Anti-scraping Tool: A system that blocks automatic scrapers through Computer Test and Human Test

Paul Alejandro, Algene de la Paz, John Guevara, John Ong David, Alexis Pantola
Computer Technology Department, De La Salle University

Introduction

- E-commerce is defined as the dealings and transactions that occur over the Internet [1]. The availability of the products information along with its currency value over the Internet has become the target of competing companies. Competitors use a method website scraping where they extract the information from a website for data manipulation [2]. Automatic scrapers called bots run program that automatically harvest huge amount of information at a rapid rate. They also affect the performance of the website because of performing requests at a very fast rate. In order to counter scraping tools, anti-scraping tools and techniques were developed.

Sentinel Anti-scraping tool

- Sentinel is an anti-scraping tool that does not solely depend on predetermined IP addresses. The system also diminishes the occurrences of scrapers attacking with new authentic IP addresses. This paper discusses three of the five main modules of the system. The first module, called the Rate Limiter Module, is responsible for limiting the requests coming from the users and checking if the speed of the request is suspicious. The second and third modules are Computer Test Provider Module and Computer Test Checker Module respectively. These modules are responsible for verifying if the user is an automatic scraper through a reverse CAPTCHA called HoneyPot CAPTCHA, a CAPTCHA that is hidden from legitimate users.

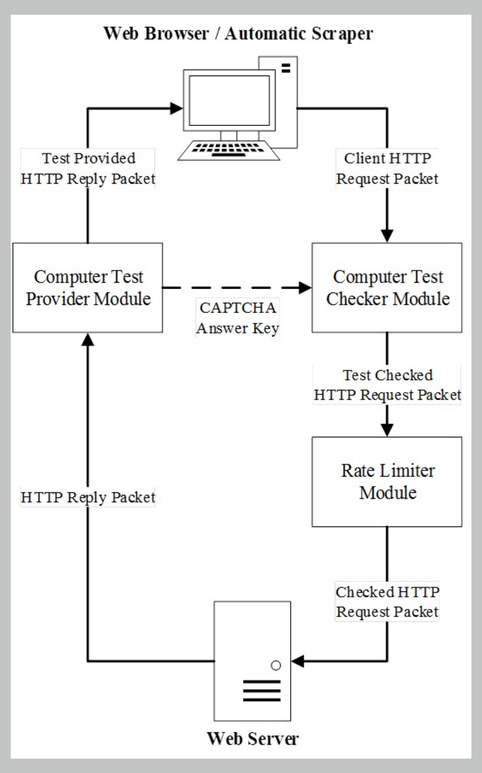

Figure 1: Architectural Design

Rate Limiter Module

- This module limits the speed a user is allowed to send HTTP Requests. The Token Bucket method is an algorithm that can be used in any language, but in this case the module uses Java. This method is executed by giving each user a bucket with a set amount of tokens inside. The user can then spend a token to send a request, but the system gives the users a set amount of tokens every so often, to avoid the depletion of the supply. If the bucket of a user should become empty, then the said user is unable to send requests until the system refills his/her bucket.

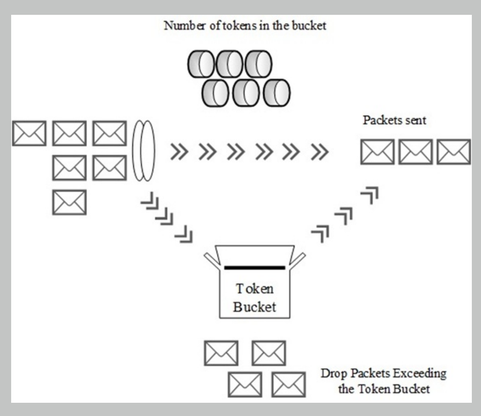

Figure 2: Token Bucket Algorithm [3]

Computer Test Provider and Computer Test Checker Module

- These modules are in charge of inserting tests into HTTP Reply Packets to verify a users authenticity and checking if the tests are answered correctly. The Computer Test Provider Module inserts an additional test depending on the level of the user involved, and the Computer Test Checker Module labels users as scrapers once a specific test is failed. The tests implemented in these modules are CAPTCHAs.

Results: Table

- Results of Rate Limiter Only (RLO)

Request/Sec	Result: RLO	Average No. Replies
10	Detected	12.8
9	Detected	12
8	Detected	12
7	Detected	12.6
6	Detected	11.6
5	Detected	13.8
4	Detected	14.6
3	Detected	17.4
2	Detected	21.8
1	Not Detected	100*

Table 1: Test Result: RLO (* Note: 100 since it scrapes every web page)

- Results of Sentinel

Request/Sec	Result: Sentinel	Average No. Replies
10	Detected	1
9	Detected	1
8	Detected	1
7	Detected	1
6	Detected	1
5	Detected	1
4	Detected	1
3	Detected	1
2	Detected	1
1	Detected	1

Table 2: Test Result: Sentinel

Results: Figure

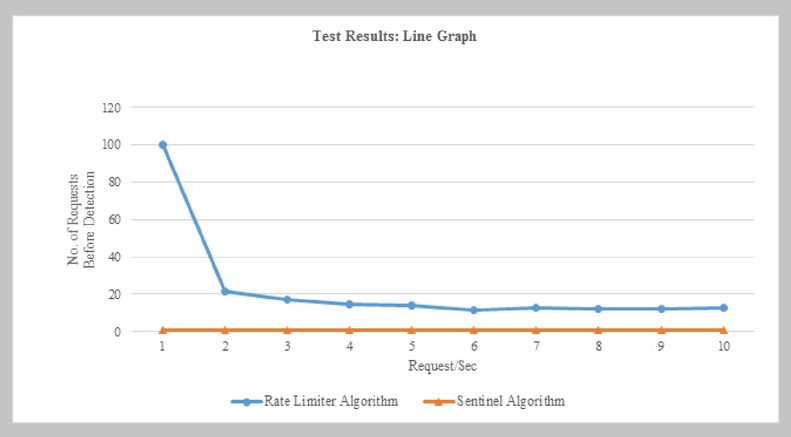

Figure 3: Figure caption

Conclusion

- Using only rate limiting requires a specified speed limit to be broken, which leads to false negatives once scraping tools are able to abide by the speed limit. Sentinel have the Computer Test Provider and the Computer Test Checker which both use the Honey Pot CAPTCHA for verifying a users authenticity and this has proven to be more effective in the detection of scraping tools. This approach was able to successfully detect the scraping tools regardless of the number of requests per second. Furthermore, the approach also keeps the number of scraped web pages to a minimum compared to only using the rate limiter method.

References

[1] What is electronic commerce?, Webopedia, [online] n.d., http://www.webopedia.com/TERM/E/electronic_commerce.html (Accessed: 28 January 2013).
[2] What is Web Scraping?, Webopedia, [online] n.d., http://www.webopedia.com/TERM/W/Web_Scraping.html (Accessed: 28 January 2013).
[3] Qos Introdcution, H3C, [online] n.d., http://www.h3c.com/portal/res/200705/31/20070531_107793_image004_195599_57_0.jpg (Accessed: 30 October 2013).