usage Data: How Much Can We Learn About the Consumer From Web Logfiles?

Abstract

We discuss information needs of marketers on the World-Wide Web and present a classification of types of information one can possibly get about Web site visits based on its ease of gathering. We then analyze how information needs can be satisfied using different categories of information gathered with these varying degrees of ease. We conclude that, although some of the information needs of marketers can be satisfied using data that can be automatically gathered with state-of-the-art Web tracking methods, many others cannot. We also discuss relevant issues and potential solutions to this problem.

Center for Digital Economy Research
Stern School of Business
Working Paper IS-96-18

1. Introduction

Marketing has been revolutionized by the availability of new and increasingly diverse sources of information on consumers and consumer behavior. Masses of data are being collected for use in marketing by marketing research agencies, which include the single source data (household level purchase data from supermarket scanners and demographic and television viewing behavior data from household panels) besides the more traditional consumer focus group and survey data on attitudes, purchase intention, purchase behavior, satisfaction etc. to name a few. In addition, individual firms themselves collect volumes of internal transactional data through their sales, marketing, operations, and back-office activities (e.g. through their on-line transaction processing systems).

In addition to these sources of data collection, one of the newest and perhaps most unique sources of marketing data are the *logfiles* generated as a result of consumers and clients accessing information from an organization's World Wide Web (WWW) site. These logfiles *automatically* record tracking information pertaining to a Web site *visitors'* activities at the Web site and may contain potentially valuable information on existing and prospective new customers.

In this paper, we investigate the degree to which a typical company's market segmentation information needs pertaining to their Web site visitors might be satisfied by the information in standard Web logfiles, and the augmented logfiles created using more advanced Web usage tracking technologies. Since Web logfiles contain data primarily on visitors' browsing behavior at the site, this data has to be coupled with other types of data collected through the traditional methods, for marketers to get a more comprehensive picture of their Web site visitors. This is analogous to scanner data on purchase behavior being augmented with other data to produce the single source data. So, in this paper we also look at how these other data may fill up some of the gaps left by the Web logfile data in providing the market segmentation information marketers need.

In order to understand how Web logfiles might be used to answer questions about consumers, we begin by discussing the types of questions a marketer may traditionally ask of a consumer for market segmentation purposes. Based on these categories of information that marketers traditionally use for segmenting consumers, we illustrate analogous Web related information that may be used for understanding Web site visitors and their Web usage behavior. Once we establish this foundation, we go on to briefly explore the various technologies that are currently available to collect information in Web logfiles, and how these logfiles may capture additional data which marketers need using technological augmentations.

We note that the technology for capturing data in Web logfiles is still immature. In addition to technological advances, the actual technical standards agreed upon by WWW standardization committees will impact greatly the growth and viability of Web logfile data usage for marketing research. Furthermore, privacy issues will also direct the course of this type of research. We touch briefly upon some of these issues in this paper.

Finally we present a mapping of the extent to which the augmented Web logfile data and the data which marketers may collect using traditional marketing research methods on their Web site visitors, may satisfy a marketer's information needs for segmenting these consumers. We also speculate on how this may change in the future when some of the privacy issues are addressed.

Finally, we believe that our paper contributes to the understanding of an area which is yet emerging, but may soon become an important information resource for those who are able to make use of it.

Marketing and the WWW

Before we examine these issues, however, it is useful to consider why the World Wide Web has captured the interest of so many firms. The WWW is the latest entrant in the communication and direct marketing media alternatives available to marketers, and in the words of Hoffman and Novak (1996), "For several years, a revolution has been developing that is dramatically altering (the) traditional view of advertising and communication media. This revolution is the Internet..." Although it was originally the purview largely of technologists and academics, the Internet and its first and current networked global implementation, the WWW, has become a medium utilized by a broad spectrum of individuals, from experienced business people to computer novices. Hoffman and Novak (1996) also note that despite the strong belief in business circles that the Web represents a phenomenal marketing opportunity, not much scholarly attention has been paid to it for understanding it as a medium for marketing communications and as markets in and of themselves.

Large numbers of commercial organizations have put up Web sites to avail of the opportunities the Web promises. By one estimate, $312 million ad dollars were spent on advertising on the WWW in 1996, and this figure is estimated to grow to $5 billion by the year 2000 (Business Week, September 23 1996). According to one survey, one fifth of randomly selected senior managers nationwide reported that their organizations currently had Web sites, and another one third indicated that they planned to launch a Web site within the next six months (Internet Marketing, May 1996).

Organizations have a variety of overlapping motivations for creating Web sites. Web sites allow firms to communicate directly with customers, to build brand awareness, position, and value, and to create new sales channels. Correspondingly, of the respondents to the above survey who have or plan to have Web sites, 31% reported that full or partial responsibility for content development fell on the marketing department, rather than the MIS department, as might be expected. The sales department and marketing communications departments had responsibility 22% and 11% of the time, respectively.

As a medium for communication the WWW's strength is that it allows businesses to receive direct feedback to their communications in a relatively effortless way (Sterne 1995). The Web

site establishes a one-to-many link between an organization and the receivers of the message (visitors to the site) whereby it is possible for both to communicate with each other. This *interactive* link with the customers which the Web site provides allows a company a fast means for conveying critical information about its products and services to customers and an equally fast means for getting input back from them.

In this paper, we discuss the challenge of harvesting the potentially rich source of marketing data represented by these Web site interactions. The remaining sections of the paper are organized as follows. In Section 2 we describe the information needs of marketers about visitors to their Web site and present a classification of these needs. In Section 3 we discuss the current technological practices used in analyzing Web site usage data. We then present a mapping in Section 4 that examines how the information needs of marketers can be satisfied using both Web site usage data and additional data that may need to be collected. We present our conclusions in Section 5.

2. Information Required by Marketers about Visitors to their Web Site

In this section we classify the customer analysis information needs of marketers pertaining to visitors to their Web sites, for doing market segmentation. Traditionally, market researchers have looked at market segmentation variables such as demographic, geographic, psychographic, behavioral, and benefit (Kotler, 1997; Churchill & Peter,1995):

In the context of the Web as a medium for marketing and advertising, marketers will be interested in learning about the consumers who visit their Web sites both with respect to the traditional segmentation variables, as well as aspects which unite or differentiate these visitors on their Web usage needs and behavior -- such as the nature of information sought at the site, their temporal usage patterns of the Web site, their navigation within the Web site and their purchase behavior at the site. We combine the Web-specific information needs with the traditional segmentation variables, and classify the information needs on Web visitors into five groups as shown below.

i. Demographic and Psychographic variables describe **Who** the consumers (visitors) are;
ii. Geographic variables describe **Where** the consumers (visitors) come from;
iii. The Behavior variables describe **What** the consumers (visitors) do and **When** they do it;
iv. The Benefit variables describe **Why** the consumers (visitors) behave the way they do.

American Demographics' characterization of the traditional market segmentation information needs is quite similar. For further discussion on this, please refer to American Demographics,

"The Insider's Guide to Demographic Know-How" (Ithaca, NY: American Demographics Press: 1990).

These five variable groups (**Who, Where, What, When,** and **Why)** specify information needs of marketers pertaining to their Web site visitors. We call this the *5W's classification* of the World Wide Web consumer information needs. Table 2.1 summarizes this:

Table 2.1. The 5 W's Classification of the WWW Consumer Information Needs.

The 5 Ws	Traditional Market Segmentation Information	Web Site Visitor/Consumer Segmentation Information
Who are they?	Demographics and Psychographics	About the visitor as a Web site user
Where are they?	Geographic	About the visitor's Internet address/access path, etc.
What do they do?	Behavior	About the specific information sought
When do they do it?		About Web site usage temporal behavior
Why do they do it?	Benefit	About the reason/motivation for visiting the Web site

The information on these 5W's may be used by marketers to make various decisions about their marketing strategy on the Web, as well as for marketing using traditional channels. For example, Levis Jeans advertises on the Web sites "Women's Wire" and "Sports Zone" and leverages the information it learns about its Web site visitors to make decisions about their entire 50M media budget (ARF Interactive Media Research Summit II, July 1996).

Tables 2.2-2.4 detail the information needs of marketers under the above categories by dividing each category of information needs in Table 2.1 into several sub-needs. Later we will evaluate which of the individual sub-needs would be feasible for marketers to satisfy from the Web log-files and other related sources.

Table 2.2 details the information needs pertaining to understanding the visitors by identifying their demographic and psychographic descriptors. The Web specific information needs consist of identifying the visitor in terms of their Internet addresses (e-mail address, homepage address etc.), understanding their Web usage personality -- i.e. are they willing to give information about themselves over the Web or buy products using their credit card, or do they consider it risky.

Table 2.2: Who Are They?

A. Traditional Market Segmentation Information Needs

On Individual Visitors:
(i) Demographic: Name, Age, Sex, Income, Education, Occupation, Family size, Stage in family life cycle, Race, Religion, etc.
(ii) Psychographic: Lifestyle (i.e. activities, interests, opinions), Personality, etc.
(iii) Level of Expertise about the product/service.

On Corporate Visitors:
(iv) Name, Size (Annual Sales, Market Share, Number of Employees), Nature of business, Product lines/Brands etc.
(v) Customer/Competitor/Others (e.g. Web site developers, content providers, maintainer of Web sites, etc.)

B. Web Site Visitor/Consumer Information Needs

(i) Visitor's email address, homepage address.
(ii) Browser software used by the visitor (browser specific features may limit the interaction that a user can have at a Web site).
(iii) Willingness to give information on themselves, as well as feedback about product over the WWW
(iv) Willingness to buy products using a credit card over the WWW

Table 2.3 details the information needs for identifying the geographic and Internet location of the visitor, such as where they are located in physical/geographical space and in "cyberspace."

Table 2.3: Where Are They?

A. Traditional Market Segmentation Information Needs

Individual:
(i) Geographic: Region, City or Metro size, Density (urban/suburban/rural), Climate.

Corporate:
(ii) Geographic details on where the organization is located.

B. Web Site Visitor/Consumer Information Needs

(i) Where did/do they access the company's Web site from -- direct (by typing in a URL) or through links advertisements or banners (from which site?)
(ii) Where from do they access the Internet (home/office/other)?

Table 2.4 illustrates information needs a marketer will have for segmenting consumers or visitors based on their product consumption or Web site usage behavior. A visitor may browse through the information on the site, buy products, add the site to their bookmark so they can revisit it easily, etc. Besides the reasons behind their visit (included in Table 2.6), the Web usage behavior can differ between visitors in terms of their level of Web expertise which determines the ease with which they can find the site and the information they are looking for in the site. First time visitors may need more help in finding information within the site; repeat visitors may want to be able to access a particular information directly rather than going through other pages in the site.

Table 2.4: What Do They Do?

A. Traditional Market Segmentation Information Needs[1]

(i) Behavior:
- -- Consumption Occasions
- -- User status (nonuser/ex-user/potential user/first-time user /regular user)
- -- Usage rate (light/medium/heavy product user)
- -- Loyalty status (None/medium/strong /absolute)
- -- Buyer-readiness stage (Unaware of/Aware of/Prefer/Intend to buy product)
- -- Attitude toward product (enthusiastic/positive/indifferent/negative/hostile)

B. Web Site Visitor/Consumer Information Needs

(i) First time or repeat visitor?
(ii) What files did/do they access? Is the information gathered related to product specifications, dealer or retailer locations, ordering product, customer service, sending mail to company, consumer interest information, etc.
(iii) If repeat visitor: how many previous visits, is the pattern of site usage (file accesses) same as in the past etc.
(iv) Level of Web expertise. For example: Did they conduct the navigation and search efficiently? Did they encounter any problems accessing information at the site?
(v) Did they download/print any files?
(vi) Did they purchase any product(s)?
(vii) Did they give any feedback/send mail to company?
(viii) Did they add the site to their bookmark?

Table 2.5 describes the Web specific information needs of a marketer for understanding the Web site usage temporal behavior.

[1] The variable in this category -- Behavior, may be determined by asking both the questions "What do they do" and "When do they do it" about the consumers. However, in case of the Web site user profile, we have split this into two separate categories.

Table 2.5: When Do They Do It?

A. Traditional Market Segmentation Information Needs
Same as in Table 2.4 above.
B. Web Site Visitor/Consumer Information Needs
(i) What time of the day did/do they visit site? (ii) How much time do they spend at site/file? (iii) Are their visits ir/regular (e.g. in terms of number of days, time of day, duration of stay, path of access etc.)

Table 2.6 illustrates the information needs for understanding the reasons or motivations behind product use or visiting the Web site. Also this table addresses questions that a marketer may have regarding the effectiveness of current advertising or promotions of their Web site.

Table 2.6: Why Do They Do It?

A. Traditional Market Segmentation Information Needs
(i) Benefits: -- The Need(s) which the product satisfies -- Quality of the product -- Service provided by the company -- Economy associated with buying the product -- Convenience or Speed of use.
B. Web Site Visitor/Consumer Information Needs
(i) Casual browser/serious searcher. (ii) Saw some specific advertisement or promotional campaign outside the Web. (iii) Benefits sought from the Web site - product information gathering and evaluation, purchasing products, customer service etc.

An Illustrative Example

We illustrate the uses of the above information in a retail setting for marketing strategy determination using the example of a wine and liquor retailer. An article in the New York Times (September 25, 1996) says "(that) it does seem that wine and the Web were made for each other" and already a search through Lycos provided them access to 31,365 documents related to wine. For the shopper and the enthusiast, it provides a reach over time and space making it possible to

get the best deals or the most esoteric store-house of facts. The author believes that in one not-too-distant day, the Internet may strikingly change the way people shop for wine, since already it is possible to use the Web to track down "the best deals on oaky chardonnays for less than $10 ... learn about wine storage..." and get other information of interest to the wine consumer which by nature may be hard to find, large scale, or ever changing -- making the Web and its interactivity an ideal medium for searching for these. Consumers can also expect better service from retailers fulfilling their purchase orders, since if anything is not right, it is much easier for the customer to complain about it on the Internet and others knowing about it.

The company we describe is among the largest retailer for wines and liquors in New York State, and the largest retailer for wines from Long Island, New York. To reach consumers, the company advertises on cable TV, newspapers and radio, and has occasionally tried direct mail -- of which their experience has not been very satisfactory. As its latest media vehicle, the company decided to advertise on the WWW. It has recently put up a Web site with information and ordering capability for its wines from Long Island (URL: http://www.mcadam-buyrite.com). The following characteristics of the Web site are noteworthy:

1. The site features a consumer product that is appealing to a broad segment of consumers.
2. The company has put up the site with the objective of informing people about the different vineyards in Long Island and the wines they produce, for selling these wines to customers over the Internet as well as building traffic for its store.
3. The site is relatively large (has several files with a lot of information) and is well designed.

According to their President, the key issues concerning the Web site for the company currently are:
1. How to increase traffic to their site
2. How to increase business from their site and their store
3. How to determine "customer satisfaction" with respect to the quality and accessibility of the information on the site
4. How to get information about the visitors to their site which will help them in planning their direct marketing strategy on the Web.

To act upon the first issue, "How to increase traffic to the site", the company needs to understand the reasons why people may visit their site, the search words and the strategy they use to access their site and other related sites, etc. This knowledge would be used to design their advertising strategy for the site in order to attract new visitors, and to design the site itself to satisfy the first-time visitors so that they become repeat visitors. Decisions on its advertising strategy include:
(a) If and where to advertise the site on the Web such as Web zines (e.g. Smart Wine Web zine) ,Web sites (e.g. Virtual Vineyards, sites for complementary products such as gourmet dining etc.), service providers (e.g. America Online's Wine Locator board) or on browser sites (e.g. Netscape, Mosaic, Microsoft's Web Crawler, etc.); and the economics of using these options.
(b) If and where to advertise in the traditional media, such as community and local newspapers, direct mail, local TV channels, etc., and the economics of using these options;

(c) Which Web sites to have links with, so that people from those sites may access its site.[2]

For the above analysis, the categories of information which the company will find particularly relevant would be:
(i) *Who*: Demographic & Psychographic information on the visitors (for determining the right advertising message and media for communicating with them); whether they are first time/ repeat visitor (to learn about what type of customers are repeat visitors for example); whether they are the company's existing store customer (to learn whether they find it convenient to purchase using the Web); etc.
(ii) *Where*: Geographic; where did they access the company's site from -- direct (using URL) or using a particular search engine (e.g. Yahoo, Infoseek, Lycos, etc.); for advertising in the appropriate places
(iii) *What*: What files did/do they access -- related to information gathering (wine listings with descriptions/ price list/ retail location address), related to ordering product, customer service, sending mail to company, etc.; did they give any feedback /send mail to the company; did they add the site to their bookmark (to learn what they visit the site for to improve the design and contents of the site.)
(iv) *Why*: Why did they visit the site -- casual/serious visitor; did they see some specific ad or came to the site with the help of a search engine; did they come for information gathering or intention to purchase; if for purchase, are they going to use the web as their sole method of purchasing, etc.

In this section we classified the information needs of marketers pertaining to their Internet customers. In order to understand how these needs can be satisfied with current technology, we first describe the state-of-the-art in these technologies.

3. Information Tracked in Web Logfiles

Each time a user accesses a web site, the server on the web site automatically adds entries to files called *logfiles*. These files therefore summarize the activity on the web site and contain useful information about every web page accessed at the site. The exact nature of the information captured, however, depends on the specific software (the *Web server*) that the web site uses. This section summarizes the nature of information that is tracked by the common log files and discusses common practices for gathering additional information about users' accesses to Web servers.

3.1 Information Tracked in Log Files

[2] This decision needs careful consideration since typically linkages are mutual, i.e. the company would also be obliged to provide links to these other sites from their own. Therefore the company would want to select sites which are compatible with its image and positioning for example.

The four common log files generated by Web servers are *access_log*, *error_log*, *referrer_log* and *agent_log*. Earlier versions of Web servers such as NCSA httpd1.0 maintained access and error logs only. More recent versions of several popular Web servers track additional information by also creating referrer and agent logfiles. This section describes the information tracked in each of these logfiles.

Access_log File.

In the context of tracking web access, the access_log file is the most important log file, since it keeps details of transactions involving every single file on the web server that was transmitted to a client accessing the site.

Access statistics generated by Web servers provide some information on each *hit* to a web site. A hit is a request for any file maintained on a Web server. A hit is *not* equivalent to a request for a Web page since a Web document could consist of a collection of different files. For example, if a web page consists of some text and an image (e.g. a "gif" picture), the text and the image are stored in separate files and both files need to be transmitted to the user when the page is accessed. This could therefore result in two hits recorded in the access_log file. Each time a user requests the web page, *all* the files included in the page get transmitted to the client and each file transmission results in one hit recorded in the access_log file.

The common access_log file has the following format:

```
(remote-host, remote_log name, username, date, request, status,bytes)
```

An example of an entry in an access_log file is:

```
128.96.134.65    unknown    -    [09/Aug/1996:20:13:13    -0400]    "GET
/istemp/hmsproj.html HTTP/1.0" 200 1876
```

where the fields in the example are:

(1) *remote host* = 128.96.134.65.
This is the remote hostname, or *IP address,* from where the request originated. This field indicates *where* (with respect to Internet network addresses) an access came from. This field could alternately be recorded as a "domain name", such as "xyz.netscape.com" (here it may be inferred that the client is from Netscape or has access to login to the Internet from Netscape's network).

(2) *remote_log name* = "unknown" and *username* = "-".
These fields are intended to be placeholders for the user's machine name and the user's name, as determined by an authentication protocol titled "Request For Comments #931" (see http://ds.internic.net/rfc/rfc931.txt for a detailed discussion). These fields were included to permit the client to *optionally* send in these values to a server. In practice, very few servers supply any of these fields and the values returned are usually as shown in this example ("unknown" and a "-"). Rarely do these fields contain useful information, and

most of the Web Site Usage tracking tools that attempt to infer the user do not rely on these fields to provide them with any relevant information.

(3) *date* = 09/Aug/1996:20:13:13 -0400.
This field contains the time and date when the access was recorded. The "-0400" in the example indicates that the time recorded is in a local time which is four hours behind GMT.

(4) *request* = "GET /istemp/hmsproj.html HTTP/1.0".
The *request* consists of three parts:
(i) The URL received by the Web server. On the user's browser this may appear in greater detail such as "http://is-2.stern.nyu.edu/istemp/hmsproj.html". The access_log records only that portion of the URL that specifies the physical location of the file within the Web server (in this case, "/istemp/hmsproj.html" - the preceding part, "http://is-2.stern.nyu.edu" only serves to specify the Web server name and is not recorded since this portion would be the same for every hit recorded in the access_log file).
(ii) The specific Web protocol that the client used ("HTTP 1.0") and
(iii) The nature of request that the client made to the web server. For example, when the client retrieves a complete document this field would be "GET" and when the client interacts with the Web server by filling out a form on the site this could be "POST".

(5) *status* = 200.
This field indicates the status of that specific transaction, which may for example indicate whether the transmission was successful or if any error had occurred. A successful transaction generates a "200" response code. If an error had occurred, more detailed information will be logged on to the error_log file.

(6) *bytes* = 1876
This field contains the number of bytes transmitted. In this case this indicates the size of the file *hmsproj.html* that was sent to the client.

Figure 3.1 contains some entries in an access_log file created by a Web server (NCSA httpd 1.1). The entries listed permit the identification of the domain from which the request originated, the date and time and the file accessed (including the number of bytes transmitted and a status code). However, some of the common problems and limitations in tracking accesses from plain access_log files manifest themselves in these entries. These include:

- Inability to trace each access to a specific user. The access_log file indicates the hostname from where a request originated. A hostname may be as general as "xyz.netscape.com". Typically there are a large number of users who may access a web site from the same host, and the hostname, therefore, cannot be used to identify individual users.
- Inability to track client sessions. A session consists of a set of web pages that a user accesses during a continuous period of time that the user spends at the web site. It is not possible to decide whether the last two entries (the third and the fourth records) in Figure 3.1 came during the same client session. Both these accesses are from the same hostname

(jacobs90.nmsu.edu), but from the entries, it cannot be determined if these accesses were from the same user session.

- Incomplete tracking of access information. The third record in Figure 3.1 indicates that a file called "chess-page.html" was requested by a user from the domain "jacobs90.nmsu.edu" on August 9 at 20:14:03. If there had been another request for the same file on the same day at 20:14:06 (minutes *after* this recorded request), it is possible that the new request *may not appear* in the access_log file. This may happen due to the *caching* of documents by another server that maintains copies of files recently accessed from the web site. The advantages of caching recently accessed documents by an intermediary server include faster response time, lesser processing load on the web server and lower network load. The problem with respect to tracking web usage is that this request would not get recorded in the Web server's access_log since the request never reaches the Web server (the intermediary server processes this request).

```
128.96.134.65    unknown    -    [09/Aug/1996:20:13:13    -0400]    "GET
/istemp/hmsproj.html HTTP/1.0" 200 1876

117.new-york-001.ny.dial-access.att.net unknown - [09/Aug/1996:20:13:42 -
0400] "GET /~mbloch/dilbert.htm HTTP/1.0" 200 698

jacobs90.nmsu.edu    unknown    -    [09/Aug/1996:20:14:03    -0400]    "GET
/~vboykod/classes/chess/chess-page.html HTTP/1.0" 200 4544

jacobs90.nmsu.edu    unknown    -    [09/Aug/1996:20:14:05    -0400]    "GET
/~vboykod/classes/chess/knight.gif HTTP/1.0" 200 850
```

Figure 3.1. Examples of entries in an access log file

Error_log File.

The error_log file is useful to track errors that occur during an interaction between a client and a server. The common format for error_log files is:

```
(Timestamp, Error Message)
```

where *Timestamp* specified the time when an error occurred and *Error Message* specifies details pertaining to why the error occurred. Figure 3.2 lists entries in an error log file generated by NCSA httpd1.1.

```
[Mon Jul  8 21:20:58 1996] httpd: access to /usr/local/stern/web/cgi/get
failed for crab.icsi.berkeley.edu, reason: file does not exist

[Tue Jul  9 07:33:29 1996] httpd: send timed out for cha-
nc210.ix.netcom.com

[Mon Jul  8 21:22:29 1996] killing CGI process 29001
```

Figure 3.2 . Examples of entries in an error log file

The errors listed in the records in Figure 3.2 indicate unsuccessful transactions due to non-existence of a requested file, incomplete transfer of a file and the termination of a program on the server respectively. Web site administrators frequently analyze error_log files as part of managing the web site. The information tracked in the error-log file is therefore relevant primarily for web site administrative purposes.

Referrer Log.

The referrer_log file lists the web page that a user accessed prior to accessing any web page on the site. The information tracked in the referrer_log file can be used to determine *where* (from which site or web page) a user came to this site from and can therefore be of value for advertising purposes. For example, if a web site has an advertisement placed on some other page on the Web (e.g. at http://xyz.ad-site.com), the referrer_log file can be used to determine the visits to the web site from that page (e.g. visitors coming to the site from http://xyz.ad-site.com). The common format for referrer_log files is:

```
(Timestamp, Remote-URL, Local-File)
```

Figure 3.3 lists sample entries to the referrer_log file. The first entry, for instance, indicates that a user who was previously browsing Netscape's homepage, accessed the file "products.html" at the web site. This could indicate one of the following possibilities:
(a) There is an explicit link to the file "products.html" from the previous page (http://home.netscape.com). This explicit link could, for example, be due to an advertisement placed on the site or due to a link that the owner of the previous site decided to include on his/her page.
(b) There may be no explicit links from the previous page to the file "products.html", but while browsing through the previous page, the user explicitly opens the URL by typing the entire Web address (or even by choosing the link from a bookmark).

```
[02/Oct/1995:13:03:29] http://home.netscape.com/ -> /products.html/

[02/Oct/1995:13:03:54] http://www.lycos.com/ -> /products.html/
```

Figure 3.3 . Examples of entries in a referrer_log file

An interesting use of the referrer_log file is determining which sites on the web "point to" or have links to a web site. If a referring URL appears frequently in the referrer_log file, it could probably indicate that there is an explicit link from the referring site to a file on the web site that maintains the log. However, confirmation of this hypothesis would involve having to examine the page whose address is the referring URL.

Agent Log.

Agent log files capture the browser type of the client during an access. The common format for the agent_log is :

```
(Timestamp, Browser-Type)
```

Figure 3.4 lists example entries in an agent_log file. All the entries in the figure indicate that a user accessed the site using Netscape browsers (different versions though).

```
[02/Oct/1995:12:00:28] Mozilla/1.1N (X11; I; SunOS 5.3 sun4c)

[02/Oct/1995:12:00:29] Mozilla/1.0N (Windows)
```

Figure 3.4 . Examples of entries in an agent_log file

The information tracked in the common logfiles maintained by Web servers can be used to generate several summary statistics that summarize the activity at the web site. Many publicly available programs create summary tables and charts automatically from these data.

3.2 Problems and Limitations of Log file data

While Web logfiles contain certain information about visitors' accesses to a Web site, they do not provide some important information that marketers would need. These limitations relate primarily to the inability of identifying each access to a specific user, and the difficulty in tracking when a session begins and ends. These two limitations are discussed below:

(1) Inability to trace each access to a specific user: User identification issues arise due to limitations of the access_log file. While entries in this file indicate the hostname where a request originated, this is seldom enough to identify a particular user. A hostname may be as general as "xyz.unilever.com." The hostname in this case indicates that the access was from some computer at Unilever PLC. Unfortunately, it does not specifically indicate who at Unilever the user was. Typically there are a large number of users who may access a Web site from the same host, and as a result the hostname cannot be used to uniquely identify individual users.

Nonetheless, being able to track a unique user at a Web site is central to satisfying many of the marketing information needs outlined in Section 2. Fortunately, this limitation may be surmounted to some extent. There are two different levels at which a unique user could be tracked : (a) anonymously, where a user is nothing more than a unique user ID (that remains the same whenever this user accesses the site) or (b) in addition to a user ID, some knowledge about the user (such as his/her name, email address etc.) is also maintained. There are some privacy concerns regarding tracking individual users on the Internet. We will discuss some issues related to privacy concerns in Section 4.3.

(2) Inability to track client sessions: Session identification issues arise for slightly different rea-

sons than user identification issues. A session consists of a set of Web pages that a user accesses during a continuous period of time that the user spends at the Web site without leaving it[3]. For example, say that a certain company's Web site allows users to purchase products while browsing through a catalog of Web pages. In order to keep track of the products purchased by each user during their respective "shopping sessions", it is essential that the site keeps track of all individual user sessions (including identification of the beginning and the end of a user's session at the site, the list of products that a user has added to his or her personal "shopping basket" in that session etc.). In this case it is not necessarily important to identify the users, but only to be able to track specific sessions using a unique session ID.

Though plain logfile data do not permit the tracking of sessions or users, there are certain technical methods that can be used to attempt both session-tracking and user-tracking. Tracking user-sessions, however, is "easier" than tracking individual users (since it is a sub-problem of tracking users).

3.3 Technical Methods for Gathering Additional Information

In this section we describe a few common techniques by which Web servers attempt to gather additional user access information. The two primary problems that these methods aim to address are (a) identification of individual users as they browse a site and (b) identification of user sessions. In this section we describe techniques such as using *tokens* and *cookies* and discuss the need for heuristics in session identification.

3.3.1 Tokens

A web page may contain several "links" inside the page that a user could follow. Each link contains a URL address that is sent to the web server when the link is selected. A URL typically contains the name of a webpage on the server.

Tokens are strings of characters that are appended to these URL addresses in links. By appending the same token to every link that a user follows within the site, a Web server can infer the user's path through the site and can, therefore, track the user's session. A Web server uses tokens to track sessions as follows. When the web server receives a request for a file (in the form of a URL address) the server parses the request to separate the token from the URL address. Before sending the requested page back to the user, the server inserts this parsed token into the links in the new (requested) web page. In this manner, if any link on the new page is selected by the user, the same tokens are sent back to the web server.

It is important to note that in order for the web server to parse out tokens from URL addresses and to make inferences based on the value of a token, the server must use additional software or a

---

[3] In practice it is usually a non-trivial task to identify when a user's "session" ends. A heuristic used by some commercial packages assumes that a user's session ends if the time difference between successive accesses by the same user is greater than thirty minutes.

program that performs this preprocessing of URL addresses. These programs are often not part of standard web server software and may need to be written by administrators and programmers at each web site or purchased from a vendor.

For tokens to be able to identify unique users, the web server would have to infer the individual identity of users and append this identity string as a token. This is possible if the user conveys his/her user ID (presumed unique) to the server during the first access in a session. The server could then use this user ID as the token that gets appended to links in web pages. Web sites that require users to "login" to the site by providing a username (and password) can use the supplied user ID as a token. Such sites would therefore be able to identify both unique users and user sessions at the site. A drawback of this method is that requiring users to register and login to access the content of a web site may prevent many users from visiting the site.

3.3.2 Cookies

Cookies are an open specification supported by Netscape Communications Corporation that can be used to maintain *state*[5] between a client (a web browser) and a web server. Netscape browsers (version 2.0 or higher) maintain a file called "cookies.txt" on the client side. This file can be read from and written to by servers on remote web sites. For obvious security reasons, any remote site can read only that information that it has written in the client's cookies.txt file. This restriction therefore prevents a server on one site from reading any registration information pertaining to the client's accesses to some *other* web site. Each time a user accesses any page in a web site, if there is any information in the cookies file that had been written by the server at that site, then that information is automatically sent to the site (along with the request for the page in the site).

Cookies can be used to identify users and sessions as follows. When a user accesses a web site, information in the cookies file specific to that site gets transmitted to the server at the site. The server can then use this information to identify the user. If no such information exists in the cookies file (this may be the first time that the user is ever visiting the site), the server could generate a (unique) userid and any other information deemed relevant that can then be written into the client's cookies. Similarly the server can write a session ID into the user's cookies and substitute this session ID with a new session ID each time the server decides that the prior session has "expired".

There are some drawbacks to identifying unique users in this manner. First, the client has to maintain such information for *each* web server accessed (that uses the cookies file). For users who browse a large number of sites, this may result in having very large cookie files. Second, since these files reside on the client side and are browser-specific, cookies cannot be used to differentiate between different users who may timeshare and use the same computer (same browser) to access the web. This is often the case for PC or Mac users who may use different computers to

[5] Hypertext transfer protocol (HTTP) defines the protocol for communication between a browser and a web server. This protocol is "stateless" in that the server disconnects itself from the client after a single request is processed. If the same client accesses two different pages on a web site, the statelessness of the protocol results in the server being unable to infer that the successive requests came from the same client.

access the web at different times. Third, not all web browsers support the cookies feature. These are currently a feature that Netscape Communications Corp. defines and supports.

3.3.3 The Use of Heuristics in Session Identification

Though we had previously discussed how tokens and cookies could be used to identify user sessions at a site, a fundamental problem with identifying a session remains. Often, it may not be clear when a user's "session" ends. If a user accesses several pages at a site and comes back to the site a few hours later to the same site, is the user continuing the previous session or has the user started a new session?

A method of identifying the start and the end times of a session would involve having the user to *logout* from the web site when the user has completed a "session". Since most web pages do not require that a user *logout* of a site, web log analysis may involve the use of heuristics to identify the "end" of any session. If the end of every session can be identified, any subsequent access by the same user can be considered as the "start" of a new session.

In the absence of conclusive evidence regarding the end of a session, heuristics need to be used. A heuristic used in commercial web log analysis tools is the use of "thirty minutes" as the maximum allowable "idle period" within any single session. This method deems the last access *prior* to a thirty minute idle period as the end of the previous session.

Though methods described in this section can be used to identify client sessions and unique users, it must be emphasized that these do not reveal the *identity* of the user because the sessions and users, defined by these methods, are essentially dummy identifiers (ID numbers) created by web servers. The user ID thus created can be used to identify an anonymous web user but contains no information about user's name or any other personal information.

3.4 Commercial Solutions for Web Usage Analysis

A primary source of revenue for web sites is advertising revenue. Advertisers pay fees to place advertisements on web sites. This results in a demand for two distinct kinds of services provided by several commercial organizations:

(a) *Software that can measure web traffic at a site*. There are several commercial software tools that can be used by web servers to measure traffic at the site. Some of the popular packages include analysis tools such as I/Count from Internet Profiles Corporation, NetCount from NetCount LLC, Interse Market Focus, from Interse Corporation, SiteTrack from Group Cortex Inc., Web Reporter from Open Market Inc., WebTrac from Logical Design Solutions, Web Tracker from Cambridge Quality Management Inc. This list is not intended to be a comprehensive list of commercial vendors, but serves primarily to illustrate the proliferation of firms that develop products for web traffic measurement and usage analysis. All of these packages analyze web logfiles using proprietary technology and methods (including extensive use of tokens and cookies) to provide an understanding of web traffic at a site.

(b) *Independent auditing of web usage data.* The analysis and statistics compiled by packages that measure web traffic at a site are crucial for advertisers for deciding on whether or not to advertise at a site during media planning. However not all of these packages agree on a standard set of measurement units that would assist in a fair comparison of traffic across different web sites. Further, there may not be a shared definition of the measurement vocabulary across different packages. For example, "visits", "sessions" and "hits" could be interpreted differently by different packages. In order to address these concerns and arrive at a common understanding of the measurement vocabulary (or "measurement units") and to develop objective guidelines for the measurement of these units, CASIE (a joint project of the Association of National Advertisers, Inc., and the American Association of Advertising Agencies), with the support of Advertising Research Foundation (ARF), has created the Guiding Principles of Interactive Media Audience Measurement (ARF, 1996). However, there is no unanimity yet on the measurement definitions and standards in the industry. Thus, some firms such as Audit Bureau of Circulations (ABC). and I/PRO (through their I/Audit tool) offer independent third-party auditing of web usage statistics at web sites. These and other firms above are also involved in the development of measurement standards.

The reader will note that the methods described in this section can be used to identify client sessions and unique users, but that they specifically *do not reveal the identity* of the user. The session IDs and user IDs, as defined by tokens and cookies and in access_logfiles, are essentially dummy identifiers (ID numbers) created by Web servers. While the user ID can be used to uniquely identify an anonymous Web user, it contains no information about the user's name or any other personal information.

The limitations of Web logfile information may be overcome by marketers who choose to elicit this information directly from their users. These limitations have also generated opportunities for third-party companies to act as information brokers and provide marketers with more detailed visitor information and these approaches are discussed in more detail in the next section.

4. Meeting Information Needs

In Section 3, we discussed the Web usage tracking information that can be gathered from the common logfiles augmented by several technical methods. In this section we examine which of the 5 W's classification of the WWW consumer segmentation information needs (outlined in Section 2) may be satisfied using this data in Web logfiles. In addition to the logfiles, marketers may be able to collect relevant data from other sources such as consumer surveys, data from commercial vendors or brokers, and secondary data from publicly available sources. Some of these data are "easily" available to the marketer (e.g. data from Web logfiles or publicly available secondary sources), while others require cooperation on the part of the consumer.

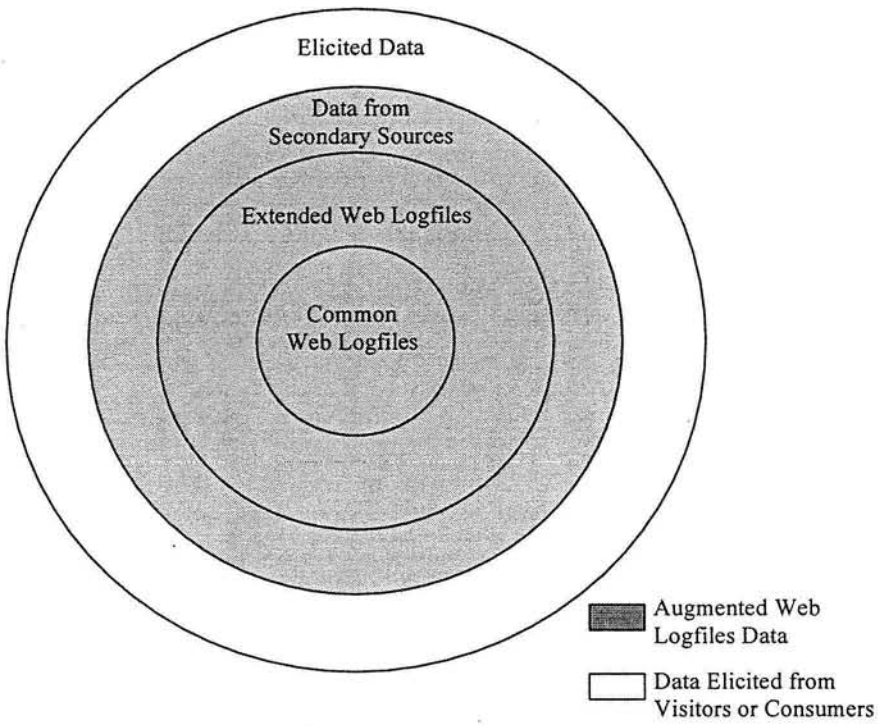
In this section, we examine how the easily available data can be used independently or in conjunction with other types of data to satisfy market segmentation information needs. In section

4.1 we illustrate the sources of data available to marketers. In Section 4.2 we analyze how the information needs described in Section 2 may be satisfied using these data. Through this analysis we propose a mapping of a marketer's Web site consumer segmentation information needs into two categories: those satisfied by *readily or easily available data* (including Web logfiles data), and those *requiring the collection of additional data from the visitors themselves* (e.g. data collected from individuals or households as in the case of single source data). Thus the latter category consists of information needs that will be more difficult or may even be impossible to meet in some cases.

4.1 Classification of Data Based on Ease of Gathering.

Whether or not the data has to be specifically *elicited* from a Web site's visitor determines how easily a marketer may be able to gather information. The various technical methods described in Section 3 can be used to automatically track data about access to Web sites. In addition, there is much secondary data about organizations that is publicly available at little or no cost. Such data do not require any interactions with customers and are, therefore, typically easy to gather.

Figure 4.1 Levels of Web-Related Data



We assume that all other relevant data which needs the visitor's involvement and cooperation (e.g. demographics, motivations, attitudes, purchase intentions, product/service feedback etc.) require additional effort and are more difficult to gather than the data automatically tracked in Web logfiles or publicly available. For example, a consumer may browse a company's Web site but

may be reluctant to fill out its questionnaire on-line. Or, a visitor might be willing to provide product feedback at a company's Web site but may be less inclined to provide demographic information about him or herself. However, in this article we do not distinguish between potentially different levels of difficulty in gathering consumer-specific data, but treat the entire category as being more difficult to obtain. Thus, we distinguish between two broad categories of data -- *augmented Web logfile data* and the data *elicited from visitors and consumers.*[5] These two broad categories of data are illustrated in Figure 4.1 and are described in detail below.

(a) *Augmented Web Logfile Data:* This class of data contains types of data that can be gathered *easily* and without having to involve customers in the process of gathering this information. We include in this class the following three sources of information:

(i) Information available in common Web logfiles. This information was described in Sections 3.1 and 3.2.

(ii) Information on Web site accesses that can be added to common logfiles using any of the technical methods described in the Section 3.3, including tokens and cookies. We use the term *extended logfiles* to refer to the augmentation of common logfiles with any additional information that can be tracked using currently available technical methods.

(iii) Publicly available information. This includes any information that is available through secondary data sources (e.g. Industry Directories, Company Annual Reports, 10-K and 10-Q filings of companies made publicly available by the Securities and Exchange Commission (SEC) in print or at the Electronic Data Gathering and Retrieval (EDGAR) site, http://www.sec.gov)

(b) *Data Elicited From Visitors or Consumers:* This class of information contains sources of information that are not readily available and require the visitor's or consumer's, involvement. This data may be gathered using both traditional as well as new (Web-specific) methods for gathering information, by the company itself or by marketing research firms or information brokers. The different sources for this information are:

(i) Visitor Registration Data. The information which users provide if the organization that owns the Web site asks them to register for the use of their site. A company may limit itself to asking for the user's name, physical and email address, or it may ask visitors for more detailed information on their demographic and psychographic profiles, their attitudes, preferences and the benefits they are seeking from the product or service, etc.

(ii) Marketing Research Data. Traditional marketing research data (from interviews, focus groups, cross-sectional or panel surveys of individuals or households) may be used in

---

5. Our categorization of information is related to a classification, referred to by practitioners, as "customer-centric" and "site-centric". Our classification of information into "augmented logfile" and "elicited" is primarily motivated by the *ease* of gathering information, rather than *where* the information is gathered.

conjunction with logfile data to build a more complete picture of the visitor. In addition to these traditional methods, the ease of communication with individuals over the Web has resulted in the use of this new medium for collecting marketing research information. The advantages of using the Web as a medium for data collection also include its being fast, inexpensive and having a global reach (see Hoffman and Novak, 1996). Beyond simply asking for information once during registration, marketers have now begun conducting periodic on-line surveys of visitors who come to their site for the purpose of tracking segment level trends (Gupta 1996). Companies may also form on-line consumer panels (e.g. Hotweb) and elicit information from them on a periodic basis, by providing suitable incentives. By building a long term relationship with a group of consumers in the panel, a company can track individual level trends in needs, preferences and behavior.

The potential of this medium has also attracted some marketing research firms to start conducting online focus groups and surveys about topics Web marketers are interested in (some of these firms are PC-Meter LP, Cyber Dialogue, Perception Research Services, Inc., Modem Media). These companies are finding that the willingness of consumers to give demographic, psychographic and other information is generally encouraging if the information is used anonymously (Kevin Mabley, 1996). Some of the reasons for this willingness may be that people perceive the Web to be a personal medium as compared to mail (Hoffman and Novak 1996); having a Web site demonstrates the customer oriented-ness of companies (Maddox and Mehta 1996); the consumers who come to a Web site are self selected and have some affinity/interest in the products or services featured there; and the ability for companies to provide the respondents added value to their Web use experience by providing customized content (some sites which provide this are The Wall Street Journal Interactive, Sportszone, Pathfinder, Kraft Interactive kitchen, Hotwired, etc.), free services or software, contests and giveaways, clubs etc. in return for the information the visitors provide them. Of course, the fact that the medium primarily attracts higher educated or younger people may also be a factor (Hoffman, Kalsbeek and Novak 1996). However, users do not want to share information if they are unclear how it will be used or how sharing it may benefit them (GVU's 6th WWW User Survey, November 1996).

(iii) Data from Web Site Information Brokers (or Web site traffic measurement companies). In addition to the in-house data gathering efforts, organizations with a Web site can also look to companies involved in Web site traffic measurement or third party auditing of site traffic to provide a variety of information services to their clients. With the growth in Web sites and Web advertising, the demand for measurement of Web site traffic and independent audit of this has also increased. As a result, commercial software tools that can be used by Web servers to analyze Web logfiles to determine Web traffic and usage have proliferated. The analysis and statistics compiled by such packages can be crucial for advertisers deciding on whether or not to advertise at a certain site.

However, since there is no unanimity yet on the measurement definitions and standards in the industry, some firms such as Audit Bureau of Circulations (ABC) and I/PRO (through their I/Audit tool) offer independent third-party auditing of Web usage statistics at Web sites. These firms and others are also involved in the development of measurement stan-

dards[6]. Besides measuring and auditing traffic, firms are branching out into visitor profiling (e.g. I/PRO through their I/CODE service). Such firms provide anonymous demographic and psychographic information about individuals who have registered with them and agreed to provide this information. From these information brokers, businesses may get a more complete profile of anonymous visitors to their site in terms of their usage, demographic and psychographic descriptors.

In subsequent sections we refer to the two broad categories described here as *augmented logfile information* and *elicited information.*

Figure 4.2 Satisfying Information Needs With Augmented Logfile Information and Elicited Information using Current Techniques.

	Traditional Marketing Segmentation Needs		**Web-Site Specific Marketing Segmentation Needs**	
Information needs category	**With augmented logfile information only**	**With additional elicited information from visitors**	**With augmented log-file information only**	**With additional elicited information from visitors**
Who	(iv)* company information. (v)* competitor/ customer etc.	(i) individual's demographics (ii) individual's psychographics, (iii) product expertise level	(ii) browser software	(i) email/homepage (iii) product expertise (iv) willingness to give information.
Where	(ii)* company's geographic location.	(i) individual's geographic location.	(i) where Web page access came from (direct /through links/ other).	(ii) whether access was from home/office etc.
What		(i) behavior.	(i)* first/repeat visit (ii) files accessed (iii)* usage pattern (vi) product purchased? (vii) feedback.	(iv) Web expertise level (v) download files? (viii) bookmarked?
When			(i) time of visit, (ii)* time spent, (iii)* regularity of visit.	
Why		(i) benefits		(i) casual/serious, (ii) came due to external ads/promotions, (iii) benefits sought.

Note: The numbers in the table correspond to the specific information needs under each of the 5W categories as outlined in Section 2. Text beside the numbers are brief descriptions of the needs listed in Table 2.2 through Table 2.6 An asterisk beside a number indicates that the need is only partially or conditionally satisfied.

[6] More recently Hoffman and Novak (1996) have proposed a set of interactivity and outcome metrics based upon the idea that the best measures of ad value in the Web medium are based upon the degree to which the visitor interacts with the ad.

4.2 Satisfying Information Needs of Marketers

Figure 4.2 lists the nature of information required to satisfy the specific categories of information needs of marketers outlined in Section 2. The figure summarizes how these information needs are satisfied using current state-of-the-art technology (including cookies and tokens) described in Section 3. The two broad columns of the matrix divide the Web site consumer information needs of marketers into traditional market segmentation needs and Web site specific segmentation needs. Within each of these columns, the information needs are split into two sub-categories: (i) those that can be satisfied using augmented logfile information (shaded area in Figure 4.2), and (ii) those that need (additional) elicited information (white area in Figure 4.2).

The rows of the table divide the information needs into the categories discussed in Section 2. For example, the upper-left cell in the table indicates that sub-categories (i), (ii) and (iii) of traditional market segmentation based information needs pertaining to the "Who" category (as detailed in Section 2) can only be satisfied using elicited information. The asterisks besides some of the categories signifies that this category is only partially or conditionally satisfied and therefore involves some assumptions which will be discussed in the next section.

4.3 Description of the Map.

In this section we explain our mapping of the specific information needs listed in Section 2 into those that can be satisfied using augmented logfile information and those that need elicited information. The analysis here is organized by the rows in the matrix.

(a)Who are they?

(i) *Traditional Market Segmentation Information Needs.*

The first three sub-categories of information needs involve individual demographics, psychographics and expertise-levels, none of which is captured in the augmented logfile information available. These needs involve having to collect this information from the individual consumer directly and thus would have to be satisfied using elicited information. The last two sub-categories however pertain to information about firms accessing the Web site. In many cases the IP number tracked in the access logfile can be mapped onto a company name. The domain name "igw2.merck.com" can be used to infer that someone from Merck accessed the site. From publicly available information on Merck it would be possible to obtain information regarding their sales, nature of business, product lines, number of employees and much more. Since Merck is a publicly listed firm, the disclosure requirements mandated by the Securities and Exchange Commission would guarantee the availability of such information in some detail. Furthermore, many of these filings recently became electronically available and can, therefore, be gathered quite easily.

However, there are situations where it may not be possible to gather this data or where potentially erroneous conclusions may be made. First, there are situations when IP numbers tracked in the access log file cannot be mapped onto a company name. Second, the company may not be a publicly listed firm and thus the information available on their activities may be very limited and non-trivial to obtain. Third, when users access a Web site using an access provider (such as Compuserve), the domain name recorded in the logfile may indicate that the access is from a user subscribing to the Compuserve network. In such cases, it may be erroneous to conclude that someone in the firm "Compuserve" is accessing the Web site. For example, even if Compuserve can be treated as a "non-competitor" to the specific Web-site, it is unclear if the user who accessed the site is a competitor. Similarly it will not be possible to identify the user correctly, when individuals access Web sites using their company's Internet accounts for their own and not their company's needs. In such cases, the logfile will only tell us the domain names of their companies' computers, e.g. as in the above example "igw2.merck.com". Therefore we marked sub-categories (iv) and (v) with asterisks ("partially or conditionally satisfies") in Figure 4.2.

(ii)*Web Site Specific Information Needs.*

The specific browser that a user accesses the site with can be tracked in the agent logfile. Hence Sub-category (ii) can be satisfied with augmented logfile information. However the rest of the sub-categories in this class (willingness to give information and make purchases over the Web, their homepage/email address etc.) involve knowledge of the user or interaction with the user and thus need elicited information to be satisfied.

(b) Where are they?

(i) *Traditional Market Segmentation Information Needs.*
Information pertaining to an individual user's geographic location would have to involve either knowledge of or interaction with the user. However, a firm's geographic location can, in most cases, be obtained from publicly available information. The problems of doing so are the same as discussed previously in Section 4.3 (a) (i).

(ii) *Web Site Specific Information Needs.*

The referrer logfile would indicate the previous Web page from which a user accesses any specific page. If this previous page contains any pointer to the specific page mentioned, it could indicate that the user followed the link (or banner) in the previous page. If the previous page does not contain any pointer to the specific page accessed, it could be used to infer that the user accessed the site "directly" (by typing in a URL or using a bookmark).

However inferring whether the user accessed the site from home or from work for instance cannot be made with equal strength by just examining the augmented logfile information. Even if a user is at home, the user might login to a server at work before accessing the Internet and in such cases it would be impossible to determine if the access was from work or from another place. The user may have to furnish this information herself/himself.

(c) What do they do?

(i) *Traditional Market Segmentation Information Needs.*

All of the traditional market segmentation needs outlined in this class require knowledge of the user's preferences and hence need elicited information.

(ii) *Web Site Specific Information Needs.*

In cases where the IP number could be used to represent a single user or when cookies can be used to identify unique users, it would be possible to infer if the user is a first time visitor or if the user has visited the site earlier (repeat visitor). However, not all browsers support the cookies feature and only some of the IP numbers can be used to identify a unique user (in cases where users access the site through an Internet access provider, for example, IP numbers may be assigned dynamically, and it will be erroneous to associate a specific IP number with a single user). However since many browsers support cookies, we decided to tabulate this category into one that may be satisfied with augmented logfile information only.

Keeping track of what information is accessed by visitors is trivial since the file that was accessed gets recorded directly in the access logfile. For example, if the access logfile indicates that the file "homepage.html" was accessed and if this page lists customer benefits, it could perhaps be inferred that the user accessed information on "Customer Benefits". Similarly the files accessed can be used to infer whether a product was purchased at the site or if the user provided feedback through the Web site. The user's pattern of accesses to the site can be inferred by identifying the user each time (similar to the identifying the regularity of visits described previously) and identifying the session. Sub-categories (ii), (iii), (vi) and (vii) can therefore be satisfied using augmented logfile information. However augmented logfile information is not sufficient to infer whether the user added a page to their bookmarks, whether they printed any page or what their level of Web expertise is.

(d) When do they do it?

(i) *Traditional Market Segmentation Information Needs.*
These were combined with the "What" category (Table 2.1) and the mapping is therefore the same as in Section 4.3 (c)(i) above.

(ii) *Web Site Specific Information Needs.*

The time of day when a user accesses a page is recorded directly in the access logfile. The other two sub-categories listed in this class (amount of time spent at a site, regularity of visits etc.) involve having to identify each entry in the logfile with a unique person. Cookies can be used for

this purpose (the pitfalls are the same as those outlined previously). The amount of time that the user spends at a page can be approximately computed by calculating the difference between the time instants when the user accesses a page in the site and the time instant when the user accesses the next page in the site. This method will therefore not work for the "last" page that the user accesses in any session. Hence knowledge of the "session" is essential to satisfy this need. Once the access can be identified with a specific ID (through cookies), the regularity of visits can be approximately inferred. However this would require the identification of *both* the session and the user (in a unique manner).

(e) Why do they do it?

(i) *Traditional Market Segmentation Information Needs.*
All of the traditional market segmentation needs outlined in this class require knowledge of the user's preferences and hence need elicited information.

(ii) *Web Site Specific Information Needs.*

Satisfying the various needs pertaining to why the user accessed the Web site needs interactions with the user. Hence the satisfaction of all three sub-categories of information needs of this class listed in Figure 4.2 involves having to gather the appropriate elicited information.

4.4 Implications of Privacy Issues on the Web

The mapping presented in Figure 4.2 illustrates which of the information needs outlined in Section 2 can be satisfied *now* (using current Web usage tracking technologies). Satisfying many of the information needs require the ability to infer user IDs and session IDs of Web site visitors. As discussed in Section 3, current technology permits the tracking of user IDs and session IDs partially or conditionally. Future advances in technology could make Web usage tracking easier (requiring less reliance on browser types or proprietary technology) and more conclusive. This would result in "more" augmented logfile information since, currently, user IDs and session IDs can only be partially or conditionally tracked.

It is also possible that future technology may permit the automatic tracking of additional kinds of information than is done currently (e.g. such personal details that a user chooses to make "publicly available", as say in their .profile or .plan files in unix systems, giving certain information about themselves to anyone who "fingers" them). However these advances would have to be viewed in the light of industry standards. As we briefly mentioned in the previous section, not all Web server software and Web browser software can track individual users using cookies for example. Until new standards on tracking access information on the Web emerge, it may not be clear by how much and in what areas future technology may augment the tracking information contained in logfiles. The W^3 Consortium is an informal governing body of the Internet and is

currently working toward establishing standards for an "extended logfile"[7] that can track more user information in Web logfiles.

However, solving the technological challenges is only one aspect of what the future will determine, that will impact the ease of availability of information. Privacy issues on the Internet is an a important and complex question, and technology that keeps track of individual users' Web accesses and personal information would have to address several privacy matters including ethical, social and legal issues. Laudon (1996) argues that there has been a continued erosion of privacy brought about by technological change, institutional forces and increasingly outdated legal foundations of privacy protection and that the cost of invading individual privacy is far lower than the true social cost of invading that privacy. The article proposes the consideration of market-based mechanisms based on individual ownership of personal information and a National Information Market (NIM) in which individuals can receive fair compensation for the use of information about themselves. A Business Week article (December 1996) discusses some privacy concerns with the use of cookies at Web sites, but concludes that "the technology is too useful to abandon just because a few site owners are unscrupulous". The article also discusses the role of "third parties", such as an organization called eTrust (http://www.etrust.org), that can provide "privacy ratings" to various Web sites. In a recent survey of over 15,000 Web users, "privacy" ranked second in importance among the concerns that they had (GVU's 6th WWW User Survey, November1996).

In the future, privacy issues will be resolved in favor of tracking either more or less user-specific information. In order to examine the marketers' ability to satisfy their information needs in the future, in this section we consider two possible future scenarios: one with fewer privacy restrictions and one with more privacy restrictions. Figure 4.3 illustrates the direction of movement of the amount of information needs that would be satisfied in the two possible futures[8].

Figure 4.3 Amount of Information Needs Satisfied in Two Possible Futures.

	Future with Lesser Privacy	**Future with Greater Privacy**
With Augmented logfile Information	Increases	Decreases
With Augmented logfile + Elicited Information.	Same	Same

[7] See http://www.w3.org/pub/WWW/TR/WD-logfile.html for a detailed discussion of these issues.

[8] The terms *increases* and *decreases* are relative to the current state of individual privacy on the Web, as determined by the amount of information that can be tracked by logfiles and technological methods possible now.

It is possible that in a future with fewer privacy constraints, both user IDs and session IDs will unambiguously be tracked. The amount of augmented logfile information that can be tracked would, therefore, increase. This could increase the amount of information needs of marketers that can be satisfied using only augmented logfile information. The number of information needs that would be satisfied using both augmented logfile and elicited information would remain the same (bottom-left quadrant in Figure 4.3). However lesser effort may have to be expended to obtain the necessary elicited information since more information gets automatically tracked.

In a future with greater privacy, it may not be possible to track user IDs. Future technology may also give users the option of preventing some Web sites from being able to track them using technical methods. For example, users may have the option of disabling sites from writing onto their cookies files which would make the identification of users and sessions using cookies less feasible. In a future with greater privacy the amount of augmented logfile information that can be tracked would, therefore, decrease. Hence the number of information needs that can be satisfied using augmented logfile information would also decrease (upper-right quadrant of Figure 4.3). The number of information needs that would be satisfied using both augmented logfile and elicited information would still remain the same, though greater effort may have to be expended to obtain the necessary elicited information since less augmented information would be available.

4.5 Implications of the Mappings

There are some general implications of the mappings in Figure 4.2. First and most importantly, few traditional market segmentation information needs (the first column in the table) can be satisfied using augmented logfile information alone. Furthermore, some of those that can be satisfied, have conditional clauses that were described in Section 4.3. This is not surprising since the scope of the traditional market segmentation information needs is broad, and Web logfiles provide us primarily with information on a visitor's behavior at the Web site. Thus we find that similar to any other data on consumer behavior, logfile data needs to be augmented for any usable analysis (e.g. as scanner purchase data is augmented with demographic and psychographic data from consumer panels, data on the prevailing store variables, etc.). Second, even the medium-specific information needs (Web site specific needs) of marketers cannot be completely answered using only the augmented logfile information. Satisfying quite a few of the information needs still requires gathering additional information to supplement the information that is tracked automatically in the extended logfiles using tokens and cookies.

Though elicited information may be required to satisfy several information needs, the interactive nature of the Web makes it technologically feasible to obtain this information electronically at a Web site, instead of having to conduct surveys or polls through traditional media. Not only can this elicited information be gathered faster, cheaper and more efficiently over the Web, marketers have been finding that if Web site visitors can be shown the advantage of responding, their cooperation can be expected.

5. Conclusions

A Web logfile is an important tool for analyzing consumer behavior on the WWW, and it is crucial to understand its capabilities and how it can help marketers analyze their markets. In this paper, we examined the information needs of marketers and analyzed how they can be satisfied with the current Web technologies. To this extent, we examined the structure of the Web logfiles and specified (in Fig. 4.2) which marketing information needs can be satisfied by examining these logfiles and by obtaining other useful information from the readily available sources such as cookies, tokens as well as publicly available secondary information sources such as Company Annual Reports, Industry Directories etc. (we referred to them as augmented logfile sources).

We concluded, based on Figure 4.2, that although some of the information needs of marketers can be satisfied from the augmented logfile sources, many other needs cannot. Two solutions to this problem suggest themselves:

(1) The capabilities of the Web technologies may be extended so that Web logfiles can capture more information useful for marketers.

(2) Marketers have to acquire additional elicited information (i.e. those that require the involvement of the consumer).

With respect to extending the technology frontier of the Web logfiles, one limiting factor is not technological but rather visitor privacy related. Most of the current technological advances involve finding new ways to identify individual users and individual sessions. However, in many cases such user identification violates the privacy of the users, and, therefore may be unacceptable for the general public. Therefore, standardization committees, such as WWW Consortium, are working toward identifying solutions that are acceptable to both parties, i.e. the visitors and the Web site owners.

The second solution is to acquire information from elicited sources, such as Web registration information, surveys of individuals or panels, focus groups, etc. as described in Section 4.5. Since the technological solution has privacy as a serious limiting factor, it is important for marketers to continue to find creative and practical ways for providing incentives to consumers to cooperate in giving information on themselves.

Thus, the future of consumer analysis using Web usage data depends on the direction of the development of Web logfile standards and on the success of marketers in motivating consumers to provide the elicited information. These are challenging issues, and will require much discussion, exchange of ideas and effort on the part of the community involved. However, financial rewards and opportunities to understand their markets well through an intelligent analysis of their consumers' behavior on the Web are very high, and are certainly worth the effort for marketers.

References

Business Week, (September 23, 1996). "Making Money on the Net," by Rebello, K. and Cortese, A. pp 104-118. New York:McGraw Hill.

Business Week, (December 16, 1996). "Privacy and the Cookie Monster," by Wildstorm, S.H. pp 22. New York:McGraw Hill.

Blattberg, R.C. and Deighton, J. (1991). "Interactive Marketing: Exploiting the Age of Addressability," Sloan Management Review, Fall 1991.

CASIE Guiding Principles of Interactive Media Audience Measurement (1996). "Getting Started on Interactive Media Measurement". New York: Advertising Research Foundation

Churchill, G.A. Jr. and Peter, J.P. (1995). "Marketing: Creating Value for Customers". Illinois: Austen Press.

Georgia Tech Research Corporation. "GVU's 6th WWW User Survey". October 1996. http://www.cc.gatech.edu/gvu/user_surveys/survey-10-1996

Gupta, S. (1996), "Online Research: Matching Decisions and Data," paper presented at the Workshop on Internet Survey Methodology and Web Demographics, MIT, January 30, 1996.

Hoffman, D.L, Kalsbeek, W.L. and Novak, T.P. "Internet Use in the United States: 1995 Baseline Estimates and Preliminary Market Segments". Working Paper, Owen Graduate School of Management, Vanderbilt University.

Hoffman, D.L. and Novak, T.P. "Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations". Journal of Marketing. July 1996, Vol. 60, pp. 50-68.

Internet Marketing, (May 1996) "Rust-belt marketers plot Web strategies: How the Nation's Manufacturers reach Cyberspace" by Joe Mullich. New York: Advertising Age.

Kotler, P. (1997) Marketing Management: Analysis Planning Implementation and Control. New Jersey: Prentice-Hall, 9th edition.

Laudon, K.C. (1996). "Markets and Privacy". Comm. of the ACM. Sep. 96, V. 39, No. 9.

Myerson T., President, Interse Corporation. Speech given at the Advertising Research Foundation Interactive Media Research Summit II, July 23-24,1996, New York.

Sterne, J., (1995) WWW Marketing: Integrating the Internet into Your Marketing Strategy. New York: John Wiley & Sons. Inc.