Chapter VII
Uses, Limitations, and Trends in Web Analytics

Anthony Ferrini*
Acquiremarketing.com

Jakki J. Mohr*
University of Montana, USA

ABSTRACT

As the Web's popularity continues to grow and as new uses of the Web are developed, the importance of measuring the performance of a given Website as accurately as possible also increases. In this chapter, we discuss the various uses of Web analytics (how Web log files are used to measure a Website's performance), as well as the limitations of these analytics. We discuss options for overcoming these limitations, new trends in Web analytics—including the integration of technology and marketing techniques—and challenges posed by new Web 2.0 technologies. After reading this chapter, readers should have a nuanced understanding of the "how-to's" of Web analytics.

INTRODUCTION

Effective Website management requires a way to track not only the traffic (number of visitors) at a particular Website, but also what those visitors are doing at the particular Website. Importantly, effective Website management requires a way to map the behavior of the visitors to the site against the particular objectives and purpose of the site.

Many tools have been devised to help assess Website performance; these tools are known generally as *Web metrics*, or the indicators used to measure Website performance (Napier, et al, 2003; Napier, et al, 2001; Schneider, 2007). Many Web metrics are available from the server (the computer) on which the Website is hosted, or "served up," on the Internet. In particular, the server records data for every time a browser hits a particular Web page, and includes informa-

tion for every action a visitor at that particular Website takes; these data, known as *log files*, include, for example, who is visiting the Web site (the visitor's URL, or Web address), the IP address (numeric identification) of the computer the visitor is browsing from, the date and time of each visit, which pages the visitor viewed, how long the visitor viewed the site, and other types of information (discussed subsequently). *Log file analysis*, also known as Web log analytics or more simply *Web analytics*, is the study of the log files from a particular Website. The purpose of log file analysis is to assess the performance of the Website; software (called log analysis software, such as that available from WebTrends, Web Side Story, or Urchin Web Analytics, cf. Schneider, 2007, p. 380) pulls data from the server log files and presents the information in a variety of useful templates.

Although Web analytics can provide very useful information, it also has several drawbacks. New techniques in Web analytics have been developed to overcome some of these drawbacks. Moreover, as the Internet has evolved with the use of new *Web 2.0* technologies (such as social networking, tagging, blogging, and so forth), the ability to effectively measure the performance of a given Website becomes more complicated.

The purpose of our chapter is three-fold. First, we will discuss the current ways in which log file data are used to evaluate Website performance; in addition, we discuss some of the limitations of, and remedies for, log file analysis. Second, we discuss new techniques in Web analytics that augment traditional log file analysis, providing a more robust picture of Website performance. Third, we discuss trends in Web analytics, highlighting issues related to the complications arising from Web 2.0 technologies. After reading this chapter, readers should have a nuanced understanding of the "how-to's" of Web analytics. Importantly, we note that our chapter does not address search engine positioning and how to evaluate it; nor does our chapter address privacy and trust issues, which are important topics in and of themselves.[1] Moreover, to be maximally useful, Web analytics should be used in conjunction with a robust strategic marketing process (e.g., Mohr, Sengupta, and Slater 2005).

CURRENT USES OF, AND PROBLEMS WITH, WEB ANALYTICS

This section addresses the state-of-the-art with respect to Web analytics, and is organized around the following issues:

- What data is collected in Web analytics?
- How is it obtained?
- Who uses the data?
- For what purposes?
- What are the deficiencies and limitations with Web analytics?
- How can these deficiencies be addressed?

Data Included In, and Uses Of, Web Analytics

Table 1 provides an overview of the data that are collected in Web analytics. As mentioned previously, these data are obtained by the computer server on which the Web page resides; the server records every action each visitor takes on a particular Website.

Web logs contain potentially useful information for anyone working with a Website—from server administrators to designers to marketers—who needs to assess Website usability and effectiveness. Website administrators use the data in log files to monitor the availability of a Website to make sure the site is online, available, and without technical errors that might prevent access. Administrators can also predict and plan for growth in server resources and monitor for unusual and possibly malicious activity. For instance, by monitoring past Web usage logs for

*Table 1. Types of data in Log File Analysis**

Hit	Refers to each element of a Web page downloaded to a viewer's Web browser (such as Internet Explorer, Mozilla, or Netscape); hits do *not* correspond in any direct fashion to the number of pages viewed or number of visitors to a site. For example, if a viewer downloads a Web page with three graphics, the Web log file will show four hits: one for the Web page and one for each of the three graphics.
Unique Visitors	The actual number of viewers to the Website that came from a unique IP address (see IP address below).
New/Return Visitors	The number of first-time visitors to the site compared to returning visitors.
Page views	The number of times a specified Web page has been viewed; shows exactly what content people are (or are not) viewing at a Website. Every time a visitor hits the page refresh button, another page view is logged.
Page views per visitor	The number of page views divided by the number of visitors; measures how many pages viewers look at each time they visit a Website.
IP address	A numeric identifier for a computer. (The format of an IP address is a 32-bit numeric address written as four numbers separated by periods; each number can be zero to 255. For example, 1.160.10.240 could be an IP address.) The IP address can be used to determine a viewer's origin (i.e., by country); it also can be used to determine the particular computer network a Website's visitors are coming from.
Visitor location	The geographic location of the visitor.
Visitor language	The language setting on the visitor's computer.
Referring pages/sites (URLs)	Indicates how visitors get to a Website (i.e., whether they type the URL, or Web address, directly into a Web browser or whether they click through from a link at another site).
Keywords	If the referring URL is a search engine, the keywords (search string) that the visitor used can be determined.
Browser type	The type of browser software a visitor is using (i.e., Netscape, Mozilla, Internet Explorer, etc.)
Operating system version	The specific operating system the site visitor uses.
Screen resolution	The display settings for the visitor's computer.
Java or Flash-enabled	Whether or not the visitor's computer allows Java (a programming language for applications on the Web) and/or Flash (a software tool that allows Web pages to be displayed with animation, or motion).
Connection speed	Whether visitors are accessing the Website from a slower dial-up connection, high-speed broadband, or T1.
Errors	The number of errors recorded by the server, such as a "404-file not found" error; can be used to identify broken links and other problems at the Website.
Visit duration	Average time spent on the site (length the visitor stays on the site before leaving). Sites that retain visitors longer are referred to as "sticky" sites.
Visitor paths/navigation	How visitors navigate the Website, by specific pages, most common entry pages (the first page accessed by a visitor at a Website) and exit points (the page from which a visitor exits a Website), etc. For example, if a large number of visitors leave the site after looking at a particular page, the analyst might infer that they either found the information they needed, or alternatively, there might be a problem with that page (is it the page where shipping and handling fees are posted, which maybe are large enough to turn visitors away?).
Bounce rate	The percentage of visitors who leave the site after the first page; calculated by the number of visitors who visit only a single page divided by the number of total visits. The bounce rate is sometimes used as another indicator of "stickiness."

** Napier, Judd, Rivers, and Adams (2003); see also www.webopedia.com*

visitor activity, a site administrator can predict future activity during holidays and other spikes in usage and plan to add more servers and bandwidth to accommodate the expected traffic. In order to watch for potential attacks on a Website, administrators can also monitor Web usage logs for abnormal activity on the Website such as repetitive login attempts, unusually large numbers of requests from a single IP address, and so forth.

Website designers use log files to assess the user experience and site usability. Understanding the user environment provides Web designers with the information they need to create a successful design. While ensuring a positive user experience on a Website requires more than merely good design, log files do provide readily-available information to assist with the initial design as well as continuous improvement of the Website. Web designers can find useful information about the type of operating system (e.g., Windows XP or Linux), screen settings (e.g., screen resolution), and the type of browser (e.g., Internet Explorer or Mozilla) used to access the site. This information allows designers to create Web pages which display well for the majority of users. For instance, many major Website destinations which have a wide variety of users, like Web portals such as Yahoo or MSN, can identify the computer environment for these many visitors, and design Web pages which cater to the most common environment.

Moreover, log files can show how a viewer navigates through the various pages of a given Website, or the *click trail*, also known as *clickstream data*. Clickstream data can show, say, what goods a customer looked at on an e-commerce site, whether the customer purchased those goods, what goods a customer looked at but did not purchase, what ads generate many click-throughs but result in few purchases, and so forth (Inmon, 2001). Because the details in log files give clues as to which Website features are successful, and which are not, they assist Website designers in the process of continuous improvement by adding new features, improving upon current features, or deleting unused features. Then, by monitoring the Web logs for user reaction (increased or decreased usage of the Website's features), and making adjustments based on those reactions, the Website designer can improve the overall experience for Web site visitors on a continuous basis.

Another useful piece of information to provide input on Website design comes from analyzing the actual searches that visitors perform on the site itself. If the Website has a search form on its site (e.g., possibly it has downloaded a Google search bar for its own site visitors to use), the analyst can examine the keywords that visitors searched. This provides clues about the visitor's interests at the site, and, if enough visitors are looking for a particular piece of information, the site designer may want to add it or feature it more prominently.

Finally, Web logs are also used for marketing purposes to understand the effectiveness of various on- and off-line marketing efforts. By analyzing the Web logs, marketers can determine which marketing efforts are the most effective. Marketers can track the effectiveness of online advertising, such as banner ads and other links, through the use of the referrer logs ("referring URLs"). Examination of the referring URLs indicates how visitors got to the Website, showing, say, whether they typed the URL (Web address) directly into their Web browser or whether they clicked through from a link at another site.

In addition, marketers can assess the effectiveness of search engine listings by analyzing which search engines visitors came from and which search queries (keywords typed into the search engine) they used. Oftentimes, the best keywords to use (both for search engine positioning and paid search) are not always obvious. For example, a popcorn chain in New Jersey had been using keywords like "gourmet popcorn" and "popcorn tins." But, when it started using Web analytics, the company learned that more people were searching by "chocolate popcorn" and "caramel popcorn", so it boosted the use of those phrases, both in the site content as well as in its marketing efforts (Spors, 2007). Moreover, it found that most visitors were typing "kettle corn" as two words rather than the one word that the site was using, so it added a two-word version in its strategies as well.

Web logs can also be used to track the amount of activity from offline advertising, such as magazine

and other print ads, by utilizing a unique URL in each offline ad that is run. Unlike online advertising which shows results in log information about the referring Website, offline advertising requires a way to track whether or not the ad generated a response in the viewer. One way to do this is to use the ad to drive traffic to a particular Website. So, many advertisers place a unique URL in each offline ad that they run; each unique URL directs viewers who saw the ad to a different Web address than the Website's regular URL. Web marketers can create a unique URL by buying a completely new domain name (Web address) or by using a subdomain, such as subdomain.domain.com, or by creating unique pages within the current site structure, such as www.domain.com/unique. Any visitor traffic that enters the Website via the unique URL is assumed to have been driven there by the offline ad – the only means by which a visitor could have discovered the specific URL. So, by tracking the number of visitors to each unique URL, the advertiser can evaluate the effectiveness of different offline ads.

Limitations of, and Remedies for, Log File Data

Despite the wealth of useful information available in log files, the data also suffer from limitations, creating challenges for the people using them. The limitations of Web log files generally arise because certain types of visitor data are not logged, such as information about the person visiting the site rather than just the computer visiting the site, and some of the data that are logged may be incomplete, such as visit duration as discussed below. As a result, conclusions based on this data may lead to unsound business decisions.

For example, visit duration is a commonly-reported statistic in Web log reports. However, Web logs do not provide an accurate way to determine visit duration. Visit duration is calculated based on the time spent between the first page request and the last page request. If the next page request never occurs, duration can't be calculated and will be under-reported. Web logs also can't account for the user who views a page, leaves the computer for twenty minutes, and comes back and clicks to the next page. In this situation, the visit duration would be highly inflated.

Another source of inaccuracy is in visitor count data. As discussed in the previous section (Table 1), most Web log reports give two possible ways to count visitors – hits and unique visits. The very definition of hits is a source of unreliability. By definition, each time a Web page is loaded, each element of the Web page (i.e., different graphics on the same page) is counted as a separate "hit." Therefore, even with one page view, multiple hits are recorded as a function of the number of different elements on a given Web page. The net result is that hits are highly inflated numbers.

Visit counts are also inaccurate because most Web analytics programs define a visit as a sequence of page requests from a unique visitor within a certain amount of time, usually 30 minutes. Counting visits in this manner is inaccurate because it relies on an arbitrary 30-minute timeframe to define a visit. Any visit longer than 30 minutes is counted as another visit. So, if a Website provides extensive information, or if a visitor is researching information on a Website for more than 30 minutes, visit counts will be inflated.

Another source of inaccuracy arises from the way in which unique visitors are measured. Web log reports measure unique visitors based on the IP address, or network address, recorded in the log file. However, as discussed in Table 2, due to the nature of different Internet technologies, IP addresses do not always correspond to an individual visitor in a one-to-one relationship. In other words, there is no accurate way to identify each individual visitor. Depending upon the particular situation, this causes the count of unique visitors to be either over- or under-reported. The main reason for this problem is that several Internet technologies make it difficult to identify individual

users (or unique visitors). Table 2 describes these various Internet technologies and their impact on Web analytics.

When it comes to Web logs, decision makers must understand these potential inaccuracies caused by different technologies. Without the ability to accurately identify individual users, there isn't an accurate way to determine the exact number of unique visitors to a Website. As a result, many other items within a normal Web log report also provide inaccurate information, leading to erroneous conclusions about Website activity. For example, Web log reporting software often generates secondary reports based on the original log data. If the original log data, such as hits and unique visitors, are inflated or deflated, the secondary reports will also be inaccurate – leading to unsound business decisions. Say the secondary reports calculate the return on investment for marketing expenditure (the ratio of money gained or lost on an investment relative to the total amount of money invested). If the return on a specific marketing expenditure is computed as a function of the number of visitors the campaign attracted, and if this calculation incorporates an inaccurate visitor count, the conclusion regarding the effectiveness of the campaign will also be inaccurate. As a result, decision makers will base their decisions on misleading information.

In particular, the under-reporting of visitors is a serious issue for online advertising. If the ad is cached, nobody knows that the ad was delivered. As a result, the organization delivering the ad doesn't get paid. "*Cache busting*" is a popular

Table 2. Internet technologies and complications for Web Analytics

Proxy Servers	A proxy server is a network server which acts as an intermediary between the user's computer and the actual server on which the Website resides; they are used to improve service for groups of users. First, it saves the results of all requests for a particular Web page for a certain amount of time. Then, it intercepts all requests to the real server to see if it can fulfill the request itself. Say user X requests a certain Web page (called Page 1); sometime later, user Y requests the same page. Instead of forwarding the request to the Web server where Page 1 resides, which can be a time-consuming operation, the proxy server simply returns the Page 1 that it already fetched for user X. Since the proxy server is often on the same network as the user, this is a much faster operation. If the proxy server cannot serve a stored page, then it forwards the request to the real server. *Importantly, pages served by the proxy server are not logged in the log files, resulting in inaccuracies in counting site traffic.* Major online services (such as America Online, MSN and Yahoo) and other large organizations employ an array of proxy servers in which all user requests are made through a single IP address. This situation causes Web log files to significantly under-report unique visitor traffic. On the other hand, sometimes home users with an Internet Service Provider get assigned a new IP address each time they connect to the Internet. This causes the opposite effect of inflating the number of unique visits in the Web logs.
Firewalls	A proxy server can also function as a firewall in an organization, acting as an intermediary device, but for the purpose of security rather than efficiency. Firewalls are used by organizations to protect internal users from outside threats on the Internet, or to prevent employees from accessing a specific set of Websites. Firewalls hide the actual IP address for specific user computers and instead present a single generic IP address to the Internet for all its users. *Hence, this contributes to under-reporting unique visitor traffic in Web analytics.*
Caching	Although there are many nuances to it (such as "browser caching" and "server caching"), in general caching refers to the technique in which most Web browser software keeps a copy of each Web page, called a cache, in its memory. So, rather than requesting the same page again from the server (for example, if the user clicks the "back" button), the browser on her computer will display a copy of the page rather than make another new request to the server. Many Internet Service Providers and large organizations cache Web pages in an effort to serve content more quickly and reduce bandwidth usage. As with the use of proxy servers, caching poses a problem because *Web log files don't report these cached page views. As a result, once again, Web log files can significantly under-report the actual visitor count.*

term that refers to technologies that solve this problem. These technologies, such as "page tagging," are discussed next.

Correcting Deficiencies in Log File Data

Some remedies exist for the visitor count inaccuracies commonly found in Web analytics: cookies and page tagging.

Cookies are small bits of data that a Website leaves on a visitor's hard drive after that visitor has hit a Website. Then, each time the user's Web browser requests a new Web page from the server, the cookie on the user's hard drive can be read by the server. These cookie data can be used in several ways. First—even if multiple viewers access the same Web site through the same proxy server, for example—each viewer has a unique cookie; therefore, a unique session is recorded and a more accurate visitor count can be obtained. Cookies also make it possible to track users across multiple sessions (i.e., when they return to the site subsequently); this allows a computation of new versus returning visitors. Finally, third-party cookies – often set by advertising companies such as DoubleClick -- allow the Website to assess what other sites the visitor has visited; this enables personalization of the Website in terms of the content that is displayed.

Note, however, that cookies are *not* included in normal log files. Therefore, only a Web analytics solution which supports cookie tracking can utilize the benefits. (Alternatively, Web log files generally utilize a combination of the specific computer's numeric IP address and user agent—browser, search engine spider, or mobile phone—to identify a unique user, with the assumption that the two combined are a close estimation of a unique user.)

Due to concerns about privacy (cookies show which Websites a person has previously visited), many users dislike the idea of cookies being saved to their computer. As a result, many computer users have become savvy in removing cookies, deleting them from their hard drives on a regular basis. Many users even disable the cookie feature in their browser's security options.

As users become more sophisticated, the technologies to make it harder for users both to delete cookies and to surf anonymously become more sophisticated as well, and the cookie arena is no exception. One software program commonly used on the Internet, Macromedia Flash (which allows animation, or motion on the Webpage) offers an alternative to the traditional browser cookie that is harder for users to delete. Any computer user who has Flash software installed with their normal Web browser will have *Flash cookies* on their hard drive. These cookies are different (and separate) from the normal browser cookies. As a result, when users clear their browser cache to delete any stored cookies, the Flash cookies are not cleared out. Therefore, Flash cookies present a new opportunity for tracking unique visitors—although in the future users might also learn how to properly remove Flash cookies.

Another method for collecting information that overcomes some of the limitations in measuring Web site activity is called *page tagging* (www.BruceClay.com). This technique has its origins in hit counters, a small image at the bottom of the Web page which looks and functions much like a car odometer; the hit counter increases by one count with each additional page view. Hit counters originated with many personal and small business Websites as a simple way to track how many people were visiting the site. As hit counters evolved, Website developers and marketers learned that they could identify additional information beyond the basic number of page views on the counter. Page tagging, which uses the same basic principle as hit counters, is a more robust system that relies on embedding a small piece of Javascript software code on the Web page itself. Then, when the computer user

visits the Web page, the Java code is activated by the computer user's browser software. Referred to as "client-side technology" (because the tagging occurs on the user's computer when he loads the Web page) — as opposed to a server-side technology in which the log file records activity generated at the server — page tagging offers a significant advantage with respect to the "caching" problem found with server log files. Log files cannot track visitor activity from cached pages because the Web server never acknowledges the request. However, since page tagging is located on the Web page itself rather than on the server, each time the page is viewed, it is "tagged." Therefore, under-reporting of unique visitors is less of a problem with tagging than with Web log files. While server logs cannot keep track of requests for a cached page, a "tagged" page will still acknowledge and record a visit. Moreover, rather than recording a visit in a Web log file which is harder to access, page tagging records visitor information in a database, offering increased flexibility to access the information more quickly and with more options to further manipulate the data. Because of its increased flexibility (compared to traditional Web analytics based on server log files), most of the innovation in Web analytics is coming from page tagging. This method easily adapts to the rapidly changing Web environment and allows new ways to capture, manipulate, and display visitor information, as discussed in the next section.

Cookies and page tagging assist in an important marketing objective: identifying the most valuable customers (typically defined as those that account for a significant volume of purchases or Web-based activities). This objective can be difficult to accomplish when challenges in Web analytics software make it difficult to identify individual visitors. Flash cookies and page tagging are technologies available to deal with this problem.

NEW TECHNIQUES IN WEB ANALYTICS

Two new features of Web analytics software are site overlays and geo-mapping. In addition, other new features of Web analytics software make it easier to link the log analysis to specific online marketing activities and expenditures.

Many of the newer versions of Web analytics software provide a feature called a *site overlay*. As shown in Figure 1, the site overlay is a visual representation of the click activity on a specific page of the Website. The complete Web page is displayed as seen by the user in a browser, with the addition of the percentages of click activity for each link on the Web page. This overlay feature is a useful addition to the Web analytics software of the past. Rather than reviewing a numerical Web log report for the most popular links and paths through a site, the site overlay provides a detailed visual representation of each individual Web page, with all click activity represented. One benefit of a site overlay is that it provides an easy way to quickly identify which features visitors are clicking. Moreover, it gives a more complete picture of the activity on a specific Web page, as compared to traditional Web analytics which is usually limited to a simple list of the most popular click paths. Web developers and marketers alike can utilize a site overlay to analyze a specific Web page, and even each individual link within a Web page. For example, in Figure 1, the site overlay helps to quickly assess which fruits are the most popular and which are receiving little activity. As the Figure shows, site visitors clicked on mango much more frequently than kiwi fruit.

In addition to site overlays, another new technique in Web analytics arises from visual representations of the data. *Geo-mapping*, relying on new mapping technologies being made available by services such as Google Earth and Microsoft Virtual Earth, displays Web analytics with a richer geographic perspective. In the past, most Web analytics reports provided a

Figure 1. Example of site overlay

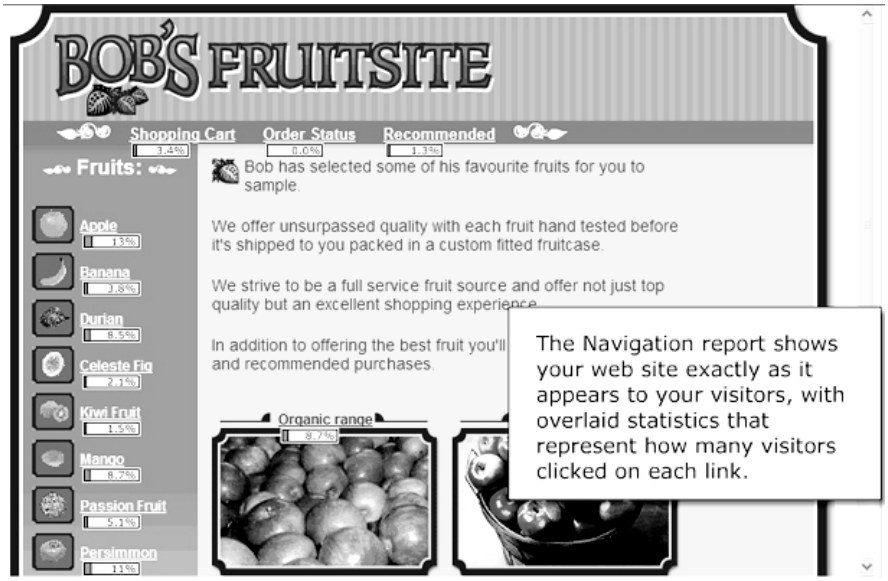


Figure 2. Example of geo-mapping

list of visitor countries (and number of visitors from each country) with little additional detail. Improvements in Web analytics and mapping software provide more detail on visitor locations. As shown in Figure 2, in addition to providing country of origin, geo-mapping provides detail on the specific cities visitors originate from, and creates a visual representation of all the visitors on a world map. This technique can be useful for tracking the penetration of a Website in a particular geographic region, or for tracking the effects of marketing activities in a specific city.

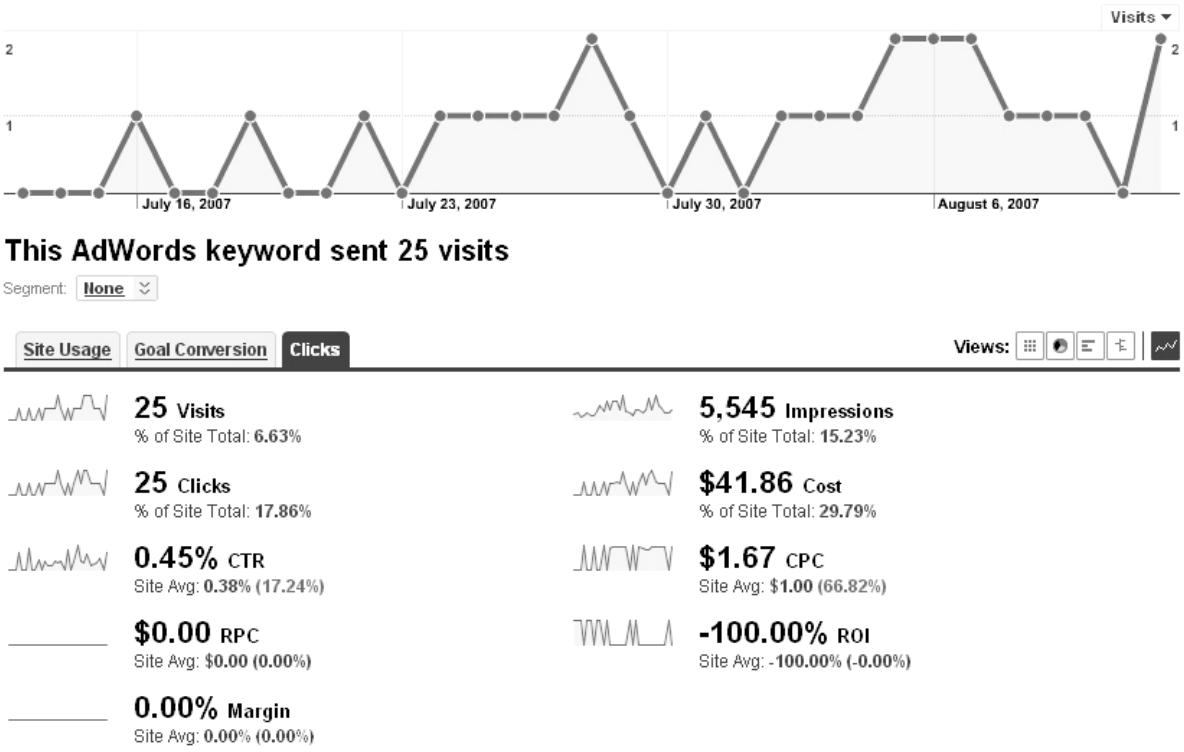
Other new tools in Web analytics provide a stronger link between online technologies and

online marketing, giving marketers more essential information lacking in earlier versions of Web analytics software. For many years, Web analytics programs that delivered only simple measurements such as hits, visits, referrals, and search engine queries were not well linked to an organization's marketing efforts to drive online traffic. As a result, they provided very little insights to help the organization track and understand its online marketing efforts. Trends in Web analytics specifically improve both the method of data collection as well as the analysis of the data, providing significantly more value from a marketing perspective. These newer tools attempt to analyze the entire marketing process, from a user clicking an advertisement through to the actual sale of a product or service. This information helps to identify not merely which online advertising is driving traffic (number of clicks) to the Website and which search terms lead visitors to the site, but which advertising is most effective in actually generating sales (conversion rates) and profitability. This integration of the Web log files with other measures of advertising effectiveness is critical to provide guidance into further advertising spending.

For example, Web analytics software (e.g., Google Analytics) has the capability to perform more insightful, detailed reporting on the effectiveness of common online marketing activities such as search engine listings, pay-per-click advertising, and banner advertising. Marketing metrics to assess effectiveness can include:

- **Cost-per-click:** The total online expenditure divided by the number of click-throughs to the site.
- **Conversion rate:** The percentage of the total number of visitors who make a purchase,

Figure 3. Example of Google Analytics: Cost-Per-Click



signup for a service, or complete another specific action.

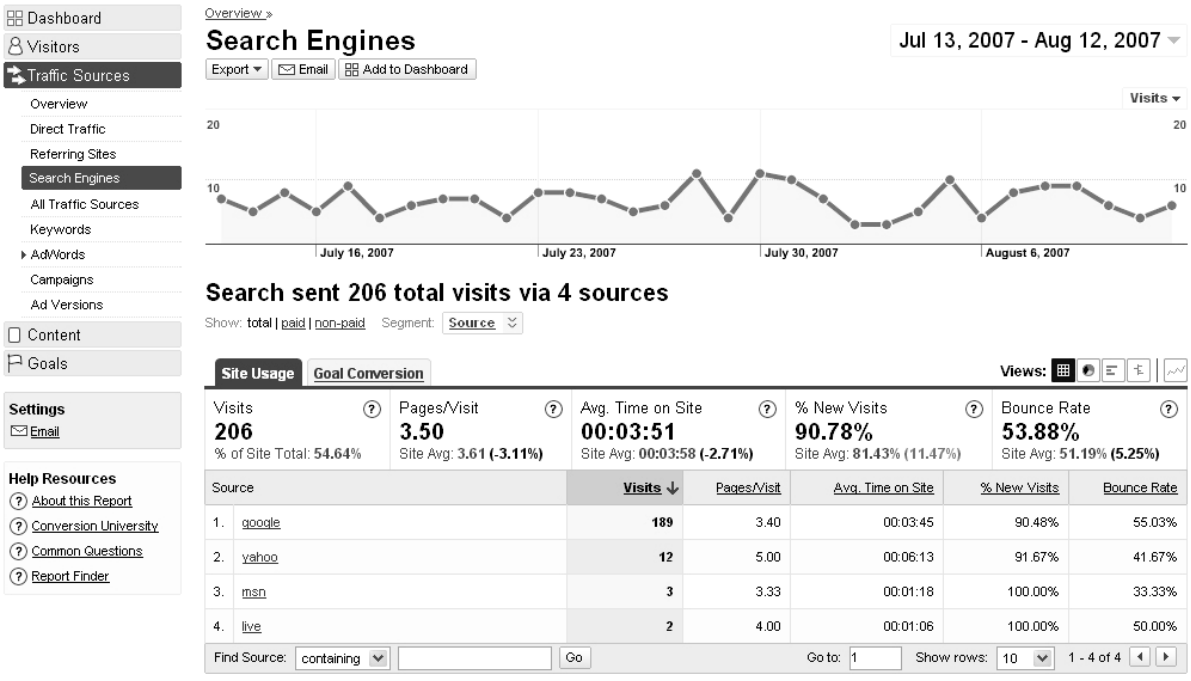
- **Return on marketing investment:** Quantifies the benefits of a marketing expenditure, calculated as the advertising expense divided by the total revenue generated from the advertising expense.
- **Bounce rate:** the number of users that visit only a single page divided by the total number of visits; one indicator of the "stickiness" of a Web page.

An example of a report that links advertising to these metrics is shown in Figure 3. This figure was generated in Google Analytics from a campaign using the Google Adwords program. The Google Adwords program allows marketers to participate in a paid search advertising campaign. The advertiser buys specific keywords at a set price-per-click, and establishes a budget maximum and duration for the campaign. The figure illustrates a report generated for a single keyword; showing the click-through rate, cost-per-click, return-on-investment, and other information which might be helpful in determining a successful marketing campaign.

Figure 4 illustrates another report generated in Google Analytics for traffic to a Website from the top search engines. This report provides the average bounce rate for all traffic in addition to the specific bounce rate from each search engine. A consistently lower bounce rate from a specific search engine might indicate more valuable visitor referrals, in terms of visitor interest. This type of reporting enables a comparison of paid search traffic to, say, organic search engine traffic (which comes from the search engine's own listings using its algorithms), helping an organization to more effectively allocate its resources (Enright, 2006).

Another recent development in linking Web analytics to marketing is "*behavioral targeting*," a

Figure 4. Example of Google Analytics: Bounce Rate



technique that allows "supersmart, supertargeted display ads" based on a person's online behavior that not only do a better job of getting a Web surfer's attention, but also can be tracked with "laserlike precision" (Sloan, 2007). For example, in 2007 Yahoo had about 131 million monthly unique visitors to its sites. By dropping cookies onto every Web browser that looks up one of its sites, Yahoo analyzes this information and combines it with data about what people are doing on its search engine. Its sophisticated model can then be used to predict consumer behavior. In one campaign, Yahoo found that visitors who saw a specific brokerage ad were 160% more likely to search in that category over the next three weeks, typing in keywords like "online brokerages." Most importantly, the visitors who previously saw the ad overwhelmingly clicked on a display for this brokerage when it appeared in Yahoo's paid search results. The benefit is a user profile that goes well beyond a particular search episode (which search string, for example), and integrates the data with a host of other surfer behaviors. Say a person's cookie profile shows that he spent time at Yahoo Auto evaluating cars on fuel efficiency, and then clicked over to Yahoo's Green Center to read about alternative fuels, and then looked at cars on eBay (a Yahoo partner) (Sloan, 2007). Yahoo's behavioral targeting program can predict with 75% certainty which of the 300,000 monthly visitors to Yahoo Auto will actually purchase a car within the next three months. And, the next time this person visits Yahoo Sports, he will see an ad for hybrid cars. Indeed, based on this analysis, Yahoo is finding that ads on sites that seemingly have nothing to do with them (where the content seems irrelevant) can perform very well, because they are based on an elaborate analysis of a user's complete Internet behavior (and not merely a group of search terms.)

Despite these advances in integrating technology and marketing activities, Web logs alone do not answer a host of important business and marketing questions. User surveys and site registration both provide a start for Website owners to reliably identify each unique visitor as well as to collect more in-depth information about the people visiting the site that goes beyond simply how many are visiting. However, important questions still remain. Meaningful data about customer satisfaction is critical, as are insights into the reasons users visit and interact with a Website. Although Web log files provide the number of clicks from a site homepage to another page on the site, they don't provide information on why the users clicked that link. Are the users genuinely interested in the content of that link? Did the user find the information she was looking for at that link? Is the user satisfied with her overall experience with the Website? Moreover, Web logs do not include information about competitors and other market forces that are an important aspect of positioning the Website and its value to prospective site visitors. So, other techniques (beyond Web analytics) must be used to supply insights into other questions and concerns. Standard marketing research methods can be very useful in this regard. Quantitative research techniques such as customer satisfaction surveys can be used as a supplement, as can qualitative research techniques such as usability testing, interviews, and so forth.

WEB 2.0 CONSIDERATIONS

The tools mentioned previously that are used to evaluate Web site performance work well when Internet users are viewing Web pages and seeking out information. However, new uses of the Internet are based on user-generated content and a more user-driven experience; they include, for example, blogging (or posting entries to a Website in the form of a diary or journal, also known as a 'web log'—not to be confused with Web log files), tagging, RSS feeds, wikis, interacting on social networking sites (such as MySpace, FaceBook, or LinkedIn) and sharing rich-media content such as videos (e.g., YouTube). Known collectively as

Web 2.0 (see Table 3), this cluster of collaborative technologies are designed to enhance the user experience on the Internet through enhanced connectivity and communications.

These new technologies pose new complications for Web analytics. First, some Web 2.0 technologies make it difficult to count Website traffic. If a person wants to determine how many readers are reading her blog, it becomes complicated when the blog is shared, say, via an RSS feed. In addition to monitoring traffic at the blog itself, one has to measure how many people access the blog via the RSS feed. The page views of the blog that occurs in GoogleReader, or Bloglines, or LiveJournal, or any place that the blog is syndicated are nearly impossible to track and count.

Second, new technologies such as AJAX and widgets make it difficult to count site traffic. *AJAX* (for **A**synchronous **J**avaScript **a**nd **X**ML) is a programming technique that allows quick,

Table 3. Web 2.0 Technologies

AJAX (*A*synchronous *J*avaScript *a*nd *X*ML)	A programming technique for Websites whose data are regularly refreshed by the user; it allows the Website to exchange small amounts of data with the server behind the scenes (rather than reloading the entire Web page each time the user requests an update), resulting in enhanced interactivity, speed, functionality, and usability.
Blogging (Blogs)	Short for We**b log**, a blog is a Web page that serves as a publicly accessible personal journal for an individual. Typically updated daily, blogs often reflect the personality of the author.
Podcasting (podcasts)	Allows subscribers to subscribe to a set of audio feeds to listen to the content on an iPod (or like device).
RSS (an acronym for *R*eal *S*imple *S*yndication)	Allows people to sign up to have news articles, blog posts, or audio interviews/podcasts from their favorite Websites sent directly to their computers—essentially, the syndication of Web content. A Website that wants to allow other sites to publish some of its content creates an RSS document and registers the document with an RSS publisher. A user that can read RSS-distributed content can then read content from a different site. Syndicated content can include data such as news feeds, events listings, news stories, headlines, project updates, excerpts from discussion forums or even corporate information.
Social networking sites	Websites whose "members" invite contacts and friends from their own personal networks to join the site. New members repeat the process, growing the total number of members and links in the network. Sites then offer features such as automatic address book updates, viewable profiles, the ability to form new links through "introduction services," and other forms of online social connections. MySpace, for example, builds on independent music and party scenes, and Facebook was originally designed to mirror a college community (though it has since expanded its scope to include high school, job-related, and regional networks). The newest social networks on the Internet are becoming more focused on niches such as travel, art, tennis, football (soccer), golf, cars, dog owners, and even cosmetic surgery. Other social networking sites focus on local communities, sharing local business and entertainment reviews, news, event calendars and happenings. Social networks can also be organized around business connections, as in the case of LinkedIn.
Twitter	A Web service that allows users to send "updates" about what they are doing at a particular moment in time via text messages (SMS), instant messaging or email to the Twitter Website; these updates can also be displayed on the user's profile page and can be delivered instantly to other users who have signed up to receive the updates. Also called "micro-blogging" because of the short nature of the frequently-updated posts. Twitter "look-alikes" include country-specific services (e.g. frazr) or sites that combine micro-blogging with other functions such as filesharing (e.g. Pownce).
Widgets	A portable chunk of code that can be installed and executed within any separate HTML-based Web page by an end user without requiring additional compilation; akin to plugins or extensions in desktop applications, these downloadable, interactive icons allow users to perform a task from their desktop without opening a Webpage.
Wiki	A collaborative Website comprised of the collective work of many authors. Similar to a blog in structure and logic, a wiki allows anyone using a browser interface to edit, delete or modify content that has been placed on the Website, including the work of previous authors. In contrast, a blog, typically authored by an individual, does not allow visitors to change the original posted material, only add comments to the original content. The term *wiki* refers to either the Website or the software used to create the site. *Wiki* means "quick" in Hawaiian.

** Web 2.0 is a second generation of Web-based communities and hosted services which facilitate collaboration and sharing user-generated content between and among Website visitors.*

incremental updates for the user without clicking a page refresh or reloading the entire Web page. Google Maps is one Website using AJAX technology. Essentially, in an interactive Web application, the Website exchanges small amounts of data with the server behind the scenes, so that the entire Web page does not have to be reloaded each time the user requests a change. As a result, the Web page's responsiveness (interactivity, speed, and functionality) are increased, and the user has a better browsing experience. However, AJAX technology that allows a page to update itself without reloading creates a problem for counting "page views." When a visitor hits a page using AJAX, only the first page view is recorded; no matter how long that person stays and interacts with the page. (Recall that a page view is typically counted every time the same visitor visits/refreshes the page, cf. Web Analytics Association 2006). Hence, the use of a page view metric for Websites using AJAX can cause problems. For example, after deploying new versions of AJAX-intensive pages, many Websites lost all their traffic in comScore and Nielsen//NetRatings page-view counts (Picard, 2006). In fact, Yahoo's homepage was once listed as the most popular page based on the page-view metric. However, when Yahoo launched its new AJAX-enabled homepage, it lost the number-one ranking to MySpace. As a result, more emphasis is being placed on newer metrics such as visit duration and user interaction. In addition, AJAX does provide some capability for tracking refreshed page views through a tagging and "call back" to the server; however, most experts today find AJAX problematic for the mainstream, commercial analytics software that most companies use.

Widgets are little bits of programming (such as Javascript or Flash) that can be downloaded from one Website and then used or displayed by another. One popular Web widget is from YouTube, whose widget allows users to place videos on their social networking profiles and blogs. Google AdSense also has a popular widget that allows Website owners to display relevant advertisements and share in the ad revenue. The developers of a widget can track how many times their widgets are loaded elsewhere, but again, simple counting may be misleading. For example, if a widget is loaded into a sidebar of a Webpage without anyone paying attention to it, does the simple count convey meaningful data?

More important than the problems in counting site traffic per se are the metrics themselves. In the Web 2.0 environment, traditional metrics used to evaluate Website performance are called into question. Prior to Web 2.0, most visitor activity could be tied to simple page views. However, some argue that, at the extreme, "page views are obsolete" (Williams, 2006) and that "there will come a time when no one who wants to be taken seriously will talk about their Web traffic in terms of 'page views' any more than one would brag about their 'hits' today" (Zedowsky, 2006). In many cases, the sheer number of visitors to a particular site matters less than how engaged the visitors are. "Most bloggers would rather be read by a handful of key influencers who provide thoughtful commentary rather than by legions of regular Joes" (Zedowsky, 2006). Or, the bloggers are interested in the thoughtfulness of a handful of responses to their blogs rather than merely the number who read the blog. As one person stated on Zedowsky's (2006) blog:

I would much rather have 100 focused people reading my site than 100,000 people mindlessly wandering through. With a strong, well-defined niche, I can advertise to it, pull advice and knowledge from it, and learn a lot. This might be [only] a handful of page views. The analogy would be an airline company that brags about how many millions of people it is moving every day. If the quality of the interaction is low and people don't have a reason to come back, bragging about some number you counted up doesn't capture the reality of the situation.

Therefore, Web 2.0 presents a challenge for measuring Web activity because much of the key user activity is more complicated than simply viewing a page. Because user activity on Web 2.0 sites can involve watching a video, listening to a podcast, subscribing to RSS feeds, or creating rather than just viewing content, new metrics must be considered. For example, Web analytics of rich-media content might include, say, metrics such as the number of times a video has been played, the average duration of viewing, and completion rates. Or, in an interactive user environment, the quality of the user base may be more important than the quantity per se. Quality might be captured by visitors who stimulate word-of-mouth, for example.

Unfortunately, the dominant Web analytics companies provide little functionality to track these more nuanced issues posed by Web 2.0 technologies. However, new companies are springing up to address these issues. While there really isn't a comprehensive application to track all of the various Web 2.0 content, an assortment of new companies can provide information on the effectiveness of Web 2.0 sites. For example, TubeMogul.com provides information on various video Websites. FeedBurner.com can provide insight on the popularity of various blogs and analysis of RSS feeds and podcasts as well.

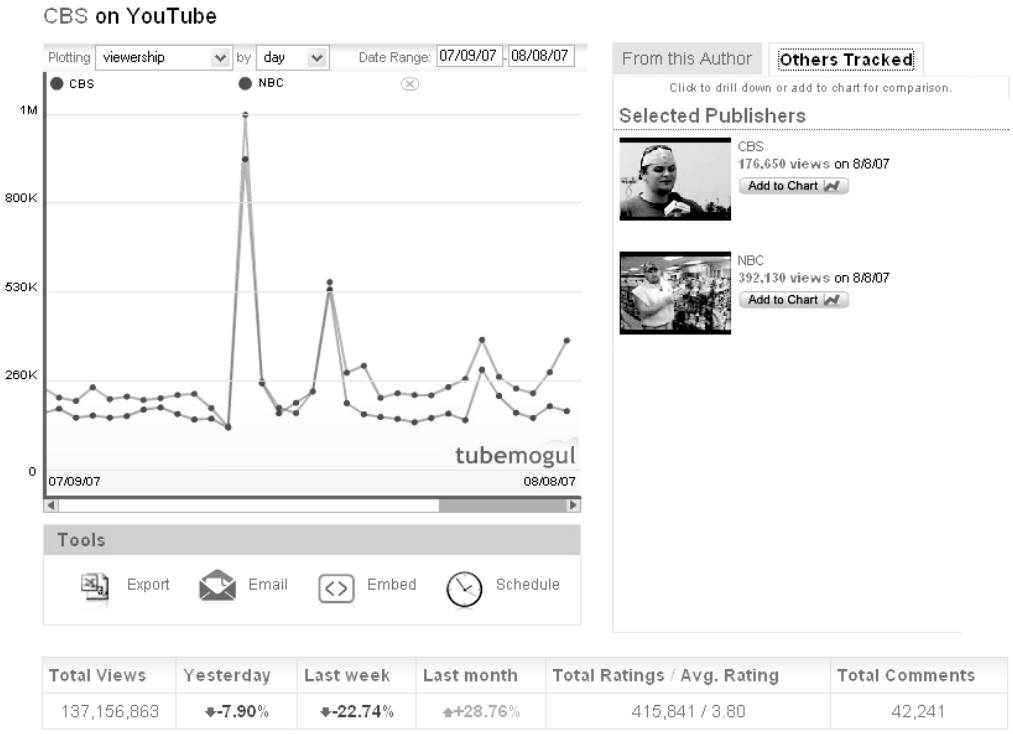
TubeMogul.com is a tool for those that publish, monitor, or advertise within online video. The service allows for viewership-related analytics that aren't provided by conventional Web analytics products. TubeMogul.com overcomes one obstacle with Web 2.0 content, related to the trend in publishing videos to popular video sites such as Metacafe and YouTube. Since this video content is published to an external site, conventional Web analytics does not track this content. TubeMogul can track the viewership of videos scattered across the popular video sites. The service will even aggregate the video comments and ratings from the various sites. Viewership is plotted over time which allows users to monitor spikes and trends. Figure 5 shows a TubeMogul report for viewership in YouTube for CBS versus NBC videos. The data indicate a close relationship between CBS and NBC in peak viewership.

FeedBurner.com, purchased in June 2007 by Google, provides a service for tracking several types of Web 2.0 media including blogs, podcasts, and RSS feeds. This service allows users to determine the number of subscribers, where subscribers are coming from, what they like best, and what they are downloading. FeedBurner, in much the same way as TubeMogul tracks video, overcomes the analytics challenge presented by blogs and other types of feeds by offering a solution to track content that is no longer contained in a single Website, but rather is distributed to other sites and feed readers across the Web. Figure 6 illustrates a FeedBurner report; it shows the most popular feed items, and the number of views and clicks for each item. This report also lists the number of feed subscribers. Feed activity is displayed visually in a graph to trend the activity over time.

As these two examples show, new companies are springing up to handle measurement and monitoring of new Websites based on Web 2.0 technologies. Although complications still exist, the evolving nature of the Internet implies that Web analytics will continue to evolve as well, providing better tools to manage such complications.

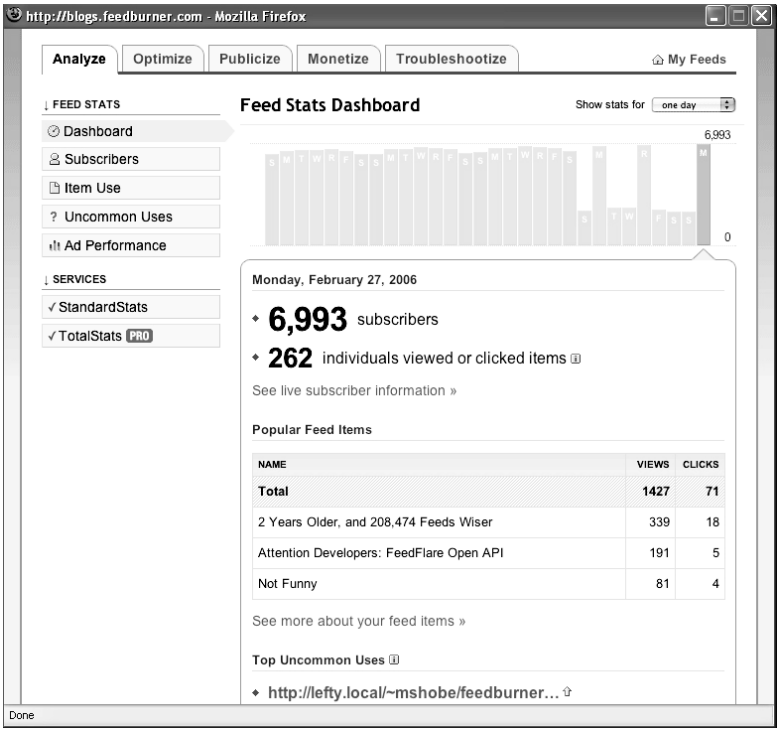
A final consideration in the metrics used for evaluating Web 2.0 sites that we address here is the concept of "the long tail" (Anderson, 2006), a reference to the tail of a demand curve based on the notion that although a relative handful of, say, blogs have many hits, "the long tail" consists of the millions of blogs that have only a handful of hits going into them. Because the "long tail" is a potentially large market, this phenomenon has many implications for current and future business models (Anderson, 2006). For example, products that are in low demand or have low sales volume can collectively make up a market share that rivals or exceeds the relatively few current bestsellers

Figure 5. Example of TubeMogul.com statistics



Total Views	Yesterday	Last week	Last month	Total Ratings / Avg. Rating	Total Comments
137,156,863	-7.90%	-22.74%	+28.76%	415,841 / 3.80	42,241

Figure 6. Example of FeedBurner.com statistics



and blockbusters, if the store or distribution channel is large enough (think Netflix). Indeed, the total volume of low popularity items can exceed the volume of high popularity items—and the distribution and sales channel opportunities created by the Internet often enable businesses to tap into that "long tail" market successfully. The implication of the long tail phenomenon for Web analytics is that current metrics (counts of page views, visitors, etc.)—especially those based on averages—simply don't capture it.

CONCLUSION

This chapter has presented an overview of the traditional metrics used in Web analytics. Web analytics are a collection of tools and techniques that create meaning from the data derived from Web server log files. They can show a plethora of information, including, for example, how Internet users visitors navigated to a particular Website, which pages they visited, where they clicked, what they responded to, what information they supplied, what purchases they made, and which Website they visited next (www.Connectusdirect.com). Web analytics allow companies to discover meaningful patterns and relationships in Web usage and online behavior. Site overlays and geo-mapping are recent developments in Web analytics that provide visual representations of the data.

Integrating the technical perspective of Web log analytics with a business/marketing perspective can highlight not just *what* insights can be gained, but *how they can be used* to guide effective decision making about the specific Website. When combined with other types of information, Web analytics can be used by companies to optimize the conversion of Web traffic to sales and to increase their return on investment from marketing expenditures. At the extreme, companies can learn what motivates customer purchases, what drives customer satisfaction, what builds loyalty, which customers are likely to defect, and even, through behavioral targeting, what a particular customer's future behavior is likely to be.

Although the state-of-the-art in Web analytics is moving in this direction, there are still problems and complications with the existing tools and techniques. Some technologies make it difficult to count and identify unique visitors. When traffic data are inaccurate, subsequent reports based on that data can be very misleading. The use of cookies and page tagging are two techniques that can be used to generate more accurate visitor count data.

Developments in technology tax existing measurement systems. At the extreme, Web 2.0 technologies challenge the very idea of Web performance and measurement. New metrics and new companies are being developed to address these challenges.

The key to successfully use Web analytics to measure Website performance is for decision makers, first, to have a clear understanding of the underlying goal and purpose of the Website itself. Then, one can choose the Web analytics that will provide meaningful answers. Importantly, no single approach or solution provides all the possible information that decision makers need. Web analytics that evolve continuously with the development of new technologies, and that use a combination of solutions to track Website performance, will ensure a rich analysis to ensure effective decision making.

REFERENCES

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. New York, NY: Hyperion.

ConnectUsDirect.com. (2006). Online newsletter on Web analytics. (April).

Dhyani, D., Ng, W. G. & Bhowmick, S. (2002). A survey of web metrics. *ACM Computing Survey, 34*(December), 469-503.

Enright, A. (2006). Real-time analytics boost ROI, accountability. *Marketing News,* (October 1), 20-24.

Immon, B. (2001). Why clickstream data counts. *e-Business Advisor,* (April)

Mohr, J, Sengupta, S. and Slater, S. (2005). *Marketing of high-technology products and innovations*, Upper Saddle River, NJ: Prentice Hall.

Napier, H. A., Judd, P., Rivers, O., & Adams, A. (2001). *Creating a winning E-business* (pp. 364-369). Boston, MA: Thomas Course Technology.

Napier, H. A., Judd, P., Rivers, O., & Adams, A. (2003). *E-business technologies*. (pp. 372-380). Boston, MA: Thomas Course Technology.

Picard, E. (2006). *Ajax Counting Nightmares*. Retrieved 9/14/2007, http://www.clickz.com/showPage.html?page=3610786

Rosenkrans, L. G. (2006). Online banner ads and metrics challenges. *International Journal of Internet Marketing and Advertising, 3*(3), 193-218.

Schneider, G. (2007). *Electronic commerce*. (7th ed., pp. 380). Boston, MA: Thomson Course Technology.

Sloan, P. (2007). The quest for perfect online ad. *Business 2.0,* (February)

Sponder, M. (2006). *ComScore working on web 2.0 metrics*. Retrieved 4/20/2007, from http://www.webmetricsguru.com/2006/12/comscore_working_on_web_20_met.html

Spors, K. (2007, May 13). For the web, choose your words carefully. [Electronic version]. *Wall Street Journal,*

Sweiger, M. (2002). *Clickstream data warehousing*. New York: Wiley.

Web Analytics Association. (2006). Retrieved 09/14/2007, http://www.webanalyticsassociation.org/attachments/committees/5/WAA-Standards-Analytics-Definitions-Big-3-20061206.pdf.

Weischedel., B., Matear, S., & Deans, K. (2005). The use of e-metrics in strategic marketing decisions: A preliminary investigation. *International Journal of Internet Marketing and Advertising, 2*(1/2), 109-125.

Williams, E. (2006). Retrieved 06/12/2007, from www.evhead.com/2006/08/pageviews-are-obsolete.asp

Zawodny, J. (2006). *Hit counter 2.0, or web 2.0 metrics*. Retrieved 02/16/2007, from http://jeremy.zawodny.com/blog/archives/007665.html.

KEY TERMS

Behavioral Targeting: A technique used by online publishers and advertisers to increase the effectiveness of their campaigns. The idea is to observe a user's online behavior anonymously and then serve the most relevant advertisement based on their behavior. Theoretically, this helps advertisers deliver their online advertisement to the users who are most likely to be influenced by them.

Cache Busting: Techniques used to prevent browsers or proxy servers from serving content from their cache, in order to force the browser or proxy server to fetch a fresh copy for each user request. Cache busting is used to provide a more accurate count of the number of requests from users.

Clickstream Data/Clicktrail: The recording of Web pages that a computer user clicks on while Web browsing or using a personal computer.

Cookies (HTTP cookies or Web cookies): Parcels of text left by a Website on the computer user's hard disk drive; these data are then accessed by the Website's computer server each time the user re-visits the Website. Cookies are used to authenticate, track, and maintain specific informa-

tion about users, such as site preferences and the contents of their electronic shopping carts.

Flash Cookies: Similar to "cookies" (above), but coded with Macromedia Flash software; Flash cookies are more difficult to remove than traditional cookies, and as a result, they tend to be more reliable.

Geo-Mapping: A visual representation of the geographical location of Website visitors layered on top of map or satellite imagery.

Log Files or Web Server Logs: A file (or several files) automatically created and maintained by a computer server on which a Website is hosted of the activity on that Website (traffic, hits, etc.). A typical example is a Web server log which maintains a history of page requests.

Log File Analysis: Analyzing log files (Web server logs) to review the aggregate results.

Page Tagging (Web Bug/Beacon): An object that is embedded in a Web page or e-mail and is usually invisible to the user but allows checking that a user has viewed the page or e-mail.

Server Logs: See log files.

Site Overlay: Any type of content that is superimposed over a Web page; for the purpose of Web analytics, the site overlay typically shows click and conversion data superimposed over the links on a Web page.

Web 2.0: A second generation of Web-based communities and hosted services, such as social-networking sites, wikis and blogs, which facilitate collaboration and sharing between users.

Web Analytics: The study of the behavior of Website visitors; the use of data collected from a Website to determine which aspects of the Website work towards the business objectives (for example, which landing pages encourage people to make a purchase).

Web Metrics: A generic term for the many types of measurements that can be made about a Website and its visitors.

ENDNOTES

* Both authors contributed equally to this project.

1 Quantifying site traffic is important for more than just an individual Website. There are many companies that exist to rank Websites based on site traffic (e.g., Alexa.com; comScore.com; HitWise.com; Nielsen NetRatings.com) (cf. Lacy, 2006). The idea is to provide some way to assess audience size in order to allow various Websites the opportunity to "monetize" their traffic by setting ad rates for banners and other forms of online marketing. Moreover, these metrics are sometimes used by investors to determine the valuation of a dot-com start-up. The issues related to measurement and auditing these measures for verified Website traffic statistics are beyond the scope of this article.