Toby Wallis Communications

Burnham House, Park Street, Ripon HG4 2BY, UK
Tel: +44(0)113-815-5318 Mobile: +44(0)7802-989-969
Email: toby@walliscommunications.com

A Guide to Visitor Statistics
(Webstats)

Published by
TOBY WALLIS COMMUNICATIONS
as a service to website owners

A Guide to Visitor Statistics
(Webstats)

CONTENTS

1. How webstats work

In this introductory document we aim to provide an overall understanding of what webstats are, how they are calculated, and how you can use them to get a view of the popularity of your website. A comprehensive description of the full complexities of webstats is beyond the scope of this brief introduction.

In common with most web hosting companies, Toby Wallis Communications' web servers maintain **Log Files** which collect information on the visits that a particular website receives. In their raw form the information in these files is pretty obscure and also, since one line is generated for every hit to a site, they can get extremely large. For example, here are the log entries for just one second of server activity on 15th January 2006:

```
riponcathedral.org.uk 82.43.149.49 - - [15/Jan/2006:07:05:09 +0000] "GET /awmdata/awmlib1.js HTTP/1.1" 200
37642 "http://www.riponcathedral.org.uk/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322)"
riponcathedral.org.uk 82.43.149.49 - - [15/Jan/2006:07:05:09 +0000] "GET /awmdata/dot.gif HTTP/1.1" 200 43
"http://www.riponcathedral.org.uk/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322)"
riponcathedral.org.uk 82.43.149.49 - - [15/Jan/2006:07:05:09 +0000] "GET /awmdata/bullets_26.gif HTTP/1.1" 200
833 "http://www.riponcathedral.org.uk/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322)"
riponcathedral.org.uk 82.43.149.49 - - [15/Jan/2006:07:05:09 +0000] "GET /awmdata/menuBG.gif HTTP/1.1" 200 848
"http://www.riponcathedral.org.uk/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322)"
riponcathedral.org.uk 82.43.149.49 - - [15/Jan/2006:07:05:09 +0000] "GET /pix/frontbanner.gif HTTP/1.1" 200
17969 "http://www.riponcathedral.org.uk/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322)"
riponcathedral.org.uk 82.43.149.49 - - [15/Jan/2006:07:05:09 +0000] "GET /pix/frontbg.gif HTTP/1.1" 200 89182
"http://www.riponcathedral.org.uk/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322)"
riponcathedral.org.uk 82.43.149.49 - - [15/Jan/2006:07:05:09 +0000] "GET /pix/quire.jpg HTTP/1.1" 200 43529
"http://www.riponcathedral.org.uk/" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322)"
```

Obviously this information must be analysed, condensed and presented in an intelligible form if it is to be of any use.

However there is a problem with the interpretation of the information in the log file. The root of this problem lies in the fact that a web page can be made up of several different files – text, separate images, media files and so on.

Here is an example to illustrate the information flow when a web page is requested. Imagine a simple web page, 'mypage.html', which is an HTML web page that contains two graphic images, 'myimage1.jpg', and 'myimage2.jpg'.

A typical server/visitor interaction might go something like this:

- The web browser (the visitor) asks for the URL mypage.html.
- The server sees the request and sends back the HTML page.
- The web browser reads the HTML file and sees that there are two inline graphic links in it, so it asks for the first one, myimage1.jpg.
- The server sees the request and sends back the graphic image.
- The web browser then asks for the second image, myimage2.jpg.
- The server sees the request and sends back the graphic image.
- The browser displays the web page and graphics for the user.

In the web server log, the following lines would be added:

```
192.168.45.13 - - [24/May/2005:11:20:39 -0400] "GET /mypage.html HTTP/1.1" 200 117
192.168.45.13 - - [24/May/2005:11:20:40 -0400] "GET /myimage1.jpg HTTP/1.1" 200
231
192.168.45.13 - - [24/May/2005:11:20:41 -0400] "GET /myimage2.jpg HTTP/1.1" 200
432
```

Here is the information we can obtain from the listing:

- The number of lines in the log file tells us that the server received 3 **hits** during the period that this log file covers. A hit is a server request.
- We can also calculate the number of hits each **URL** received (in this case, 1 hit each). A URL is a specific file definition.
- Along the same lines, we can see that the server received 3 hits from the **IP address** 192.168.45.13, and when those requests were received. An IP address is a unique identifier for the computer accessing the web page.
- The two numbers at the end of each line represent the **response code** and the **number of bytes** sent back to the browser. The response code is how the web server indicates how it handled the request. In this example, they are all 200, which means everything went OK.

Some more obscure numbers can be **calculated**, like the number of different response codes, number of hits within a given time period, total number of bytes sent to remote browsers, etc. Other numbers can be **implied** based on certain assumptions, however those cannot be considered entirely accurate, and some can even be very inaccurate.

The only information you can **accurately** determine is what IP address requested which URL and when. The total number of events logged is the total number of hits on the server.

2. What does this mean in practice?

There are therefore three types of information presented by a webstats analysis package:

- **Absolute**: specific information measured by the log files (e.g. IP number)
- **Calculated**: derived mathematically from absolute values (e.g. Number of Hits)
- **Implied**, which uses calculated numbers and certain assumptions to arrive at an estimated result (e.g. Visits – see below)

In general, reliability of the figures is less the further down the list you go; but at the same time the figures are more informative. For example, the **URL**, which is absolute, only tells you which file was requested at any particular time; but a **visit**, which is an implied value attempting to count the number of individual people visiting a web page, is potentially a much more useful figure.

So how can we use webstats to get an idea of how successful our website is? A useful working assumption is that the numbers are relative. In other words, their significance is not in their absolute value (though this is obviously a useful indicator) but in the way they develop over time. Even implied figures are always calculated with consistent assumptions; so a measure of the way Page Views move over a period of, say, six months can give a valuable indication of trends in the popularity of your site.

- If viewing a snapshot, for example stats for a single month, you should take an overview of all the different measures offered by the analysis software.

- If you are comparing your site's stats with those of another, or if an advertiser on your site asks for a specific metric, make sure as far as you can that the number is calculated the same way on each site. This is why the absolute value of hits is still used in some circumstances: although not precisely representative of the popularity of a site it is very simply defined and unlikely to have been distorted by calculation or assumptions.

- A useful procedure for assessing the popularity of your site is a combination of approaches. Many website owners do a month-to-month comparison of one or two metrics as well as an overview in (more or less) objective terms of the results for this particular month. This is why summary webstats are generally presented using a number of different metrics.

Finally, if you are serious about getting a picture of the popularity of your website, collect and record several metrics every month into a spreadsheet program. Then you can review a year's performance, identify any abnormal figures, graphically display trends, chart responses to changes in your website or promotional campaigns and so on.

3. Glossary of terms

Toby Wallis Communications servers use industry standard stats packages to produce a consistent and comparable set of statistics on the websites we host. These packages present summary results using 6 main metrics, which are in keeping with those used across the industry. It is important to understand the meaning of each metric and how the numbers are derived. The following glossary should help you interpret your webstats.

Absolute metrics are identified with [a]
Calculated metrics are identified with [c]
Implied metrics are identified with [i]

Stats are presented as both Monthly Totals and Daily Average; graphic

interpretation of the figures is also provided.

a. Main Headings

Hits	Files	Pages	Visits	Sites	KBytes

[c] **Hits** represent the total number of requests made to the server during the given time period (month, day, hour etc.)

[a] **Files** represent the total number of hits (requests) that actually resulted in something being sent back to the user. Not all hits will send data, such as 404-Not Found requests and requests for pages that are already in the browsers cache.

Note: By looking at the difference between hits and files, you can get a rough indication of [i] **Repeat Visitors**, *as the greater the difference between the two, the more people are requesting pages they already have cached (have viewed already).*

[c] **Pages** are those URLs that would be considered the actual page being requested, and not all of the individual items that make it up (such as graphics and audio clips). Some people call this metric *page views* or *page impressions*.

[i] **Visits** occur when some remote site makes a request for a page on your server for the first time. As long as the same site keeps making requests within a given timeout period, they will all be considered part of the same **Visit**. If the site makes a request to your server, and the length of time since the last request is greater than the specified timeout period (default is 30 minutes), a new **Visit** is started and counted, and the sequence repeats. Since only *pages* will trigger a visit, remotes sites that link to graphic and other non- page URLs will not be counted in the visit totals, reducing the number of false visits.

[c] **Sites** is the number of unique IP addresses/hostnames that made requests to the server. Care should be taken when using this metric for anything other than that. Many users can appear to come from a single site, or a single user can appear to come from several sites; so this metric should be used simply as a rough gauge as to the number of visitors to your server.

[a] A **KByte** (KB) is 1024 bytes (1 Kilobyte). Used to show the amount of data that was transferred between the server and the remote machine, based on the data found in the server log.

b. Common Definitions

A **Site** is a remote machine that makes requests to your server, and is based on the remote machines IP Address.

An **IP Address** is a unique identifier for any computer accessing the internet. According to circumstances it is possible for one IP address to represent many individual users, or one user to be represented by many IP addresses.

URL - Uniform Resource Locator. All requests made to a web server need to request something. A URL is that something, and represents an object somewhere on your server, that is accessible to the remote user. URLs can be of any type (HTML, Audio, Graphics, etc.)

Referrers are those URLs that lead a user to your site or caused the browser to request something from your server. The vast majority of requests are made from your own URLs, since most HTML pages contain links to other objects such as graphics files.

Search Strings are obtained from examining the referrer string and looking for known patterns from various search engines. The search engines and the patterns to look for can be specified by the user within a configuration file. The default will catch most of the major ones.

Note: Only available if that information is contained in the server logs.

User Agents are a fancy name for *browsers*. Internet Explorer, Firefox, Netscape, Opera, etc. are all **User Agents**, and each reports itself in a unique way to your server. Keep in mind however, that many browsers allow the user to change its reported name.

Note: Only available if that information is contained in the server logs.

Entry/Exit pages are those pages that were the first requested in a visit (**Entry**), and the last requested (**Exit**). These pages are calculated using the **Visits** logic above. When a visit is first triggered, the requested page is counted as an **Entry** page, and whatever the last requested URL was, is counted as an **Exit** page.

Countries are determined based on the top level domain of the requesting site. This is somewhat questionable however, as there is no longer strong enforcement of domains as there was in the past. A .COM domain may reside in the US, or somewhere else. An .IL domain may actually be in Israel, but it may equally be located in the US or elsewhere. The most common domains seen are .COM (US Commercial), .NET (Network), .ORG (Non-profit Organization) and .EDU (Educational). A large percentage may also be shown as *Unresolved/Unknown*, as a fairly large percentage of dialup and other customer access points do not resolve to a name and are left as an IP address.

Response Codes are defined as part of the HTTP/1.1 protocol (RFC 2068; See Chapter 10). These codes are generated by the web server and indicate the completion status of each request made to it.