Web analytics
From Wikipedia, the free encyclopedia

Web analytics is the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage.[1]

Web analytics is not just a tool for measuring web traffic but can be used as a tool for business and market research, and to assess and improve the effectiveness of a web site. Web analytics applications can also help companies measure the results of traditional print or broadcast advertising campaigns. It helps one to estimate how traffic to a website changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a website and the number of page views. It helps gauge traffic and popularity trends which is useful for market research.

There are two categories of web analytics; off-site and on-site web analytics.

Off-site web analytics refers to web measurement and analysis regardless of whether you own or maintain a website. It includes the measurement of a website's potential audience (opportunity), share of voice (visibility), and buzz (comments) that is happening on the Internet as a whole.

On-site web analytics measure a visitor's behavior once on your website. This includes its drivers and conversions; for example, the degree to which different landing pages are associated with online purchases. On-site web analytics measures the performance of your website in a commercial context. This data is typically compared against key performance indicators for performance, and used to improve a web site or marketing campaign's audience response. Google Analytics is the most widely used on-site web analytics service; although new tools are emerging that provide additional layers of information, including heat maps and session replay.

Historically, web analytics has referred to on-site visitor measurement. However in recent years this has blurred, mainly because vendors are producing tools that span both categories.

Contents

- 1 On-site web analytics technologies
  - 1.1 Web server logfile analysis
  - 1.2 Page tagging
  - 1.3 Logfile analysis vs page tagging
    - 1.3.1 Advantages of logfile analysis
    - 1.3.2 Advantages of page tagging
    - 1.3.3 Economic factors
  - 1.4 Hybrid methods
  - 1.5 Geolocation of visitors
  - 1.6 Click analytics
  - 1.7 Customer lifecycle analytics
  - 1.8 Other methods
- 2 On-site web analytics - definitions
- 3 Common sources of confusion in web analytics
  - 3.1 The hotel problem
  - 3.2 New visitors + Repeat visitors unequal to total visitors
- 4 Web analytics methods
  - 4.1 Problems with cookies
- 5 Secure analytics (metering) methods
- 6 See also
- 7 References
- 8 Bibliography
- 9 External links

On-site web analytics technologies
Many different vendors provide on-site web analytics software and services. There are two main technical ways of collecting the data. The first and older method, server log file analysis, reads the logfiles in which the web server records file requests by browsers. The second method, page tagging, uses JavaScript embedded in the site page code to make image requests to a third-party analytics-dedicated server, whenever a page is rendered by a web browser or, if desired, when a mouse click occurs. Both collect data that can be processed to produce web traffic reports.

In addition, other data sources may be added to augment the web site behavior data described above. For example: e-mail open and click-through rates, direct mail campaign data, sales and lead history, or other data types as needed.

Web server logfile analysis

Web servers record some of their transactions in a logfile. It was soon realized that these logfiles could be read by a program to provide data on the popularity of the website. Thus arose web log analysis software.

In the early 1990s, web site statistics consisted primarily of counting the number of client requests (or hits) made to the web server. This was a reasonable method initially, since each web site often consisted of a single HTML file. However, with the introduction of images in HTML, and web sites that spanned multiple HTML files, this count became less useful. The first true commercial Log Analyzer was released by IPRO in 1994.[2]

Two units of measure were introduced in the mid-1990s to gauge more accurately the amount of human activity on web servers. These were page views and visits (or sessions). A page view was defined as a request made to the web server for a page, as opposed to a graphic, while a visit was defined as a sequence of requests from a uniquely identified client that expired after a certain amount of inactivity, usually 30 minutes. The page views and visits are still commonly displayed metrics, but are now considered rather rudimentary.

The emergence of search engine spiders and robots in the late 1990s, along with web proxies and dynamically assigned IP addresses for large companies and ISPs, made it more difficult to identify unique human visitors to a website. Log analyzers responded by tracking visits by cookies, and by ignoring requests from known spiders. [citation needed]

The extensive use of web caches also presented a problem for logfile analysis. If a person revisits a page, the second request will often be retrieved from the browser's cache, and so no request will be received by the web server. This means that the person's path through the site is lost. Caching can be defeated by configuring the web server, but this can result in degraded performance for the visitor and bigger load on the servers.[citation needed]

Page tagging

Concerns about the accuracy of logfile analysis in the presence of caching, and the desire to be able to perform web analytics as an outsourced service, led to the second data collection method, page tagging or 'Web bugs'.

In the mid-1990s, Web counters were commonly seen — these were images included in a web page that showed the number of times the image had been requested, which was an estimate of the number of visits to that page. In the late 1990s this concept evolved to include a small invisible image instead of a visible one, and, by using JavaScript, to pass along with the image request certain information about the page and the visitor. This information can then be processed remotely by a web analytics company, and extensive statistics generated.

The web analytics service also manages the process of assigning a cookie to the user, which can uniquely identify them during their visit and in subsequent visits. Cookie acceptance rates vary significantly between web sites and may affect the quality of data collected and reported.

Collecting web site data using a third-party data collection server (or even an in-house data collection server) requires an additional DNS look-up by the user's computer to determine the IP address of the collection server. On occasion, delays in completing a successful or failed DNS look-ups may result in data not being collected.

With the increasing popularity of Ajax-based solutions, an alternative to the use of an invisible image is to implement a call back to the server from the rendered page. In this case, when the page is rendered on the web browser, a piece of Ajax code would call back to the server and pass information about the client that can then be aggregated by a web analytics company. This is in some ways flawed by browser restrictions on the servers which can be contacted with XmlHttpRequest objects. Also, this method can lead to slightly lower reported traffic levels, since the visitor may stop the page from loading in mid-response before the Ajax call is made.
Logfile analysis vs page tagging

Both logfile analysis programs and page tagging solutions are readily available to companies that wish to perform web analytics. In some cases, the same web analytics company will offer both approaches. The question then arises of which method a company should choose. There are advantages and disadvantages to each approach.[3]

Advantages of logfile analysis

The main advantages of logfile analysis over page tagging are as follows:

- The web server normally already produces logfiles, so the raw data is already available. No changes to the website are required.
- The data is on the company’s own servers, and is in a standard, rather than a proprietary, format. This makes it easy for a company to switch programs later, use several different programs, and analyze historical data with a new program.
- Logfiles contain information on visits from search engine spiders, which generally do not execute JavaScript on a page and are therefore not recorded by page tagging. Although these should not be reported as part of the human activity, it is useful information for search engine optimization.
- Logfiles require no additional DNS lookups or TCP slow starts. Thus there are no external server calls which can slow page load speeds, or result in uncounted page views.
- The web server reliably records every transaction it makes, e.g. serving PDF documents and content generated by scripts, and does not rely on the visitors’ browsers cooperating.

Advantages of page tagging

The main advantages of page tagging over logfile analysis are as follows:

- Counting is activated by opening the page (given that the web client runs the tag scripts), not requesting it from the server. If a page is cached, it will not be counted by the server. Cached pages can account for up to one-third of all pageviews. Not counting cached pages seriously skews many site metrics. It is for this reason server-based log analysis is not considered suitable for analysis of human activity on websites.
- Data is gathered via a component (“tag”) in the page, usually written in JavaScript, though Java can be used, and increasingly Flash is used. Ajax can also be used in conjunction with a server-side scripting language (such as PHP) to manipulate and (usually) store it in a database, basically enabling complete control over how the data is represented.
- The script may have access to additional information on the web client or on the user, not sent in the query, such as visitors’ screen sizes and the price of the goods they purchased.
- Page tagging can report on events which do not involve a request to the web server, such as interactions within Flash movies, partial form completion, mouse events such as onClick, onMouseOver, onFocus, onBlur etc.
- The page tagging service manages the process of assigning cookies to visitors; with logfile analysis, the server has to be configured to do this.
- Page tagging is available to companies who do not have access to their own web servers.
- Lately page tagging has become a standard in web analytics.[4]

Economic factors

Logfile analysis is almost always performed in-house. Page tagging can be performed in-house, but it is more often provided as a third-party service. The economic difference between these two models can also be a consideration for a company deciding which to purchase.

- Logfile analysis typically involves a one-off software purchase; however, some vendors are introducing maximum annual page views with additional costs to process additional information. In addition to commercial offerings, several open-source logfile analysis tools are available free of charge.
- For Logfile analysis you have to store and archive your own data, which often grows very large quickly. Although the cost of hardware to do this is minimal, the overhead for an IT department can be considerable.
- For Logfile analysis you need to maintain the software, including updates and security patches.
- Complex page tagging vendors charge a monthly fee based on volume i.e. number of pageviews per month.
Which solution is cheaper to implement depends on the amount of technical expertise within the company, the vendor chosen, the amount of activity seen on the web sites, the depth and type of information sought, and the number of distinct web sites needing statistics.

Regardless of the vendor solution or data collection method employed, the cost of web visitor analysis and interpretation should also be included. That is, the cost of turning raw data into actionable information. This can be from the use of third party consultants, the hiring of an experienced web analyst, or the training of a suitable in-house person. A cost-benefit analysis can then be performed. For example, what revenue increase or cost savings can be gained by analysing the web visitor data?

Hybrid methods

Some companies produce solutions that collect data through both logfiles and page tagging and can analyze both kinds. By using a hybrid method, they aim to produce more accurate statistics than either method on its own. An early hybrid solution was produced in 1998 by Rufus Evison.\[citation needed\]

Geolocation of visitors

With IP geolocation, it is possible to track visitors location. Using IP geolocation database or API, visitors can be geolocated to city, region or country level.\[5\]

IP Intelligence, or Internet Protocol (IP) Intelligence, is a technology that maps the Internet and catalogues IP addresses by parameters such as geographic location (country, region, state, city and postcode), connection type, Internet Service Provider (ISP), proxy information, and more. The first generation of IP Intelligence was referred to as geotargeting or geolocation technology. This information is used by businesses for online audience segmentation in applications such online advertising, behavioral targeting, content localization (or website localization), digital rights management, personalization, online fraud detection, geographic rights management, localized search, enhanced analytics, global traffic management, and content distribution.

Click analytics

Click analytics is a special type of web analytics that gives special attention to clicks.

Commonly, click analytics focuses on on-site analytics. An editor of a web site uses click analytics to determine the performance of his or her particular site, with regards to where the users of the site are clicking.

Also, click analytics may happen real-time or "unreal"-time, depending on the type of information sought. Typically, front-page editors on high-traffic news media sites will want to monitor their pages in real-time, to optimize the content. Editors, designers or other types of stakeholders may analyze clicks on a wider time frame to aid them assess performance of writers, design elements or advertisements etc.

Data about clicks may be gathered in at least two ways. Ideally, a click is "logged" when it occurs, and this method requires some functionality that picks up relevant information when the event occurs. Alternatively, one may institute the assumption that a page view is a result of a click, and therefore log a simulated click that led to that page view.

Customer lifecycle analytics

Customer lifecycle analytics is a visitor-centric approach to measuring that falls under the umbrella of lifecycle marketing.\[citation needed\] Page views, clicks and other events (such as API calls, access to third-party services, etc.) are all tied to an individual visitor instead of being stored as separate data points. Customer lifecycle analytics
attempts to connect all the data points into a marketing funnel that can offer insights into visitor behavior and website optimization.\[citation needed\]

### Other methods

Other methods of data collection are sometimes used. Packet sniffing collects data by sniffing the network traffic passing between the web server and the outside world. Packet sniffing involves no changes to the web pages or web servers. Integrating web analytics into the web server software itself is also possible.\[6\] Both these methods claim to provide better real-time data than other methods.

### On-site web analytics - definitions

There are no globally agreed definitions within web analytics as the industry bodies have been trying to agree on definitions that are useful and definitive for some time. The main bodies who have had input in this area have been JICWEBS (The Joint Industry Committee for Web Standards in the UK and Ireland) (http://www.jicwebs.org/), ABCe (Audit Bureau of Circulations electronic, UK and Europe) (http://www.abc.org.uk/), The DAA (Digital Analytics Association) (http://www.digitalanalyticsassociation.org/default.asp?page=aboutus), formally known as the WAA (Web Analytics Association, US) and to a lesser extent the IAB (Interactive Advertising Bureau). However, many terms are used in consistent ways from one major analytics tool to another, so the following list, based on those conventions, can be a useful starting point. Both the WAA and the ABCe provide more definitive lists for those who are declaring their statistics as using the metrics defined by either.

- **Hit** - A request for a file from the web server. Available only in log analysis. The number of hits received by a website is frequently cited to assert its popularity, but this number is extremely misleading and dramatically overestimates popularity. A single web-page typically consists of multiple (often dozens) of discrete files, each of which is counted as a hit as the page is downloaded, so the number of hits is really an arbitrary number more reflective of the complexity of individual pages on the website than the website's actual popularity. The total number of visits or page views provides a more realistic and accurate assessment of popularity.

- **Page view** - A request for a file, or sometimes an event such as a mouse click, that is defined as a page in the setup of the web analytics tool. An occurrence of the script being run in page tagging. In log analysis, a single page view may generate multiple hits as all the resources required to view the page (images, .js and .css files) are also requested from the web server.

- **Event** - A discrete action or class of actions that occurs on a website. A page view is a type of event. Events also encapsulate clicks, form submissions, keypress events, and other client-side user actions.

- **Visit / Session** - A visit or session is defined as a series of page requests or, in the case of tags, image requests from the same uniquely identified client. A visit is considered ended when no requests have been recorded in some number of elapsed minutes. A 30 minute limit ("time out") is used by many analytics tools but can, in some tools, be changed to another number of minutes. Analytics data collectors and analysis tools have no reliable way of knowing if a visitor has looked at other sites between page views; a visit is considered one visit as long as the events (page views, clicks, whatever is being recorded) are 30 minutes or less closer together. Note that a visit can consist of one page view, or thousands.

- **First Visit / First Session** - (also called 'Absolute Unique Visitor' in some tools) A visit from a uniquely identified client that has theoretically not made any previous visits. Since the only way of knowing whether the uniquely identified client has been to the site before is the presence of a persistent cookie that had been received on a previous visit, the First Visit label is not reliable if the site's cookies have been deleted since their previous visit.

- **Visitor / Unique Visitor / Unique User** - The uniquely identified client that is generating page views or hits within a defined time period (e.g. day, week or month). A uniquely identified client is usually a combination of a machine (one's desktop computer at work for example) and a browser (Firefox on that machine). The identification is usually via a persistent cookie that has been placed on the computer by the site page code. An older method, used in log file analysis, is the unique combination of the computer's IP address and the User Agent (browser) information provided to the web server by the browser. It is important to understand that the "Visitor" is not the same as the human being sitting at the computer at the time of the visit, since an individual human can use different computers or, on the same computer, can use different browsers, and will be seen as a different visitor in each circumstance. Increasingly, but still somewhat rarely, visitors are uniquely identified by Flash LSO's (Local Shared Object), which are less susceptible to privacy enforcement.

- **Repeat Visitor** - A visitor that has made at least one previous visit. The period between the last and current visit is called visitor recency and is measured in days.

- **New Visitor** - A visitor that has not made any previous visits. This definition creates a certain amount of...
confusion (see common confusions below), and is sometimes substituted with analysis of first visits.

- **Impression** - The most common definition of "Impression" is an instance of an advertisement appearing on a viewed page. Note that an advertisement can be displayed on a viewed page below the area actually displayed on the screen, so most measures of impressions do not necessarily mean an advertisement has been viewable.

- **Single Page Visit / Singleton** - A visit in which only a single page is viewed (a 'bounce').

- **Bounce Rate** - The percentage of visits that are single page visits.

- **Exit Rate / % Exit** - A statistic applied to an individual page, not a web site. The percentage of visits seeing a page where that page is the final page viewed in the visit.

- **Page Time Viewed / Page Visibility Time / Page View Duration** - The time a single page (or a blog, Ad Banner...) is on the screen, measured as the calculated difference between the time of the request for that page and the time of the next recorded request. If there is no next recorded request, then the viewing time of that instance of that page is not included in reports.

- **Session Duration / Visit Duration** - Average amount of time that visitors spend on the site each time they visit. This metric can be complicated by the fact that analytics programs can not measure the length of the final page view.[7]

- **Average Page View Duration** - Average amount of time that visitors spend on an average page of the site.

- **Active Time / Engagement Time** - Average amount of time that visitors spend actually interacting with content on a web page, based on mouse moves, clicks, hovers and scrolls. Unlike Session Duration and Page View Duration / Time on Page, this metric can accurately measure the length of engagement in the final page view, but it is not available in many analytics tools or data collection methods.

- **Average Page Depth / Page Views per Average Session** - Page Depth is the approximate "size" of an average visit, calculated by dividing total number of page views by total number of visits.

- **Frequency / Session per Unique** - Frequency measures how often visitors come to a website in a given time period. It is calculated by dividing the total number of sessions (or visits) by the total number of unique visitors during a specified time period, such as a month or year. Sometimes it is used interchangeable with the term "loyalty."

- **Click path** - the chronological sequence of page views within a visit or session.

- **Click** - "refers to a single instance of a user following a hyperlink from one page in a site to another".[8]

- **Site Overlay** is a report technique in which statistics (clicks) or hot spots are superimposed, by physical location, on a visual snapshot of the web page.

**Common sources of confusion in web analytics**

**The hotel problem**

The hotel problem is generally the first problem encountered by a user of web analytics. The problem is that the unique visitors for each day in a month do not add up to the same total as the unique visitors for that month. This appears to an inexperienced user to be a problem in whatever analytics software they are using. In fact it is a simple property of the metric definitions.

The way to picture the situation is by imagining a hotel. The hotel has two rooms (Room A and Room B).

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Room A</td>
<td>John</td>
<td>John</td>
<td>Mark</td>
<td>2 Unique Users</td>
</tr>
<tr>
<td>Room B</td>
<td>Mark</td>
<td>Jane</td>
<td>Jane</td>
<td>2 Unique Users</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>?</td>
</tr>
</tbody>
</table>

As the table shows, the hotel has two unique users each day over three days. The sum of the totals with respect to the days is therefore six.

During the period each room has had two unique users. The sum of the totals with respect to the rooms is therefore four.

Actually only three visitors have been in the hotel over this period. The problem is that a person who stays in a room for two nights will get counted twice if you count them once on each day, but is only counted once if you are looking at
the total for the period. Any software for web analytics will sum these correctly for the chosen time period, thus leading to the problem when a user tries to compare the totals.

**New visitors + Repeat visitors unequal to total visitors**

Another common misconception in web analytics is that the sum of the new visitors and the repeat visitors ought to be the total number of visitors. Again this becomes clear if the visitors are viewed as individuals on a small scale, but still causes a large number of complaints that analytics software cannot be working because of a failure to understand the metrics.

Here the culprit is the metric of a new visitor. There is really no such thing as a new visitor when you are considering a web site from an ongoing perspective. If a visitor makes their first visit on a given day and then returns to the web site on the same day they are both a new visitor and a repeat visitor for that day. So if we look at them as an individual which are they? The answer has to be both, so the definition of the metric is at fault.

A new visitor is not an individual; it is a facet of the web measurement. For this reason it is easiest to conceptualize the same facet as a first visit (or first session). This resolves the conflict and so removes the confusion. Nobody expects the number of first visits to add to the number of repeat visitors to give the total number of visitors. The metric will have the same number as the new visitors, but it is clearer that it will not add in this fashion.

On the day in question there was a first visit made by our chosen individual. There was also a repeat visit made by the same individual. The number of first visits and the number of repeat visits will add up to the total number of visits for that day.

**Web analytics methods**

**Problems with cookies**

Historically, vendors of page-tagging analytics solutions have used third-party cookies sent from the vendor's domain instead of the domain of the website being browsed. Third-party cookies can handle visitors who cross multiple unrelated domains within the company's site, since the cookie is always handled by the vendor's servers.

However, third-party cookies in principle allow tracking an individual user across the sites of different companies, allowing the analytics vendor to collate the user's activity on sites where he provided personal information with his activity on other sites where he thought he was anonymous. Although web analytics companies deny doing this, other companies such as companies supplying banner ads have done so. Privacy concerns about cookies have therefore led a noticeable minority of users to block or delete third-party cookies. In 2005, some reports showed that about 28% of Internet users blocked third-party cookies and 22% deleted them at least once a month.[9]

Most vendors of page tagging solutions have now moved to provide at least the option of using first-party cookies (cookies assigned from the client subdomain).

Another problem is cookie deletion. When web analytics depend on cookies to identify unique visitors, the statistics are dependent on a persistent cookie to hold a unique visitor ID. When users delete cookies, they usually delete both first- and third-party cookies. If this is done between interactions with the site, the user will appear as a first-time visitor at their next interaction point. Without a persistent and unique visitor id, conversions, click-stream analysis, and other metrics dependent on the activities of a unique visitor over time, cannot be accurate.

Cookies are used because IP addresses are not always unique to users and may be shared by large groups or proxies. In some cases, the IP address is combined with the user agent in order to more accurately identify a visitor if cookies are not available. However, this only partially solves the problem because often users behind a proxy server have the same user agent. Other methods of uniquely identifying a user are technically challenging and would limit the trackable audience or would be considered suspicious. Cookies are the selected option because they reach the lowest common denominator without using technologies regarded as spyware.[10][11]

**Secure analytics (metering) methods**

All the methods described above (and some other methods not mentioned here, like sampling) have the central problem of being vulnerable to manipulation (both inflation and deflation). This means these methods are imprecise and insecure (in any reasonable model of security). This issue has been addressed in a number of papers [12] [13]
but to-date the solutions suggested in these papers remain theoretic, possibly due to lack of interest from the engineering community, or because of financial gain the current situation provides to the owners of big websites. For more details, consult the aforementioned papers.

See also

- Mobile Web Analytics
- Eurocrypt
- List of web analytics software
- Web bug
- Web log analysis software
- Online video analytics
- Post-click marketing
- Geolocation
- Geolocation software
- Geomarketing
- Geotargeting
- Internet Protocol
- IP Address
- Website correlation
- Website localization
- Clickstream
- HTTP cookie

References

1. ^ The Official DAA Definition of Web Analytics (http://www.webanalyticsassociation.org/?page=aboutus)
2. ^ Web Traffic Data Sources and Vendor Comparison (http://www.advanced-web-metrics.com/docs/web-data-sources.pdf) by Brian Clifton and Omega Digital Media Ltd
3. ^ Increasing Accuracy for Online Business Growth (http://www.advanced-web-metrics.com/blog/2008/02/16/accuracy-whitepaper/) - a web analytics accuracy whitepaper
6. ^ Web analytics integrated into web software itself (http://portal.acm.org/citation.cfm?id=1064677.1064679&coll=GUIDE&dl=GUIDE&CFID=6649216&CFTOKEN=93187844)
Bibliography


External links


Categories: Internet marketing | Web analytics

- This page was last modified on 25 December 2013 at 09:55.
- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.