The Limitations of Server Log Files for Usability Analysis

by Karl Groves

October 25th, 2007

16 Comments

Introduction

One of the challenges faced most often by those of us in the field of usability is finding good data about user behavior quickly, accurately, and, in most cases, cheaply. In an environment where many stakeholders question the return on investment in usability, some in the industry have developed interesting ideas aimed at gathering user data. One such idea is the analysis of server log files to gather information about user behavior. On the surface, it is easy to understand the gravitation towards server logs: They're supposedly a data source which portrays what people are doing on a site. Server logs supposedly show what people click on, which pages they view, and how they get from page to page.

Unfortunately, practitioners who espouse such methods seem to lack important technical knowledge regarding the nature of the web, the Hypertext Transfer Protocol (HTTP) and the process of caching within networks, proxies, ISPs, and browsers. These technical details greatly limit the types and quality of information that can be retrieved from server logs.

In addition to the technical limitations of server log file analysis, without information regarding exactly what the user expects to find and why he makes the choices he makes, there’s no way for us to know whether he was successful in his quest and whether that quest was satisfying. Ultimately that is the usability information we seek.

Server log files are inappropriate for gathering usability data. They are meant to provide server administrators with data about the behavior of the server, not the behavior of the user. The log file is a flat file containing technical information about requests for files on the server. Log file analysis tools merely assemble them in a conjecture-based format aimed at providing insight into user behavior. In the commentary below, I will explain why the nature of the web, the HTTP Protocol, the browser, and human behavior make it impossible to derive meaningful usability data from server logs.

First, some technical background information is needed.

What is a Server Log File?

Server traffic logs are files generated by the server in order to provide information about requests to the server for data. When a computer connects to a site, the computer, browser, and network will deliver some data to the site’s server itself to create a record that a file was requested. Here’s what an entry into a log file looks like:

```
86.42.132.114 – - [31/Oct/2005:18:15:16 -0500] "GET /styles/style.css HTTP/1.1" 200 5194 "http://www.example.com/links/links.php?cat=css" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.7.12) Gecko/20050915 Firefox/1.0.7"
```

The format above is from an Apache log. Depending on the type of server the site is on, the log entries may look different. Thousands (or even hundreds of thousands) of entries such as the one above are placed into a plain text file, called the server log.

The above log entry includes the following information:

1. IP address of the requesting computer: `86.42.132.114`. This is not the user's IP address, but rather the address of the host machine they've connected to.
2. Date and time of the request: `[31/Oct/2005:18:15:16 -0500]`. That's October 31, 2005 at 6:15:16pm and the time zone is 5 hours behind GMT, which is Eastern Standard Time in the USA (this is because the server is in that time zone, not the user.)
3. The full HTTP request: `"GET /styles/style.css HTTP/1.1"`
    1. Request method: `GET`
    2. Requested file: `/styles/style.css`
    3. HTTP Protocol version: `HTTP/1.1`
4. HTTP Response Code: `200`. This particular code means the request was ok.
5. Response size: `5194` bytes. This is the size of the file that was returned.
6. Referring document: `http://www.example.com/links/links.php?cat=css`. The links.php file is referring to its embedded style sheet.
7. User-Agent String (Browser & Operating system information):`"Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.7.12) Gecko/20050915 Firefox/1.0.7"`. The user’s computer is using Firefox browser on Windows XP and the language is set to English – US.

Of all of the information in the log entry, only the time and date, HTTP Request, and the response information should be regarded as accurate. The IP address, referrer, and user-agent string should be regarded as unreliable, as it can be faked in some way by the user. For example, the user has the option with Netscape 8 to publicly identify the browser as Internet Explorer during the setup process and many other browsers offer this option in their “Options” or “Preferences” menus as well.

Analysis Tools

Many organizations use an analysis program to parse server log files so that they're much easier to understand. Imagine trying to cull anything of value from a huge text file of entries like the one above when hundreds of thousands (or even millions) of entries are present in the log file! Essentially, the analysis tool treats the log file as a flat-file database, processes it, and generates the “statistics” that are discussed throughout the rest of this commentary. In other words, “Web Analytics” software does little more than provide its own interpretation of data contained in the log file which, as stated above, could be at least partially faked.

[Note: I realize that some analytics software gather data by other means than parsing log files and may in fact contain some features meant to overcome one or more of the criticisms I outline throughout this article. I do not discuss such programs, primarily because there is little consistency between them and ultimately they are just as poor at gathering real usability data as analytic tools which parse log files.]

How the Web Works

In order to understand how Web server log files work and why they are considered inappropriate for measuring usability, it is important to get a brief background on how the web itself works.

The HTTP Protocol

HTTP, Hypertext Transfer Protocol, is “A protocol used to request and transmit files, especially webpages and webpage components, over the Internet or other computer network.” [1] HTTP dictates the manner in which computers, servers, and browsers transfer data over the Web. HTTP is known as a “stateless” protocol, meaning that an HTTP connection to a site is not continuous. The steps involved in requesting a Web page are:

1. The user requests a page by following a link or typing an address in the browser’s address bar.
2. The browser requests a page.
3. The server responds by delivering the page, which is displayed in the user's browser window.
4. The connection between the user’s computer and the Web server is severed. At this point, the transaction between the user and the website is considered "done" as far as the server is concerned.

Every request for a new page on the server initiates these steps. In fact, this process occurs for every element requested on that page. Therefore, if there is a page with 10 images and a style sheet, the above routine will repeat 12 times (1 page + 1 style sheet + 10 images = 12 requests).

The reasoning behind all this is important to keep in mind, because in the early days of the Web, computers were slow. Servers were slow and networks were slow as well. This slowness was compensated for in two ways: by opening and closing connections for each request as described above (rather than as one big stream), and by "caching" the data as well.

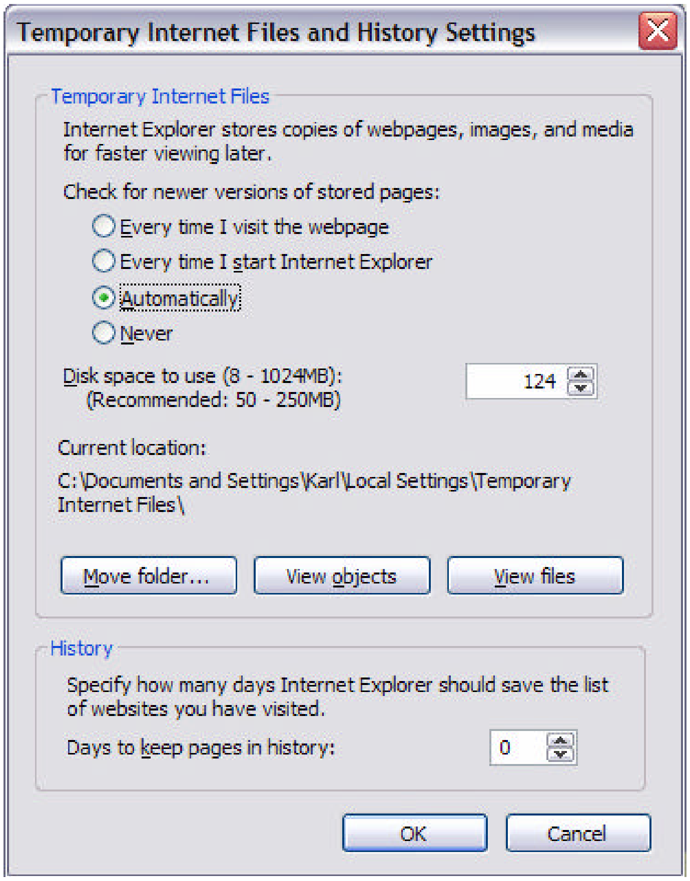
Caching

Caching is defined as “local storage of remote data designed to reduce network transfers and therefore increase speed of download. The cache is a ‘storeroom’ where the data is kept.” [2] To put it more simply: Computers store data in a cache on the user’s computer so that they can get to that data again without downloading it each time it’s needed. For the Web this is important because, as mentioned above, the historical slowness of the Web connections and computers was a severe limitation on the overall usability of the Web.

Caching, even with the speed of today’s Web connections and the power of modern personal computers, enhances performance. For example, in a very simple site where most pages have only two images and a style sheet, each page generates only four requests – again, 1 page + 1 style sheet + 2 images = 4 requests. Without caching, every page viewed on the site would require the server to send four files to the computer when only one is really necessary: the new document. This not only goes for the individual elements embedded into the page, but also the page itself. If a user visits a page, then another, then wants to return to that first page, the computer should not have to download the same page again when it has already been downloaded once. Even with the increasing number of broadband users, if everything had to be downloaded repeatedly for every page viewed, the computer, the server, and the Internet would get bogged down transferring and receiving this enormous volume of redundant data.

Who caches?

Everyone caches. Every browser of every user throughout the world has a cache generated while browsing. Default settings for the cache are set rather high by browser manufacturers in an attempt to increase the performance (and therefore usability) for the user. The screen capture below shows the default cache size for Internet Explorer 7 as 124 Megabytes:



Temporary Internet Files and History Settings

Temporary Internet Files

Internet Explorer stores copies of webpages, images, and media for faster viewing later.

Check for newer versions of stored pages:

○ Every time I visit the webpage
○ Every time I start Internet Explorer
● Automatically
○ Never

Disk space to use (8 - 1024MB): 124
(Recommended: 50 - 250MB)

Current location:
C:\Documents and Settings\Karl\Local Settings\Temporary Internet Files\

History

Specify how many days Internet Explorer should save the list of websites you have visited.

Days to keep pages in history: 0

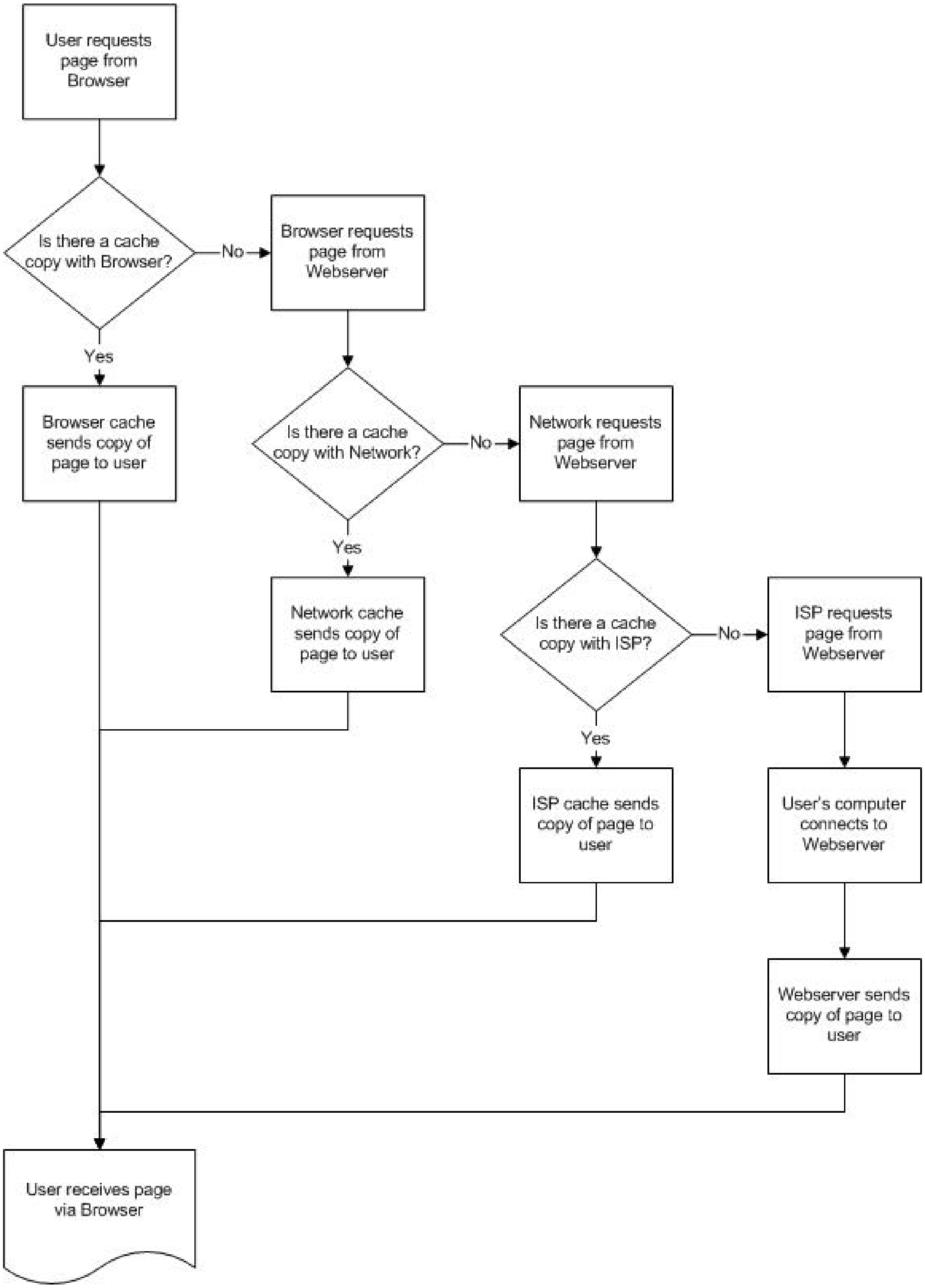
As a browser's cache fills up, more and more pages which the user views frequently (such as on their favorite sites) will be retrieved right from cache, rather than arriving via new requests to the server.

In addition, almost all corporate and institutional networks have caches (most users would experience a cache like this at their job), as do almost all Internet Service Providers (ISP), such as AOL or MSN. The larger the network or ISP, the larger the cache.

A quote from Stephen Turner, developer of Analog Stats, explains it this way: "This means that if I try to look at one of your pages and anyone else from the same ISP has looked at that page, the cache will have saved it, and will give it [the page] out to me without ever telling you [the source of the web page] about it. (This applies whatever my browser settings). So hundreds of people could read your pages, even though you'd only sent it out once."

Why cache?

Caching saves on bandwidth and hardware needs, and provides overall usability for everyone on the web. This means HTTP is really more complicated than the previous description of four steps. It now becomes:


User requests page from Browser
Is there a cache copy with Browser?
No
Browser requests page from Webserver
Yes
Browser cache sends copy of page to user
Is there a cache copy with Network?
No
Network requests page from Webserver
Yes
Network cache sends copy of page to user
Is there a cache copy with ISP?
No
ISP requests page from Webserver
Yes
ISP cache sends copy of page to user
User's computer connects to Webserver
Webserver sends copy of page to user
User receives page via Browser

Caching Wreaks Havoc on Statistics

From the image displayed above, we begin to realize the problems with looking at server logs. There may never be a connection between the user's computer and the site's server in order to fulfill the user's request to see the page.

Caching is definitely a good thing. Its importance for the overall usability of the Web cannot be overstated. But for understanding log file analysis, the purpose of this article, caching makes all site traffic stats completely inaccurate. To further invalidate the stats, the most popular pages of the site are likely to be cached more often, thus creating more problems in getting accurate log data.

Caching's Effect on Margin-of-Error

It's reasonable to take the position that we don't need 100% accurate logs to get reliable data. We just need an acceptable margin of error. If the stats in question are from a small sized sample (read as: a less popular site with little traffic), then our margin of error in these statistics will be very high. To overcome the problem of a large margin of error, a larger sample size (a more popular site) might seem appropriate. Unfortunately, the more popular the site, the more likely it is that the site will be cached by networks and proxies, and the more likely it is that AOL users will be swapping Internet Protocol (IP) Addresses with each other (more on that later). Sites with large numbers of visitors may never achieve an acceptable margin of error because so much caching occurs that the data are not accurate.

Usability Data Cannot Be Found In Log Files

It is helpful to remember that the ONLY things that log files can report are regarding requests to the server. So if the user's network or ISP has a cached copy of your page, then there is no request. If there was no request, then no log entry is created. If no log entry is created, the user's desire for that page is not accounted for in the stats.

Usability Data: Who Is Coming To Your Site?

As demonstrated in the beginning of this article, log entries do not include any demographic data about a user. The log entry cannot tell the user's age, sex, race, education, experience with computers and the Internet, experience with the organization's product or service, or any of the other information usability practitioners typically include when generating a "persona." The closest thing to "who" information is the logged IP address.

In the log entry example provided in the beginning of this article, 86.42.132.114 is an address owned by Eircom IP Networks Group from Dublin Ireland. But even this does not equate to definitive information about the identity of the user. Eircom's own web site says they offer services to businesses and individuals. So, this could be a home surfer, a secretary, or the president of a company – all three of whom could have very different demographic qualities and reasons for coming to the site. The only possible assumption one could make is that the person came from somewhere in Ireland. But even that may not be true. The IP address recorded in the log file is the IP address of the host machine that the user was connected to when they accessed the site. The user could actually be very far away from that host machine or could even be using a program which "anonymizes" him, hiding his location, and making him seem like he's in an entirely different country. For usability purposes, there simply is no way of knowing anything accurate about "who" is visiting the site – not even where he is located.

Usability Data: What Information is Requested?

Log files cannot tell how many actual times a page from the site was viewed because caching causes some "traffic" to not get counted. Even if we pretend that the page counts are accurate, the bigger issue is that "what information they're requesting" provides no data of value for usability.

- Log files don't indicate what information users want.
- Log files don't indicate what information users expected to find.
- Log files don't indicate whether that request fulfilled the user's needs.
- Log files don't indicate whether that information was easy to find.
- Log files don't indicate whether that information was easy to use once users found it.

At most, a measure of page requests can indicate that users thought they could get what they wanted there, but we still don't know exactly what 'it' was they wanted. Without information about what they wanted, there's no way of knowing whether their page view satisfactorily gave the users what they came for.

Furthermore, some page views could be artificially inflated by users going from page-to-page looking for something they cannot find. Page views of some pages could also be inflated as users drill down into a site's architecture as they seek what they really want. No matter whether the request is a success or failure, pages in the site could easily seem to have large requests even if those pages are just on the way or in the way.

Let’s imagine that the user of a newspaper site wants a medicine-related story published June 20, 2004 and that page is located at: Archives -> 2004 -> June -> 20th -> Medicine -> Story [goal]. Further, let's imagine the user's response is to follow the path above until the June 20th edition has been found. In doing so, this increases the page views to five pages to find the story in question.

If a large number of users who peruse the archives of the site use a similar scheme, in the overall traffic measure they have the affect of "artificially" increasing hits to all of the pages at the top levels as user after user branches off of them. Ultimately that data does not provide usability information. The measure of a site's usability is whether the user succeeded in doing that which he set out to do and whether he felt it was an easy thing to do. Page requests have no direct bearing on any true aspects of usability and indicate neither success nor satisfaction.

Usability Data: Can Log Files Indicate How People Navigate on a Site?

Another misunderstanding is found in claims that log files can describe “how people navigate” on a site. As before, the problem with this claim lies in the issue of caching. Even in a best-case scenario the only requests that would be counted would be requests for pages the user had not seen already. Caching means that anytime the user goes back to a page they’ve already seen, that page view is not counted. So, with that understanding, how can log files tell us “how people navigate”?

Tracking a participant's usage of the Sears.com website's tool section, we saw this example played out in real time. The pages visited were as follows:

1. Home
2. Tools
3. Air Compressors
4. Automotive Air Tools
5. Drills
6. Craftsman 1/2 in. Professional One-Touch Drill
7. "Back" to Drills – comes from cache
8. Chicago Pneumatic 3/8 in. Angle Drill
9. "Back" to Drills – comes from cache
10. Craftsman 1/2 in Heavy Duty Reversible Drill
11. "Back" to Drills – comes from cache

12. "Back" to Automotive Air Tools – comes from cache
13. Grinders
14. Craftsman 7 in. Angle Grinder
15. "Back" to Grinders – comes from cache
16. Craftsman 1/4 in Die Grinder kit
17. "Back" to Grinders – comes from cache
18. Craftsman Die Grinder
19. "Back" to Grinders – comes from cache
20. "Back" to Automotive Air Tools – comes from cache
21. Sanders
22. Craftsman Dual Action Sander
23. "Back" to Sanders – comes from cache
24. Chicago Pneumatic Dual Action Sander
25. "Back" to Sanders – comes from cache
26. Craftsman High Speed Rotary Sander

As the page list above demonstrates, as a user browses through the site, the number of actual requests may very well diminish because more and more pages are getting cached as the user navigates through the site. Thus, as they browse and eventually return to pages they've already seen their requests for those pages are not counted in the server's log file. Using log files to track a user's visit as a way to tell how people navigate would result in numerous gaps between what pages the log files indicate were visited and what pages the user actually did visit. So, log files are not a way to identify trends in how visitors navigate.

Moreover, what usability data could a list of pages visited provide? Even if there was no caching anywhere, all we'd have is data about a series of requests that a user made during a visit to the site. To repeat we are left with the following stark realities:

- The user's path does not tell us what the user wanted.
- The user's path does not tell us what information they expected to find.
- The user's path does not tell us whether that request fulfilled their needs.
- The user's path does not tell us whether that information was easy to find.
- The user's path does not delineate distractions that interfered with an initial purpose.
- The user's path does not tell us whether that information was easy to use once it was found.

Usability Data: What Element(s) Did the User Click?

Some sources state that you can determine which element (link, icon, etc.) a visitor clicked on the page to go to the next page. There is simply no component of web sites, web servers, or server log analysis programs which would tell exactly which link/ icon/ button the user clicked to navigate.

Instead, analytics programs "guess" at this information by interpreting the referring pages for a request. For example, if a user follows a link to the site's "News" page, and the referring document was the "Home" page, then the analytics program will locate the "News" link on the "Home" page and say that's where the user clicked. This data is rendered completely unreliable in any instance where there are two links on the page that go to the same destination.

There are those who've proposed methods using DOM Scripting or Ajax to gather this information. Unfortunately, such proposed methods typically involve means which would cause duplicate requests to the server. Such methods – while interesting – are inappropriate for a production environment as these duplicate requests are likely to cause performance problems for the site. These methods would be excellent for a very brief A/B test, but long-term data gathering in such a manner should be avoided.

Usability Data: How Long Did the User View the Page(s)?

Caching again makes it impossible to reliably generate such a statistic. Using the trip to Sears.com as an example again, we can see numerous times when our user visits a new page, and then returns to a page he has already seen, then a new page again. What can happen is that, as far as the logs are concerned, the time the server thinks the user spends viewing page(s) will be artificially increased because it isn’t counting the time he spent re-reading a page he has already seen. The traffic analysis tool then makes an assumption that the gap between new requests (because the other pages were in cache) is the time that was spent on the previous page. This can make it seem like users are spending longer on the pages than they really are, and not actually counting the pages on which they actually spent their time.

If you could tell how long the user "viewed" the page, what conclusion could you draw?

When our user was on the Sears.com web site looking for a part for his air compressor, he realized that he also needed a new pressure gauge because the old one wasn't working. So, he went to Sears.com and clicked on "Parts" on the left. The user then arrived at a new site for the Sears Parts Store. As soon as he got there, he saw that he could enter the actual part number for the gauge itself rather than browsing for it. He then left his computer, went to the basement, and spent approximately the next 5 minutes going through a stack of owner's manuals to get the part number for the gauge. Returning, he entered the part number, went about finding the part on the site and continued with his visit.

With the web server logs as the only data about what our user did, what conclusion could be drawn? As far as the server logs go, he spent 5 minutes at the home page of the site before going to the page listing the part.

- Does that 5 minute page view time mean he had difficulty understanding the page? Or,

- Does that mean he found the page particularly interesting?

Of course, neither is true. He wasn’t even at the computer for those five minutes. But the server logs don’t register, “User walked away to look for something.” Server logs don’t know whether the user went to get a cup of coffee, walk the dog, write something down, print something off, or left the house to go shopping. How long a user spends on a page means nothing without additional information about what the user was doing, why the user was doing it, and whether the experience was successful and satisfying. If someone isn’t with the user, there is no way to know what is contributing to the length of time users spend between new requests.

Usability Data: When Do Visitors Leave Your Site?

Server log analysis tools report the last page the user has requested during their visit as their "exit point." These programs establish a certain amount of time that they assume to be long enough to constitute a single user’s session. Most analysis tools allow the server administrator to set what is called a "Visit Timeout" after which point if there is no more activity from that IP address, it regards the user’s visit as over. Any additional activity from that IP address will be regarded as a new visit and likely to be counted as a "repeat visit" by the analysis tool. Most analysis tools’; timeouts are set, by default, to 20 or 30 minutes. Even if there weren’t any issues with caching, the analysis tools make an assumption that a user’s interaction with the site is over when, in the user’s mind, it might not actually be over. If our user’s trip to find the owner’s manual had taken 31 minutes, Sears.com would have registered him as a repeat visit, but in his mind it was the same event. If the logs can’t reliably tell use when the user left the site, we can't possibly come to any conclusion about why they left the site, and that is really what we really want from that data.

Stats Do Not Represent How Many Visits (or Visitors) the Site Had

Essentially, everything said above means any statistics from log files about number of "visits" are going to be unreliable. If we’re faced with caching issues and visit timeouts, that means we can’t expect to get a reliable statistic of how many visitors have come by or how many visits have happened. In fact, Stephen Turner, developer of Analog Stats refuses to generate visit stats for this exact reason.

Stats Cannot Tell You Where the Visitors Came From or Where They Entered

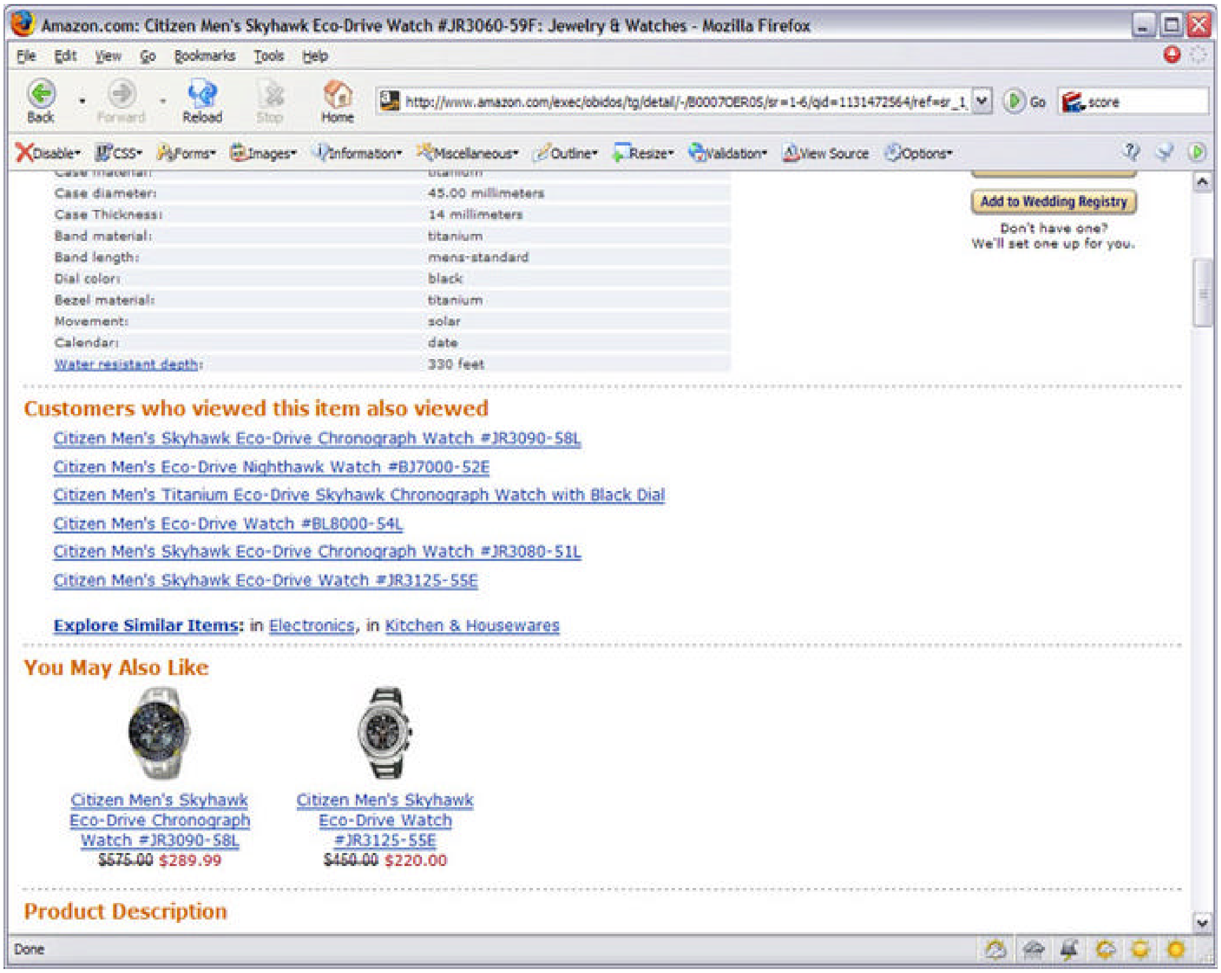
Because the site's pages may or may not be cached, a user may have viewed several pages on the site before actually needing to request a page from the server. The server logs will show that first request as the user's entry point, which may not be the user's actual entry point. This is all the more true for visitors who frequently come from bookmarks or who have the site entered as their "home" page, for these first pages are very likely to be cached in the user's browser. The site's most loyal visitors may very well be skewing the statistics by not generating an accurate count of these popular entry pages. Instead, data might make it look like other pages are more popular than they really are.

Further complicating this matter, referrer data may not even be available. If the site is running under the secure HTTPS protocol (as should be the case for web stores and any page which contains a form to collect information), referrer data will be unavailable, as this is a feature of that secure protocol. Even without the secure connection, many browsers on the market do not pass referrer data. Even in cases where they do, this capability can often be turned off in the browser as a feature intended to protect users' privacy. A user could also be using a "de-referrer" service such as UltiMod to hide their referrer.

Stats Cannot Measure Users' Online Success

There are those who would claim that stats can tell you how "successful" the site's users are. Their claim is that items purchased, files downloaded, and information viewed are concrete indicators of user success. By stating this, they're making an assumption that every user comes to a site with a singular, specific goal in mind – buy something or look for some very specific information – and then leave.

Realistically, nobody uses the Web only to complete a singular task on each site they visit. Most users do not come to a site for only one purpose and leave when they're done. The user's goal may be as well-defined as "order a new gauge for my air compressor" or as open-ended as "look at all the neat tools I could buy if I won the lottery". Frequently, a user will come for one reason and stay for a completely different one. In fact, most organizations (should) hope that users do exactly that. Amazon.com is the undisputed champion of facilitating this type of interaction.


Amazon.com: Citizen Men's Skyhawk Eco-Drive Watch #JR3060-59F: Jewelry & Watches - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back
Forward
Reload
Stop
Home
http://www.amazon.com/exec/obidos/tg/detail/-/B0007OER05/sr=1-6/qid=1131472564/ref=sr_1
Go
score
Disable
CSS
Forms
Images
Information
Miscellaneous
Outline
Resize
Validation
View Source
Options
Add to Wedding Registry
Done

Case material:	titanium
Case diameter:	45.00 millimeters
Case Thickness:	14 millimeters
Band material:	titanium
Band length:	mens-standard
Dial color:	black
Bezel material:	titanium
Movement:	solar
Calendar:	date
Water resistant depth:	330 feet

Don't have one?
We'll set one up for you.

Customers who viewed this item also viewed

Citizen Men's Skyhawk Eco-Drive Chronograph Watch #JR3090-58L
Citizen Men's Eco-Drive Nighthawk Watch #BJ7000-52E
Citizen Men's Titanium Eco-Drive Skyhawk Chronograph Watch with Black Dial
Citizen Men's Eco-Drive Watch #BL8000-54L
Citizen Men's Skyhawk Eco-Drive Chronograph Watch #JR3080-51L
Citizen Men's Skyhawk Eco-Drive Watch #JR3125-55E

Explore Similar Items: in Electronics, in Kitchen & Housewares

You May Also Like

Citizen Men's Skyhawk Eco-Drive Chronograph Watch #JR3090-58L
~~$575.00~~ $289.99

Citizen Men's Skyhawk Eco-Drive Watch #JR3125-55E
~~$450.00~~ $220.00

Product Description

In the screen cap above, you can see that while the user is viewing a specific watch, Amazon is recommending alternate items to view in case the user has determined that the item he’s viewing does not fit his needs. If Amazon assumed that the user only had one goal (look for a specific item and buy it), then Amazon would miss out on a ton of sales, cross-sales, and follow-up sales. What if the user found the specific watch he was after, but decided not to buy it? Amazon loses money. However, by offering alternatives, they support users who don’t have only one specific goal.

Website Success Cannot Be Measured Online Even If Such Stats Were Possible

Different types of users have different goals, many of which can not be measured online. Based on the author’s personal experience as well as surveys conducted by the Pew Internet & American Life Project, there are three different types of users:

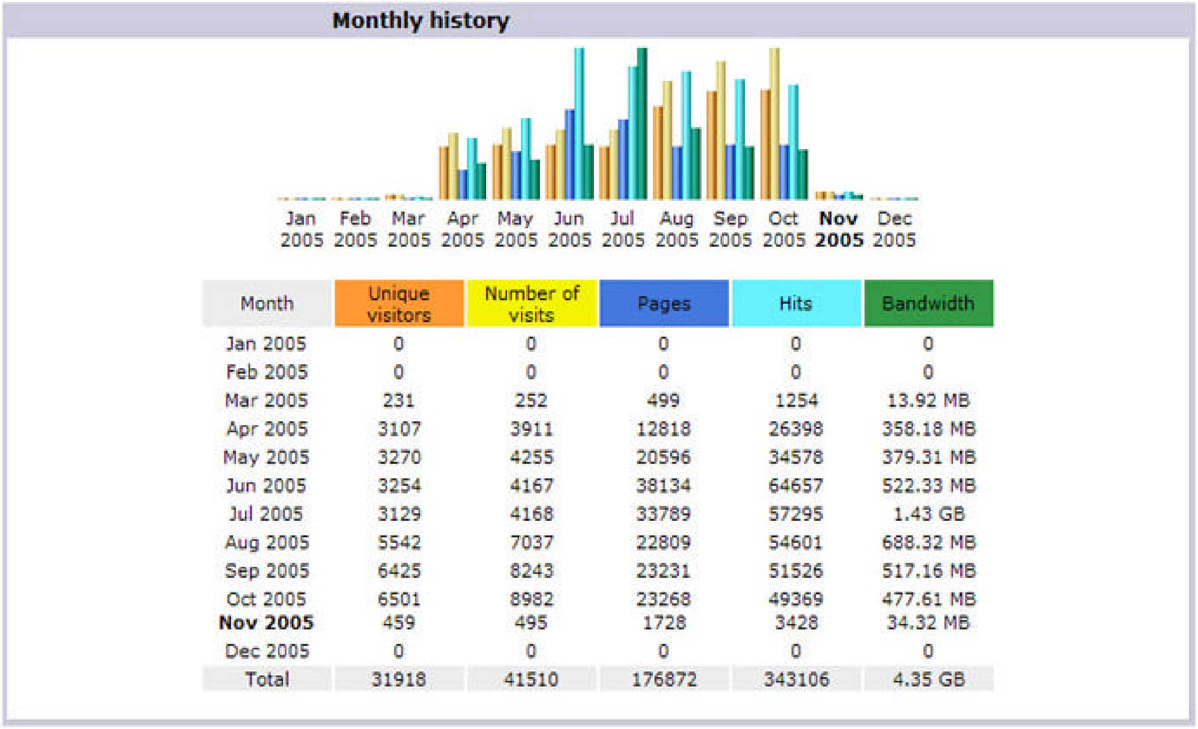
- **Users who prefer to deal with you in person.** They would rather come in to a brick and mortar location and transact their business and/or gather their information face-to-face.
- **Users who prefer to deal with the organization over the phone.** They would rather call than spend all that time traveling to and fro. The telephone is close to them, dialing is easy, the interaction feels comfortable, and they still enjoy the benefit of speaking with someone who can give them personal attention.
- **Users who prefer to deal with the organization online.** The Internet is always there regardless of how early or late it is, and they do not need or want the assistance of another human.

We shouldn't assume, however, that people strictly stay in one of the categories above at all times. Surveys show that people will often switch between these modes based upon a variety of factors such as the type of product, their level of expertise, proximity of the brick and mortar location (if one exists), their level of desire, and their overall goal. One study by Pew indicates that even people who frequently make purchases online are not more likely to do things like manage their investments online than those who don't buy online as frequently. [3] The reason the survey respondents gave for this different approach is easy to understand. They say that handling their entire financial well-being online isn't exactly the same as buying a book, CD, or gift for someone.

Generally, people will choose what type of interaction they want based on what is needed and available for their particular situation and how comfortable they are doing it online. If a user only goes to the Financial Consultant Locator Tool on the Solomon Smith Barney web site to find out where to visit the local office, then subsequently has SSB handle his investments, then the site was successful – but the log files do not track that. Likewise, saying "X amount of people visited the Locator Tool" is certainly no measure of success. Some people, after visiting that page, may decide not to go. Other people might feel the nearest consultant was too far away. Further other people might not be able to use the Locator Tool. Measuring requests to the Financial Consultant Locator Tool – which is simply an imaginary "end-point" – is no measure of success without knowing what the user's criteria for "success" actually is, and without also knowing what contributed to any failures. In the case I've just outlined, it is certainly not the site's fault if the nearest consultant is too far away. In that case, the Locator Tool would have performed perfectly. The lack of generated business may still be considered a “failure,” but not one attributable to the site.

Tracking Usage Trends Provides No Useful Data

Some people who recognize the weaknesses of web stats still argue that the log files can be used to look for trends in site usage. Unfortunately, this is also untrue. Again, caching is just as much of a spoiler here as it is elsewhere. The problem is compounded by the fact that network administrators and ISPs are constantly working to improve their system's performance. Their end-goal is to ensure their networks run quickly and efficiently for the benefit of their customers. This could mean that they could suddenly choose to place a larger (or smaller) amount of data in their cache and may choose different items to cache as well. Such activity could result in vast swings in the number of requests your site receives and in the number of the hosts making requests.



Monthly history

Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2005	0	0	0	0	0
Feb 2005	0	0	0	0	0
Mar 2005	231	252	499	1254	13.92 MB
Apr 2005	3107	3911	12818	26398	358.18 MB
May 2005	3270	4255	20596	34578	379.31 MB
Jun 2005	3254	4167	38134	64657	522.33 MB
Jul 2005	3129	4168	33789	57295	1.43 GB
Aug 2005	5542	7037	22809	54601	688.32 MB
Sep 2005	6425	8243	23231	51526	517.16 MB
Oct 2005	6501	8982	23268	49369	477.61 MB
Nov 2005	459	495	1728	3428	34.32 MB
Dec 2005	0	0	0	0	0
Total	31918	41510	176872	343106	4.35 GB

The figure above is a "Monthly history" generated by AWStats for a semi-popular personal site. Notice that traffic in August, September, and October is almost double that of April, May, June, and July? Despite the apparent evidence, absolutely nothing changed about the site at that time. Nothing was added and nothing was taken away. This could happen if some network administrator or ISP tweaked how their system was caching pages, resulting in more requests to the site.

Actions by a site's own management or the management of the organization can also change trend data. Anytime a site's structure changes, grows, or shrinks, the "trends" will as well. If the organization adds a new section and puts a big announcement about it on the home page, there will be a surge in traffic to that area of the site. If a new product is announced in the company's newsletter or on commercials then the site's overall traffic will grow. If the settings on the search engine are tweaked, pageviews may increase in some sections and decrease in others. Ignore the site for several months and the traffic will lull. Completely redesign the site or re-organize the content, and "trends" are no longer trends. Merely by managing the site, an organization invalidates the usefulness of the trend data by creating variation in the site. The more frequently new content is added to the site, the more opportunities exist for variation in the “trends.”

The Special Case of AOL

AOL further complicates statistics by assigning new IP addresses to their users in mid-session. Since AOL is the largest online service in the world, AOL is often a site's largest source of users. Since server log analysis tools often rely on the visitor’s IP address and/or hostname as a unique identifier, this means that a site’s logs can show data attributed to multiple "users" from AOL that actually belong to one person. Depending on how much time the person spends on the site, one AOL user may look like dozens of users.

One person, in a post to the Usenet Newsgroup alt.www.webmaster posted: "Here's a section of my access log that shows an AOL user requesting one page, followed by requests for the images on that page:" (edited for privacy)

```
195.93.21.98 – - [15/Mar/2006:12:44:37] "GET /xxxx/…
195.93.21.42 – - [15/Mar/2006:12:44:37] "GET /images/…
195.93.21.3 – - [15/Mar/2006:12:44:37] "GET /images/…
195.93.21.36 – - [15/Mar/2006:12:44:37] "GET /images/…
195.93.21.36 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.99 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.68 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.135 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.73 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.38 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.132 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.137 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.137 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.69 – - [15/Mar/2006:12:44:38] "GET /images/…
```

```
195.93.21.34 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.106 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.72 – - [15/Mar/2006:12:44:38] "GET /images/…
195.93.21.130 – - [15/Mar/2006:12:44:38] "GET /images/…
```

As you can see above, one visitor registers 16 different IP addresses in the log by requesting only one page. This is because the individual requests for each image increase the number of requests significantly, therefore seeming to indicate multiple users. Caching confuses traffic data on how long people view the pages, but AOL confuses it even more. AOL's use of dynamic IP addresses changing mid-session muddles everything in a site's traffic statistics. AOL even admits to the challenges this creates for site statistics. Located on AOL's Webmaster FAQ, they state:

> Q. Can I use the IP address of the request to track a member's access to my site?
> A. No. Because AOL uses proxy servers to service the requests made by members, webmasters see the IP address of the server, not the Dynamically Assigned Host Address (DAHA) of the member in their web site log files. The problem with trying to use the IP address to track access is that there may easily be multiple members assigned to a proxy server. All of the member requests would appear to be coming from one member if you assumed a relationship between member and IP address. In addition, members may be reassigned to a different proxy server during a session.

Problems created by AOL's dynamic IP addressing practice can include:

- The analytics tool may show many more users and visits than the site actually had.
- In the case of AOL, it may also show less users than the site actually had. As they state above, multiple users could potentially use the same IP address.
- The data on entry and exit points will be unreliable. The more AOL users the site has, the less reliable this data. Users will look like they've left when they've simply gotten a new IP address. Their next page view will make them

look like a new user.
- The data on length of visit and time on each page will also be unreliable.

Conclusion – Server Log Analysis Is an Unreliable Tool for Usability

It is recommended that an organization not spend extensive amounts of time and money to gain usability data from server logs. An organization would be better served by hiring an experienced human factors engineer to perform an expert review or conduct a formal study with users. The results would be much quicker, more accurate, and more informative.

[1]. HTTP. (n.d.). The American Heritage® Dictionary of the English Language, Fourth Edition. Retrieved September 30, 2007, from Dictionary.com website: http://dictionary.reference.com/browse/HTTP

[2]. Cache Planet Science. Retrieved September 30, 2007 http://www.scienceyear.com/about_sy/index.html?page=/about_sy/help/glossary.html

[3].Pew Internet and American Life Project. http://www.pewinternet.org/PPF/r/131/report_display.asp

Posted in Big Ideas, Case Studies, Discovery, Research, and Testing, Methods | 16 Comments »

16 Comments

Steve Fleckenstein The title of this article should be “Log files: 99% Bad” because, October 25, 2007 at 12:34 pm

like Nielsen’s article of similar name this one seems to have been written in 2000.

While you focus on log files — which I agree have a lot of limitations– your article completely ignores client side analytics tagging (which addresses many caching issues and which is used by thousands of sites). And while I agree that it is difficult to measure online success via analytics (but that too is possible to some degree with advanced features of some web analytics tools that can track a purchase funnel) I can still get a lot of great info out of web analytics tools that help me improve the site for users. A more balanced view would have been helpful– "log files aren't all that helpful but client side tagging is a viable workaround."

Karl Groves October 25, 2007 at 2:09 pm

As I say at the top of the article: "I realize that some analytics software gather data by other means than parsing log files and may in fact contain some features meant to overcome one or more of the criticisms I outline throughout this article. I do not discuss such programs, primarily because there is little consistency between them and ultimately they are just as poor at gathering real usability data as analytic tools which parse log files."

Regardless of how the data is captured, it is unrealistic to expect to glean usability data from analytics. Without knowing what task a specific user is attempting to do, there's no way of judging whether they've actually done it. Analytics tools only record requests. They can't record whether that request did or did not satisfy a user's need – only the user themselves can tell you that.

henrik persson October 25, 2007 at 3:55 pm

The article makes a great point and backs it up very well. I was hoping for a possible solution or alternative though.

Steve Fleckenstein October 25, 2007 at 4:29 pm

Here's an example. I discover from analytics (regardless of method) that a certain detailed content page buried in my site is requested many thousands of times. I'm using the word "request" specifically here to indicate my agreement about the inherent limitation of analytics. I also know that a link to this page does not appear anywhere in the top 3 levels of the site. Even without user testing (which is important, I'm not trying to debate that) would it be reasonable to draw the conclusion that users would be better served if I surfaced this content to a higher level?

Karl Groves October 25, 2007 at 5:08 pm

No, it would not be reasonable to do so. First, how do you know those requests are being performed by actual people and not bots? If you attempt to say that you filter out bots, I will show you one I've written which identifies itself successfully as IE. Second, you say this content page is being requested many thousands of times. To most proponents of analytics, this would seem to indicate success, would it not? After all, analytics tools only report on successful requests, right? So if it is successfully being requested, why move it? Last, and always, without knowing why the user is doing something, there's no way to determine whether the request was successful or not successful. Exactly what brand of analytics tool, through client-side means or server-side, does it tell you the user traveled to that page and then realized it was or was not what they were looking for? Without that information, any decision to modify the site due to analytics is purely driven by conjecture.

Cennydd Bowles October 25, 2007 at 5:42 pm

I agree with a lot of what you say but, to me, this article reads a little bit like one of those "BREAKING NEWS: Wikipedia may not be entirely accurate" stories. If you write a Masters' dissertation based on a Wikipedia article you're a fool. If you base any kind of big design decisions solely off log files you're a fool. But both *can* be a useful guide, a starting off point for future investigation. IAs are pretty adept at gathering data from numerous sources and appreciating the strengths and weaknesses of each observation. I'd content that analytics can contribute to this. To write it off completely, as this article does, seems a little over-zealous.

Cennydd, well stated. Karl, thanks for the response to my earlier comment. I agree

Steve Fleckenstein

October 25, 2007 at 7:28 pm

with your last statement– we shouldn't make changes based solely on analytics. I like Cennydd's statement that analytics provide a good starting off point for future investigation (especially if combined with search log analysis and other information). Bots, caching, etc. all complicate things but again that doesn't mean I should ignore web analytics completely.

Steve Fleckenstein

October 25, 2007 at 8:24 pm

P.S. And Karl, my apologies for the inflammatory opening line to the first comment– it deserved the negative rating it received. There's lots of good analysis in the article and it was irresponsible of me to toss the baby out with the bathwater, which I… er… think I accused you of doing.

Nick Besseling

October 25, 2007 at 9:01 pm

Hi Karl/

I tend agree with Cennydd and Steve.

While your in-depth analysis does really well to point out the limitations of analytics I think the idea that one would replace analytics with a usability expert or vice-vera is a wrong. They both serve different purposes and both have their faults and I really can't see how they are directly interchangeable.

I also think a lot of the analyitcs software around these days adds a lot more value and are 'smarter' than what it used to be.

I see analytics as another tool that adds weight to certain directions and provides valuable trend and user system data.

In the IA/Usability field I tend to feel there is far too many idelogical skews for and against certain approaches (I have been/am guilty of it myself) and feel that collecting all the information (both qualatative and quantative) possible that time/cost allows will be far more helpful in coming up with a suitable direction and solution.

Striking out the only available source of full scale system/user qualatative data (despite the limitations) seems a little strong.

Flexibility and a broad perspective is always the best approach and that nothing should be ruled out (unless it really is stretching the boundries of usefulness).

But depsite these points the article is a useful one as it does point out a log of specific analytic limitations I wasn’t fully aware of.

Cheers.

Peter Meyers

October 26, 2007 at 1:24 pm

You certainly make some valid points about the limitations of web logs, but I really think you’re throwing out the baby with the bathwater when it comes to analytics. Yes, analytics aren’t very useful for getting a window on any individual user, but they are useful as one more tool for understanding people in the aggregate. A lot of the caching issues do work themselves out in the wash, just as noise does in any system where you’re collecting large amounts of data. Personally, I have gleaned valuable usability insights by tracking long-term trends in exit pages, bounce rates, jumps in error-page activitiy, etc. In some of these cases, analytics helped me pinpoint a problem that I didn’t know was there, and that problem was completely verifiable (in other words, it wasn’t an illusion of bad data).

Look at the flip-side as well. Is testing a single user (or a couple of users) in a traditional usability test perfect, statistically speaking? Not remotely. Using a couple of users to generalize to thousands or tens of thousands is wildly unreliable, especially since those users may not even be representative. Does that mean the technique is useless? Of course not. In fact, using these techniques in tandem helps reduce noise even further: if you take the hypotheses from testing users individually and reference them against the clues you get from aggregate data, you’ll end up with information that is all the richer and more valuable.

My grad advisor used to say that there were three levels of understanding statistics: (1) knowing just enough to be dangerous, (2) knowing enough to play by the rules, and (3) and knowing when the rules don’t matter. I’ve met statisticians who can tear apart absolutely any real-world analysis someone could ever run, and they’re all level 2 people. The problem is, if you carry that too far, you can’t get anything done. I think it’s important to look at any tool objectively, know its faults, but then move forward and try to extract the best value you can. Web analytics aren’t a new, unproven tool, and while many people misuse them, I also know some flat-out brilliant people deriving real value from them for both clients and end-users.

Karl wrote: “First, how do you know those requests are being performed by actual

Peter Krantz

October 26, 2007 at 1:40 pm

people and not bots? If you attempt to say that you filter out bots, I will show you one I've written which identifies itself successfully as IE."

Well, analytics software typically relies on javascript to log a request when the page is viewed in the user's browser. This means that bots can be filtered from the collected data and that caching between the user and the web server are irrelevant. It is unlikely that bots implement a javascript engine. Saying that you can construct a bot to mimic that is like saying that I can instruct a user to lie in a usability test situation.

I agree that analytics software do not replace testing with real users but I still believe that you can get valuable data from them that can be indicative of user behaviour. This may help you design test scenarios better.

Melissa Robison

October 26, 2007 at 4:56 pm

Thanks for the article Karl, You provide good and specific examples outlining the limitations of using server logs to accurately analyze a site's usability, success, and/or user flows.

I agree with Peter and others above about the importance of looking at all available tools objectively, understanding their individual limitations, and then moving forward to extract the best value from all tools at hand.

While I'm not personally a fan of using server logs to measure usability, I've had to defend my position and discuss the pros and cons of using server log data with clients and peers. Often, I'm also in the position of recommending the best approach for clients to extract accurate user data from existing sources prior to planning a site redesign. Sometimes, the server logs or poorly implemented analytic software are the only data sources we have to review. The examples in your article will enable people in similar situations to analyze the data in an informed manner.

I think it's also important to keep in mind that less experienced User Experience "experts" are consulting with clients or working on sites every day. In my opinion, this discussion is a valuable one for them.

Speaking as a professional web traffic analyst (no longer practicing due to other

Andrea Wiggins

October 31, 2007 at 12:32 am

ambitions), the use of server logs is definitely “deprecated” at this point – no web analytics professional today will recommend using server log data for much more than system function monitoring, unless it’s an extremely unusual situation. Client-side data collection significantly improves on ALL of these issues; appropriate vendor solutions make it relatively simple to sessionize even AOL visitors, identify returning visitors with reasonable accuracy over very long periods of time, and filter out bots, regardless of whether they can identify themselves as IE. And a whole lot more.

I appreciate the thorough analysis of why server logs stink. They do. That’s what the web analytics community has been saying for years, and why they are trying so hard to get better data collection methods and analysis standards adopted. Appropriately collected data and careful analysis (by professionals who know the weaknesses of the methods and tools) minimizes threats to validity and provides reliable insights. Not all usability insights require advance knowledge of user goals; many branches of social science have spent centuries inferring goals from behaviors, so this is not a sufficient reason to discount the potential usefulness of this particular behavioral data for usability.

I think that this is a sadly one-sided perspective which is ignorantly dismissive of client-side analytics, and implore every reader to investigate the vast improvements to web analytic technology that have come about in the last two years. Things have really changed a lot with respect to data quality; the attitude of this article is very frustrating because it may prevent people from exploring the full range of options with respect to user data. Usability tests are almost entirely contrived, and web analytic data is almost entirely naturalistic; I can see no reason not to use both in method triangulation, as each data source balances the weakness of the other.

Alistair Harper

November 2, 2007 at 3:55 pm

I am a bit surprised by the focus of this article. Do many people still spend time analysing web server logs? The limitations of this form of analysis seem to have been well understood for sometime.

I agree that there are limitations in using web analytics tools in usability analysis. Fundamentally, these tools tell you WHAT a customer did on the site not WHY they did it or HOW frustrating the found the experience.

I still think that there is great value in web analytics, but not so much in providing answers, but in helping provide a focus. An example could be the analysis an extended application process. A web analytics tool can be used to identify the points at which users are giving up on completing the process, down to a specific form field on a specific page.

This can then be reviewed. Maybe that page needs to be redesigned, maybe the field asks for information that users need to go and look up, or maybe the field is asking for personal information that some users do not want to give out.

Analytics may not provide many answers – but it can help you understand exactly what questions need answering

Rob S. November 3, 2007 at 8:15 am

Great piece, Karl. Clearly a lot of time and research went into this article. I'm betting that if someone needs this much convincing that server logs are the wrong place for thorough usability analysis, they may be beyond saving 😉

I was happy to see Alistair (the previous comment) cite a specific example where analytics would potentially be useful. However, this could have also been gleaned from server logs (re: exit pages). Granted the level of granularity would be less (re: form field focus), but I still fail to see, in all of the "Woah Karl, that's too extreme!" comments – some examples of how exactly analytics are being used to gather usability data, for any purpose.

Arve Kvaloy November 20, 2007 at 10:37 pm

This article would have been groundbreaking six to seven years ago, when advanced users of analytics switched from analyzing serverlogs to pagetaging solutions. The screenshot from good ol' Awstats illustrates the historic value of the article.

All modern analytics solutions rely now on pagetaging: embedding a small javascript file on each page:

- Pagetagging remedies caching: I can save a pagetaged file localy on my computer, the pagetag script would still be set off when i open the page
- Robots and spiders? They dont set off pagetaging scripts so they want get counted
- Pagetags remedy proxies by placing a cookie which helps identify the user (and simple beacons for those that have javascripts turned off)

Even if the world of analytics is vastly improved from the above 2000/2001 scenario re-enacted, no serious analyst would try to suggest that you could establish real persons behind the term “Unique visitor” unless the user identified him-/herself by logging on to the site. But what analytics gives you are trends and patterns that can help you test and optimize your site.

And to some ecxtent you can infer visitor intent for example by:
- search phrases used to find your site (why do they visit the site?)
- search phrases used use in internal search (could show a lack in the navigation structure)
- combining clickstream with online surveys

A serious quantitative analysis should actually be mandatory before doing an expert review (at least on medium to large sites) so as to give the reviewer some possible problem areas to focus on instead of relying solely on iterating some heuristics, it could also give a more specific focus for user testing (analytics could point out the where and what, the user test could explain the why).

Sorry, comments are closed.

Karl Groves

Most-commented Stories

HTML’s Time is Over. Let’s Move On.
105 Comments

The Lazy IA’s Guide to Making Sitemaps
63 Comments

Blasting the Myth of the Fold
58 Comments

Welcome to Boxes and Arrows
55 Comments

Support our sponsors!

Become a sponsor

Stories

People

Jobs

About

Login

Copyright © 2014 Boxes and Arrows - All Rights Reserved