The Emerging Threats of Web Scrapping to Web Applications Security and Their Defense Mechanism

Rizwan Ur Rahman
Maulana Azad National Institute of Technology, Bhopal, India

Danish Wadhwa
JayPee University of Information Technology, Solan, India

Aakash Bali
JayPee University of Information Technology, Solan, India

Deepak Singh Tomar
Maulana Azad National Institute of Technology, Bhopal, India

INTRODUCTION

Nowadays the Internet is at its peak everything is available online or is going to be available soon. So the Internet has provided us with some facilities like online Shopping, Bookings of trains and buses tickets, education and many more. However, there is always another side to the coin, with the facilities comes the cyber attacks of various types like DDOS, Man in the Middle, SQL Injection etc. One of them is Web scraping which is a very serious issue nowadays it's affecting the market of online e-commerce at very great extent.

It is an ongoing threat that aims to take sensitive data from a victim or from web applications. According to the Automated Threat Handbook for Web Applications published by the Open Web Application Security Project, web scraping is exploited at companies in industries including education, financial institutions, government agencies, hospitals, and retail (Munzert et al., 2014).

The main objective of this chapter is to scrutinize to what degree web scraping can cause a threat to web application Security. In first section the terms in the chapter are defined, and an adequate overview of web scraping in context to web application security is presented in order to provide the reader with an understanding of the background for the remaining sections.

The next section examines the classification of web scraping such as content scraping, web scraping, price scraping, data aggregation, database scraping in general and reviews the most widely used scraping tools such as Visual Web Ripper, Web Content Extractor, Mozanda Web Scrapper and Screen Scrapper.

A section dedicated to Defense Mechanism including detective and preventive mechanisms are presented. Subsequently, the aim of this chapter is to provide review of vulnerabilities, threats of web scraping associated with web application applications and effective measures to counter them.

DOI: 10.4018/978-1-5225-9715-5.ch053

WEB SCRAPING

4

Web scraping is also known by some other names like web harvesting and web data extraction basically is used for extraction of data from the websites on the WORLD WIDE WEB. In other words, it can be defined as the process consisting of the extraction and combination of content gathered from the web in a systematic manner (Vargiu & Urru, 2012).

Software applications are available for doing the web scrapping which may do their work of accessing the World Wide Web using Hypertext Transfer Protocol or web browser. Web scraping can also be done manually by the user but is preferably done in an automated fashion implemented using a bot or web crawler. In this, some software also known as web robot is mimicking the browsing between the web and the human in a conventional web traversal.

This robot may gather the data from as many websites as needed and the parsing of the contents is done to easily find and fetch the data required and stores them in the structures as desired.

Generally, this task of web scraping is somewhat similar to copying; in this particular data is collected and copied from the Internet into some manageable and readable storage structure like some spreadsheets or databases.

In this process, the web page is downloaded or fetched (it happens whenever the browser opens up some pages) first and saved for later use and then the data is extracted from it. Hence we can say that web crawling is an important component of the process.

At the second step of the process the content present in the page is parsed, searched or some type of reformatting is done to understand the content for the data to get it inserted into the spreadsheets or database by copying. Generally, the web scrapping software may sometime take a part of the page which can be useful for the authority for some other purpose.

Web Scrapping is being used in various things in today's life like in advertisements and marketing generally by contact scraping and also an important part of the application made for data mining and web mining, and sometimes used to do some price comparisons, for online price change monitoring, weather data monitoring, research and for providing a service to the user where the content comprises of more than one source also known as web mashup for instance, like trivago and mybestprice applications.

Basically, these web scrapers are APIs which are used to extract data from a web page or a website present on the Internet. Also, some big companies like Amazon Web Services and Google provide web scrapping tools free of cost to end users.

Nowadays a new form has been also used for web scrapping which consists of listening or monitoring the data feed from the web servers. And also some web scraping systems are also using DOM parsing techniques, computer vision and NLP to simulate human browsing as to pass the checks for bots that some websites are using to prevent web scrapping.

Types Web Scraping

There are various types of web scraping. This section explores the categories of widely used web scrapping techniques. These are Data Scraping, Content Scraping, Price Scraping, Database Scraping, News Scraping, Article Scraping and Email Harvesting