Chapter 12

Web Access Patterns of Actual Human Visitors and Web Robots: A Correlated Examination

Dilip Singh Sisodia
National Institute of Technology Raipur, India

ABSTRACT

Web robots are autonomous software agents used for crawling websites in a mechanized way for non-malicious and malicious reasons. With the popularity of Web 2.0 services, web robots are also proliferating and growing in sophistication. The web servers are flooded with access requests from web robots. The web access requests are recorded in the form of web server logs, which contains significant knowledge about web access patterns of visitors. The presence of web robot access requests in log repositories distorts the actual access patterns of human visitors. The human visitors’ actual web access patterns are potentially useful for enhancement of services for more satisfaction or optimization of server resources. In this chapter, the correlative access patterns of human visitors and web robots are discussed using the web server access logs of a portal.

INTRODUCTION

Web robots are automated programs mainly designed to traverse websites over the internet. Web robots are used for a variety of functions including searching, indexing, hacking, scraping, spamming and spying, etc. (Gaffan, 2012). With the advent of web 2.0 services, web robots are playing a key role in everything that we do online and shaping our web experience. It is believed that first web robots were introduced in 1993 and since their origin, they are escalating with the unprecedented rate. Web robots are very simple to create as well as offer a great job by circumventing the collection of information (Tan & Kumar, 2002). Depending on their core functionality Web robots are also known as following (Derek Doran & Gokhale, 2011):

DOI: 10.4018/978-1-5225-3870-7.ch012
Web Access Patterns of Actual Human Visitors and Web Robots

1. **Indexers (or Search Engine Crawlers):** This seeks to harvest as much web content as possible on a regular basis, to build and maintain large search indexes.

2. **Analyzers (or Shopping Bots):** It is used to crawl the web to compare prices and products sold by different e-Commerce sites.

3. ** Experimenters (Focused Crawlers):** This seeks and acquires web pages belonging to pre-specified thematic areas.

4. **Harvesters (Email Harvesters):** This is used to collect email addresses on behalf of email marketing companies or spammers.

5. **Verifiers (Site-Specific Crawlers):** It is used to perform various website maintenance chores, such as mirroring web sites or discovering their broken links.

6. **RSS Crawlers:** It is used to retrieve information from RSS feeds on a web site or a blog.

7. **Scrapers:** It is used to create copies of websites for malicious purposes automatically.

The normal perception is that the major chunk of web server resources is used to handle human visitor's generated traffic for any web portal. This perception is changed if we observed the recent reports (Gaffan, 2012) which state that major portion of web traffic is generated through automatic software agents. Most website owners simply rely on web analytics tools (“Google Analytics,” 2013) to track who’s visiting their site. However, these tools don’t show you 51% of your site’s traffic including some seriously shady non-human visitors such as web robots.

**RELATED WORK**

The various studies thoroughly studied web access pattern induced by human visitors. Arlitt and Williamson proposed a series of metrics to describe the aggregate web traffic (Arlitt & Williamson, 1996). Crovella and Bestavros discussed the self-similarity of Web access pattern (Crovella & Bestavros, 1997). Almeida et al. proposed a hierarchical approach for workload characterization of the request generated by robots at the different timescale. Characterization was done at the session, function and request levels (Almeida et al., 2001). Yu et al. classified a particular web site’s visitors into different groups according to their purchase habits (Yu, Ou, Zhang, & Zhang, 2005). Li et al. performed cluster based analysis to classify a large number of sessions into several coherent classes that efficiently describe web server workloads (Li, Goševa-Popstojanova, & Ross, 2007). The inclusive study of human visitors induced web traffic leads to design for various solutions such as effective and efficient website design (Spiliopoulou, 2000), optimal cache replacement policies (Shyu, Haruechayiasak, & Chen, 2006) (Almeida et al., 2001) according to the variations in human navigation patterns (White & Drucker, 2007) (Lin, Quan, & Wu, 2008). Dikaiakos et al. statistically investigated the various properties of crawler-induced HTTP traffic. This study includes the distribution of HTTP requests and reply-codes, the type and size of resources sought and retrieved, the distribution of crawler requests across a Website, the frequency, and pattern of crawler re-visits, and the temporal characteristics of crawler activity (Dikaiakos, Stassopoulou, & Papageorgiou, 2005). Lee et al. carried out an empirical study on a very large data to classify various robots by their access pattern (Lee, Cha, Lee, & Lee, 2009). Huntington et al. studied robots pattern in scholarly information environment (Huntington, Nicholas, & Jamali, 2008). Jacob et al. presented the inherent access pattern of crawlers on online social network sites (OSNS) (Jacob, Kirda, Kruegel, & Vigna, 2012). Doran et al. performed a comparative analysis between web robot sessions and humans...