WebSideStory case study:

HitBox Enterprise vs. log-file analysis

WebSideStory's HitBox audience-analysis technology and traditional log-file analysis tools may produce greatly different traffic statistics for the same site. To examine this difference, we compared page-view counts by HitBox Enterprise and a popular commercial log-file analysis tool for an actual commercial site. This comparison showed that the differences in traffic statistics are to a great extent the result of difference in data-collection methods, and that inaccuracies inherent in log-file analysis can produce significant errors even for typical sites.

10182 Telesis Court
San Diego, California 92121
(858) 546-0040
www.WebSideStory.com

Introduction

WebSideStory's HitBox audience-analysis technology and traditional log-file analysis tools may produce greatly different traffic statistics for the same site. This is not surprising – by collecting traffic data directly from the browsers of actual users rather than from the log files generated by Web servers, HitBox is generally more accurate than log-file analysis, and this difference in accuracy can result in significant differences in statistics.

To examine the differences between the two techniques, we recently conducted a study of page-view counts for a single day for an actual commercial site, comparing HitBox Enterprise, our premium audience-analysis service, to a popular commercial log-file analysis software tool. (In the interest of privacy, the site and log-file analysis tool are not identified here.)

The difference in page-view counts was considerable: the log-file analysis tool reported **19,360** page views for the day, while HitBox Enterprise reported just **4,739**.

This case study examines the following potential sources of differences in page-view counts:

- HitBox code
- HTML frames
- Artificial traffic
- Limited-display devices
- Other factors

As discussed below, this examination demonstrates that the HitBox Enterprise count was in fact accurate, and shows how various inaccuracies inherent in log-file analysis resulted in a large overcount for even this fairly typical site.

HitBox code

To implement HitBox on a site, the site owner inserts a section of HitBox code in the HTML for each page to be monitored. When a page containing this HitBox code is displayed on a user's browser, the code collects page-view and other traffic data.

Two issues regarding HitBox code are relevant to this study: pages that do not contain HitBox code, and the location of the HitBox code within each page.

Pages without HitBox code

When implementing HitBox on a site, the site owner may choose to include the HitBox code on certain pages and omit it from others. HitBox collect statistics only for those pages that include the HitBox code. In contrast, log-file analysis tools usually provide statistics for all pages unless configured otherwise. This can be a significant – and often overlooked – source of differences between statistics from HitBox and log-file analysis.

In our study, some of the pages on the subject site did not include the HitBox code. The log-file analysis tool counted 3,153 page views for these pages. To restrict our comparison to the pages monitored by both HitBox Enterprise and the log-file analysis tool – comparing apples to apples – we subtracted these 3,153 page views from the log-file analysis total of 19,360, giving an adjusted total of 16,207 and reducing the difference between the two techniques.

Location of HitBox code

In general, it is desirable to insert the HitBox code near the beginning of the HTML for the page. This ensures that the HitBox code is executed whenever the page is displayed. If the HitBox code is located later in the HTML – especially after a large graphic or other slow-loading element – a user might enter and leave the page before the HitBox code is executed. In this event, the page view would not be recorded, resulting in an undercount.

Before we began our study, we verified that the HitBox code was located near the beginning of the HTML for each page in our study. By doing this, we were able to confirm that unexecuted HitBox code was not a factor in the lower page-view count reported by HitBox Enterprise.

HTML frames

Frames are independently controllable areas on a web page, used to provide added flexibility in display and functionality. Typically, there is a separate HTML file for the page itself and for each frame on the page.

Frames can be a problem for log-file analysis. When a user requests a page containing one or more frames, the typical server log records one request for the page itself, plus one additional request for each frame. Log-file analysis tools generally count each request as a separate page view even though only one page is displayed to the user, resulting in a significant overcount.

HitBox does not have this problem. When applied correctly, the HitBox code appears just once in the HTML for the page and all its frames, so it is executed only once for the entire page. As a result, HitBox records the entire page as a single page view regardless of the number of frames it contains.

Since some of the more frequently viewed pages in our study contained multiple frames, this effect turned out to be the largest source of overcount by the log-file analysis tool – a total of 9,579 extra page views, about half of the total count! Subtracting these artificial page views reduced the log-file count to 6,628, much closer to the HitBox Enterprise count of 4,739.

Artificial traffic

Another common source of excess page-view counts by log-file analysis is artificial traffic – automated programs that request Web pages from servers but do not display those pages to users. Log-file analysis tools generally count such requests as page views even though they are not true views by users, resulting in an overcount. We examined the effects of two types of artificial traffic: monitoring tools and robots.

Monitoring tools

Many sites use proprietary or commercial tools such as SiteScope to monitor various aspects of site performance. Such tools may request pages from the Web server. Log-file analysis tools incorrectly count these requests as page views.

HitBox does not count such requests as page views. This is because the HitBox code is technically an image. Since monitoring tools typically do not execute image code, HitBox correctly records no page view for the pages they request.

In our study, the monitoring tool was configured to monitor just one of the pages in our study – the base page for a page containing multiple frames. When we decreased the log-file page-view count to eliminate frame effects as described in the previous section, we eliminated the extra page views caused by the monitoring tool. However, if the monitoring tool had been configured differently, the log-file count would have included hundreds of extra page views caused by the monitoring tool.

Robots

Robots (also called "spiders" or "crawlers") are programs that surf the Web automatically, following hypertext links and scanning site content. Since robots are not actual users, their activities need to be excluded from traffic statistics.

This is difficult with log-file analysis: in order to identify the activity of a robot, a log-file analysis tool needs to know about the robot, much as anti-virus software needs to know about a virus in order to detect it. Since there are thousands of robots – and new ones appear every day – log-file analysis tools cannot identify every one. (In fact, recent studies by WebSideStory have identified hundreds of robots not detected by popular log-file analysis tools.)

HitBox does not have this problem. Like the monitoring tools described above, robots do not execute the HitBox image code. As a result, HitBox automatically excludes robots' activities from its traffic statistics without the need to identify specific robots.

Identifying log-file page views caused by robots was a challenge. To accomplish this, we created a special software tool that identified potential robots by examining the USER_AGENT string in each log-file entry, which describes the *agent* (the browser or other tool) making the request. Our tool detected likely robots by analyzing the number of unique Class C network addresses and domains associated with each agent.

Using this tool, we identified about 1,500 log-file page views caused by robots – nearly three times the number that the log-file analysis tool detected using its robot list. Subtracting these artificial page views reduced the log-file count to about 5,700 – even closer to the HitBox Enterprise count of 4,739. It is important to note that the log-file page-view count may have included additional page views caused by robots that we did not detect.

Limited-display devices

PDAs and other limited-display devices are often configured not to display images. This does not affect log-file page-view counts – the log file records each page request by such devices just as it would for any device, and the log-file analysis tool counts each request as a page view. In contrast, since the HitBox code is technically an image, it is not executed when the page is displayed without images. As a result, HitBox does not count page views for such devices.

In our examination of page views caused by robots, we identified many page views from a service that downloads pages for viewing on limited-display devices. These downloads were counted as page views by the log-file analysis tool, but not by HitBox Enterprise, contributing to the difference between the two techniques.

This difference cannot be unequivocally identified as an overcount by log-file analysis or an undercount by HitBox; this depends on the requirements of the site owner. If site owner's intent is to display advertising (usually in image form) to users – which is typical for commercial sites – then HitBox is correct to omit these non-image page views from the count.

Other factors

To paraphrase the old adage, the Internet is only as fast as its slowest link. Although many users enjoy high-speed access to the Web, others contend with slow access due to low-speed connections, congested networks and other impediments.

In extreme cases, conditions like these may prevent HitBox from recording traffic statistics for these unfortunate users. Although we did not observe clear evidence of this effect in our study, it may have contributed to the difference between the two techniques.

Conclusion

Our comparison of HitBox Enterprise and a popular log-file analysis tool underscores two important points:

- The differences in traffic statistics between HitBox and log-file analysis tools are to a great extent the result of their different data-collection methods.
- Inaccuracies inherent in log-file analysis can produce significant errors even for typical sites.

This study is by no means an exhaustive comparison of the two techniques. There are many other differences, including accuracy of other statistics, level of detail, speed, accessibility, reliability, ease of operation – and ultimately, value to the site owner. Future studies will compare HitBox Enterprise and log-file analysis in some of these other important areas.

WebSideStory is the world's leading source of real-time Internet intelligence services for e-business. We offer a unique combination of technical and business capabilities: the innovative HitBox technology for real-time analysis of Internet traffic and e-commerce activity; a vast compilation of Internet usage data; and an expert service-provider business approach. We combine these capabilities in HitBox Enterprise, the industry's most powerful high-volume traffic-analysis service, and a variety of related services, products and sites. Together, these offerings give e-businesses the Internet intelligence they need to enhance the effectiveness of their Web sites and maximize the return on their marketing investment.