

Web Analytics Overview

Guangzhi Zheng

Southern Polytechnic State University, USA

Svetlana Peltsverger

Southern Polytechnic State University, USA

INTRODUCTION

Web analytics is the technology and method for the collection, measurement, analysis and reporting of websites and web applications usage data (Burby & Brown, 2007). Web analytics has been growing ever since the development of the World Wide Web. It has grown from a simple function of HTTP (Hypertext Transfer Protocol) traffic logging to a more comprehensive suite of usage data tracking, analysis, and reporting. The web analytics industry and market are also booming with a plethora of tools, platforms, jobs, and businesses. The market was projected to reach 1 billion in 2014 with an annual growth rate more than 15% (Lovett, 2009).

Web analytics technologies are usually categorized into on-site and off-site web analytics. On-site web analytics refers to data collection on the current site (Kaushik, 2009). It is used to effectively measure many aspects of direct user-website interactions, including number of visits, time on site, click path, etc. Off-site analytics is usually offered by third party companies such as Twitalyzer (<http://twitalyzer.com>) or Sweetspot (<http://www.sweetspotintelligence.com>). It includes data from other sources such as surveys, market report, competitor comparison, public information, etc. This article provides an overview of on-site web analytics, with a focus on categorizing and explaining data, sources, collection methods, metrics and analysis methods.

BACKGROUND

Log files have been used to keep track of web requests since World Wide Web emerged and the first widely used browser Mosaic was released in 1993. One of

the pioneers of web log analysis was WebTrends, a Portland, Oregon based company, which conducted website analytics using data collected from web server logs. In the same year, WebTrends created the first commercial website analytics software. In 1995, Dr. Stephen Turner created Analog, the first free log file analysis software. In 1996, WebSideStory offered hit counter as a service for websites that would display a banner. Web server logs have some limits in types of data collected. For example, they could not provide information about visitors' screen sizes, user interactions with page elements, mouse events such as clicking and hovering, etc. The new technique of page tagging is able to overcome the limitation and gets more popular recently.

The fundamental basis of web analytics is collection and analysis of website usage data. Today, web analytics is used in many industries for different purposes, including traffic monitoring, e-commerce optimization, marketing/advertising, web development, information architecture, website performance improvement, web-based campaigns/programs, etc. Some of the major web analytics usages are:

1. **Improving website/application design and user experience:** This includes optimizing website information architecture, navigation, content presentation/layout, and user interaction. It also helps to identify user interest/attention areas and improve web application features. A particular example is a heat map that highlights areas of a webpage with higher than average click rate and helps determine if intended link/content is in the right place.
2. **Optimizing e-Commerce and improving e-CRM on customer orientation, acquisition and retention:** More and more companies analyze

DOI: 10.4018/978-1-4666-5888-2.ch756

website usage data in order to understand customers' needs to increase traffic and ultimately increase their revenue. Different sites can have different goals like selling more products and attracting more users to generate more income through advertisements. Websites want to keep visitors longer (reducing bounce rate) to encourage users to return and to make every visit end with completion of targeted action (conversion).

3. **Tracking and measuring success of actions and programs such as commercial campaigns:** To bring value, web analytics must differentiate between a wide variety of traffic sources, marketing channels, and visitor types. A common question is: "where did visitors learn that information?" For example, parameters used in tracking direct traffic from email, social media, or mobile devices allow correlation of traffic sources with marketing campaign cost, which helps to evaluate return on investments.
4. **Identifying problems and improving performance of web applications:** The study performed by Tag Man shows a significant correlation between page-load time and the likelihood of a user to convert (TagMan, 2012). Web analytics helps to address this issue. Page loading metrics such as average page load time by browser and geographic location are used to measure performance. Both real-time and historical performance analysis allow proactive detection, investigation, and diagnosis of performance issues. Improvements may range from simple image optimization to modification of the expiration date in the HTTP headers to force browsers to use cached website content. A heat map might help to reveal website errors, such as that users click on buttons or images without links. The same techniques can be used by developers of web based applications and games to add/modify software features.

DATA COLLECTION AND ANALYSIS

Data and Sources

The fundamental goal of web analytics is to collect and analyze web traffic and usage patterns. A common

way to study this data is to use the dimensional model (Hu & Cercone, 2004). Under this model, there are two major types of data: facts or measurement data and dimensional data that describe facts from different aspects and levels. Facts data are mainly about usage count and time. The most basic measure is a page view, which is a single request for a web page. Count of user actions such as mouse clicks can also be used as a measure. Various metrics are calculated based on basic measures and dimensions. Dimensional data are much more complex. Major types of dimensions include time, content, location, user client information (such as operating system, browser type, screen size, etc.), and user or session.

Both measurement data and dimensional data come from a number of sources, which can be categorized into the following 4 types:

1. Direct HTTP request data.
2. Application level data sent with HTTP requests.
3. Network level and server generated data associated with HTTP requests.
4. External data.

Direct HTTP request data directly come from HTTP request messages. An HTTP request is a message sent by a web client (browser) to a web server to request a resource (a web page or a web page element like an image). Traditionally, web traffic measurement is directly based on web resource visits (commonly called page view). Then each request is further described by a number of dimensions, such as page, visitor, technology, etc. The format of the HTTP 1.1 request is specified in IETF RFC 2616 (Fielding, Gettys, & Mogul, 1999). A typical HTTP request message is shown in Figure 1.

An HTTP request consists of a request command (the first line) and HTTP headers. The request command includes the required URI (unified resource identifier) information. A URI generally includes a host's domain or IP and a directory path. If the host information is not included as a part of the URI, then the "host" header has to be provided. The URI is the key information that leads to the count of a page/resource views. HTTP headers are pairs of field names and values. HTTP 1.1 specification defines a set of headers that can be included. These headers describe request and client characteristics. Most of the header

8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/web-analytics-overview/112470

Related Content

State of the Art and Future Trends of Datacenter Networks

George Michelogiannakis (2015). *Encyclopedia of Information Science and Technology, Third Edition* (pp. 1883-1892).

www.irma-international.org/chapter/state-of-the-art-and-future-trends-of-datacenter-networks/112593

I-Rough Topological Spaces

Boby P. Mathewand Sunil Jacob John (2016). *International Journal of Rough Sets and Data Analysis* (pp. 98-113).

www.irma-international.org/article/i-rough-topological-spaces/144708

Communities of Practice as a Source of Open Innovation

Diane-Gabrielle Tremblay (2018). *Encyclopedia of Information Science and Technology, Fourth Edition* (pp. 5027-5035).

www.irma-international.org/chapter/communities-of-practice-as-a-source-of-open-innovation/184205

Towards a Minimal Realisable System Dynamics Project Model

A. S. White (2012). *International Journal of Information Technologies and Systems Approach* (pp. 57-73).

www.irma-international.org/article/towards-minimal-realizable-system-dynamics/62028

Secure Electronic Healthcare Records Management in Wireless Environments

Petros Belsis, Christos Skourlasand Stefanos Gritzalis (2013). *Interdisciplinary Advances in Information Technology Research* (pp. 202-219).

www.irma-international.org/chapter/secure-electronic-healthcare-records-management/74542