2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON) | 978-1-6654-7512-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEMECON56962.2023.10092327

Solution to Web Scraping

Chandan Biswas
Master of Computer Application
University of Engineering and Management, Jaipur
Jaipur, India
cbiswas790@gmail.com

Rahul Mallick
Master of Computer Application
University of Engineering and Management, Jaipur
Jaipur, India
rahulrkmallick@gmail.com

Subrata Paul
Master of Computer Application
University of Engineering and Management, Jaipur
Jaipur, India
subratak389@gmail.com

Prof. Dipta Mukherjee
Dept of Computer Science & Engineering
University of Engineering and Management, Jaipur
Jaipur, India
dipta.mukherjee@uem.edu.in

Abstract— Data retrieval from a website, frequently automatically and without the owner's consent, is known as scraping. This information can also be utilized however thescraper sees fit. The action is considered criminal, but the legality has changed because hasn't stopped others from following suit. anti-scraping instruments are not used though. Anti-scraping Solutions are offered as fairly expensive services that are both slow, and effective services. This essay list offers suggestions for reducing impediments. the development of an anti-scraping System as a Product app for small-to medium-sized websites. are expensive.

Keyword - Scraping, Web scraping, Anti-web scraping.

I. INTRODUCTION

Website scraping is the practice of periodically (which might be regular, completely arbitrary, or somewhat arbitrary within a range) obtaining data from a certain website, and is typically carried out by robots. Human participation is often restricted to choosing the time interval and, sporadically, creating bot-detecting countermeasures and web analysis methods, in the event that the target is concerned about security [1].

Medium-sized and smaller websites are more vulnerable to scraping because of construction, scraping bots are simple to create, although expensive anti-scraping tools aren't used. In this field, scraping mostly entails acquiring competitor data in order to increase one's own volume of data and, as a result, the number of customers to one's own website. There have been several court decisions and precedents established.

web scrapers can harm us by theft our important/personal data from our website using web scraping techniques. So, we have to stop/block those web scrapers from our website. We have created an anti-web scraping technique by using this anti-web scraping technique we can protect websites from web scrapers.

At this time when everything is based on the Internet, the possibility of stealing people's important data is very high. Hackers, third-party organizations, etc. are always ready to steal data from websites without the owner's permission (scrap websites) [2]. Anti-web scraping techniques/Anti-web scraping programs are very helpful to protect your data from web scrapers.

II. WEB SCRAPING

The unauthorized process of collecting data from web pages is called web scraping.

Web scraping was initially employed by financial analysts to forecast stock market patterns, but other businesses might benefit from this method of data acquisition. Since this is an autonomous procedure, businesses can easily collect data and concentrate their efforts on data analysis and business strategy development [3].

Anti-web scraping techniques/Anti-web scraping programs are very helpful to protect your data from web scrapers [2].

Key Components of Web Scraping When a scraper gathers and extracts and transforms data into a format to use [3].

Step 1: Sending HTTP request to the server [3].

Step 2: parse and extract the website code [3].

Step 3: Download Data and Save [3].

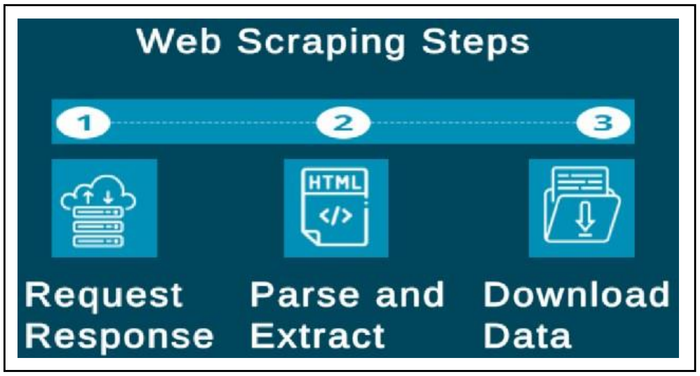


Fig 1 : Web Scraping Steps

978-1-6654-7512-9/23/$31.00 ©2023 IEEE

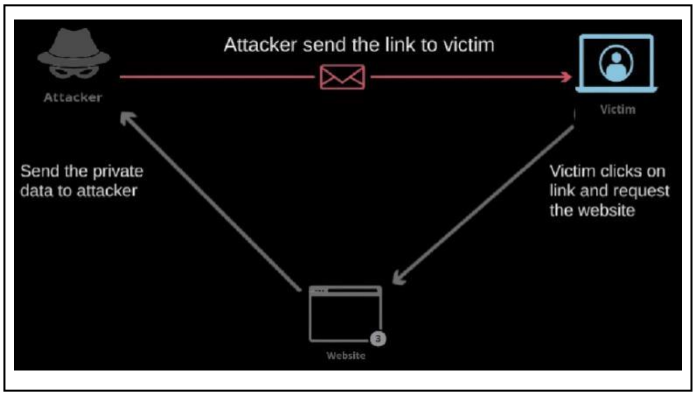

Attacker send the link to victim
Attacker
Victim
Send the private data to attacker
Victim clicks on link and request the website
Website

Fig 2 : How the Attacker Sends the Link to the Victim

III. ANTI-WEB SCRAPING

The technique of protecting websites from web scrapers is called Anti-web scraping. Web scrapers always try to theft data from websites using Scraper bots, so anti-web scraping should be applied to websites to protect those important data [1].

Web scraping bots are frequently blocked using anti-scraping techniques, which prevent the information they collect from being publicly accessed [2].

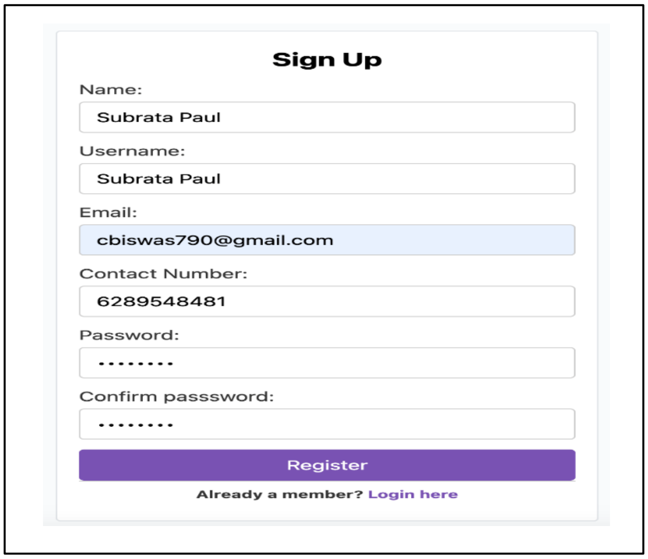
we used some different techniques/methods to achieve anti-web scraping. those techniques/methods are

A. Sign-up and login verification
B. Captcha verification
C. OTP verification

D. Honeypot
E. Web Application Firewall.

A. SIGN-UP AND LOGIN VERIFICATION

sign-up is like registration or ID creation. Using those ids, we can log in and access websites. To sign up (Create an id) you must have to enter your user id like Rahul (Your Name), and password, and confirm the password. and other things are your name, DOB, email, contact number, etc. [4].



Sign Up

Name: Subrata Paul

Username: Subrata Paul

Email: cbiswas790@gmail.com

Contact Number: 6289548481

Password: ••••••••

Confirm passsword: ••••••••

Fig 3 : User Registration

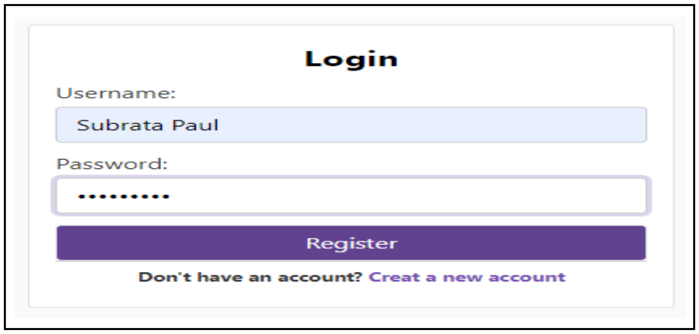
978-1-6654-7512-9/23/$31.00 ©2023 IEEE

A user must submit their identity information into a system during login to gain access to that website. It is a crucial part of our website security protocols.

User names and passwords are always the first two main parts of information needed for login [5].

A user name, often known as an account name, is used to identify a person specifically. User names may be random, identical to or connected to the real names of users or both.

A password is also a string, but unlike a user name, it is meant to be kept private and known only by the user and maybe the system administrator [6].



Login

Username: Subrata Paul

Password: •••••••••

Fig 4 : User Login

B. CAPTCHA VERIFICATION

It is a well-liked method of protecting data against online scraping. In this instance, entering captcha text is required for access to the website. The main drawback of this approach is the inconvenience it causes to normal users who are made to type captchas. As a result, it applies most commonly to systems where data is accessed sparingly and only in response to specific requests [7].

Using captcha recognition software and services, captcha may be avoided. When compared to a one-time payment when software is acquired, the human-based alternative is typically more effective, however, in this situation, money is made for each captcha that is recognized [8].

A captcha is an acronym for Fully Automated Public Turing Machine to Identify Computers and Humans Apart. Public automated software can tell whether a user is a robot or a human [7]. This application would present a variety of problems, including distorted visuals, fill-in-the-blanks, and even equations that supposedly only human beings can solve.

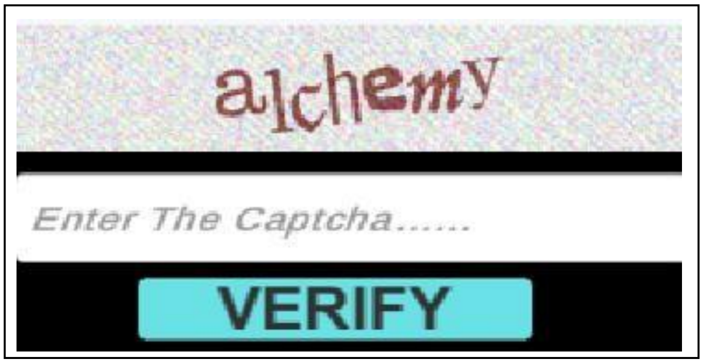


Fig 5 : Captcha Verification

C. OTP VERIFICATION

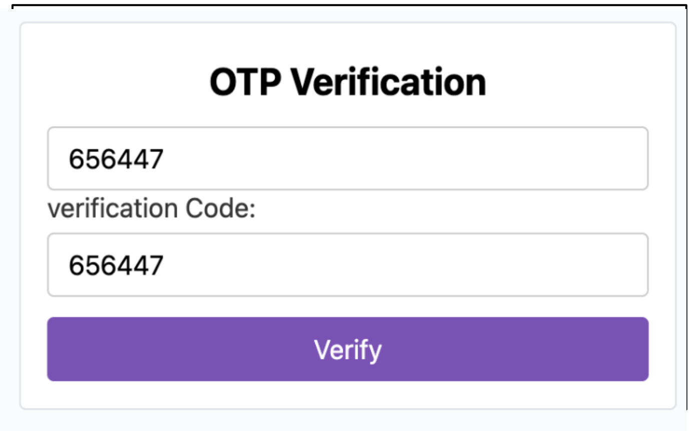
A Time Code (One Time password), or OTP, is indeed a string of letters or numbers that are created by a computer program and intended to be used only once for logging in. One-Time Passwords will reduce the possibility of malicious login attempts and, consequently reduce the possibility of data theft. In OTP verification the computer-generated string always sendsin your email, mobile number, etc. Then only the user can access websites using that OTP [9].

When a person (user/visitor) wants to access the websites then the person has to go through the OTP verification techniques.

In OTP verification always a newly computer-generated string sends to your email, mobile number, etc. Only then, the user can access websites using that OTP [10].

Every OTP always has a time period, the person has to use the OTP at that particular time, Otherwise, if the OTP expires then the person can't log in/access the websites using that expired OTP. After that, if the person wants to try again to log in/access the website, then the person has to request the website to send a new OTP (Resend OTP) [10].

Then the person will get a new OTP to log in/access the website.



OTP Verification

656447

verification Code:

656447

Fig 6 : OTP Verification

D. HONEYPOT

A security measure known as a honeypot sets up virtual traps to entice intruders. Attackers can take advantage of a computer system's flaws that have been purposefully compromised, allowing you to research them and strengthen your security measures.

A honeypot is a technique used to detect and trap bad requests on a website. The majority of these spam attempts target websites by way of spam form submissions or searches for security holes. They frequently jam into it by filling up the fields on our website's contact page, entry form, or product inquiry form, leaving us with possibly tens of thousands of spam submissions each day. As a result, the foundation of my idea is a honey-pot-based spam defense system. A field will be placed on the registration form as a security measure that users cannot see because of CSS and Js (hide the trap's position). The honeypot system recognizes and traps the bots when they scan the code and fill inside the secret gap. In our project, we used email as a hidden field and more sophisticated hiding strategies. In a sense, the goal of my research is to use cutting-edge hidden techniques to detect, trap, block, and generate a list of spam messages, making the honeypot an advanced honeypot. An agile model was employed as the methodology. The most effective and popular methodology for website development is agile. My project's desired outcome is for my honey-pot system to be able to recognize, seize, and stop spam bots in order to keep them from accessing my website. Additionally, to create a list of spam emails that anti-spam organizations can use to confirm spam emails. In a word, the goal of this project is to locate spam bots, catch them, and prevent them from accessing my website. In addition, using the information about the spam bots that were captured, a list can be used as an awareness as well as a verification list for regular users is also created. [11].

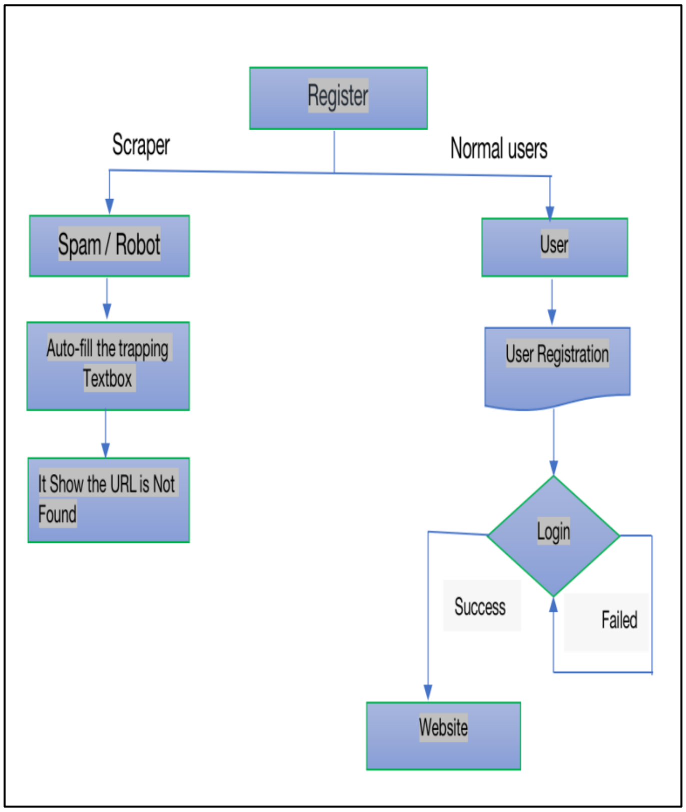


Fig 7 : Applying Honeypot

E. WEB APPLICATION FIREWALL.

By filtering and keeping track of web traffic between a website and the Internet, a WAF (web application firewall) aids in the protection of web applications. A firewall that is specially made to manage "web" traffic is known as a web application firewall. A web application firewall's job is to examine all HTTP traffic going to a web server, filter out any "bad" requests, and forward any "good" traffic.

We need to protect your web server and its content from cyber-attacks. Like: - Cross-Site Scripting, Layer 7 Dos attacks, Web Scraping, Third-party, etc.

A web application firewall guards against harmful HTTP/S traffic entering and exiting your website by filtering, analyzing, and blocking it. It also stops any unauthorized data from leaving the app. It accomplishes this by abiding by a set of guidelines that assist distinguish between safe and malicious communications. Every HTTP/S request on the application

978-1-6654-7512-9/23/$31.00 ©2023 IEEE

level is examined by a web application firewall, which safeguards the application layer. Web application firewall can be thought of as the bridge that connects the user and the web app, filtering all communications before they reach the user or the top app.

Some web application firewalls: -
Cloudflare
App Trana
Akamai
Citrix
F5 Advanced
SiteLock
Sucuri Website Firewall

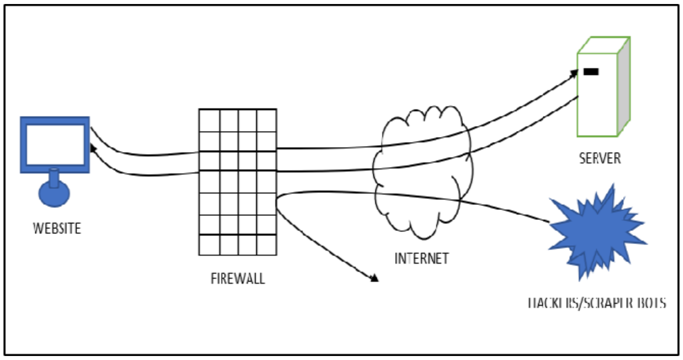


Fig 8 : Applying Firewall

- Cloudflare Firewall

We have used the Cloudflare firewall on our website.

a cloud firewall is a protection device that filters out potentially harmful network traffic. Cloud firewalls were hosted on the cloud, as opposed to conventional firewalls. Firewall-as-a-Service is another name for this firewall delivery model that utilizes the cloud (FWaaS).

Like traditional firewalls do for an organization's internal network, cloud-based firewalls provide a virtual wall across cloud platforms, architecture, and applications. On-site infrastructure can be protected by cloud firewalls as well.

Cloudflare web application firewall (WAF) guards against zero-day threats, such as SQL Injection and Cross-Site Scripting (XSS), on your website.

Risks and vulnerabilities affecting the application layer as recognized by OWASP.

Customers include the Top 50 Alexa-ranked websites, banks, e-commerce businesses, and large corporations. Our WAF prevents millions of assaults per day and is completely integrated into our DDoS defense. It automatically learns from every new threat. By using a backend server to duplicate and cache websites, Cloudflare serves as an intermediate between such a server and a client. It can reduce loading times by saving web content for transmission on the nearest edge server. It can also alter information, like graphics and rich text, in order to function better. A level of security filtration is also provided by Cloudflare using this intermediary approach. It can block spam and bot traffic, stop distributed denial-of-service attacks, intercept bot attacks, and detect harmful communication by being in between the user and the hosting server.

- ❖ FLOWCHART OF ANTI-WEB SCRAPING TECHNIQUE

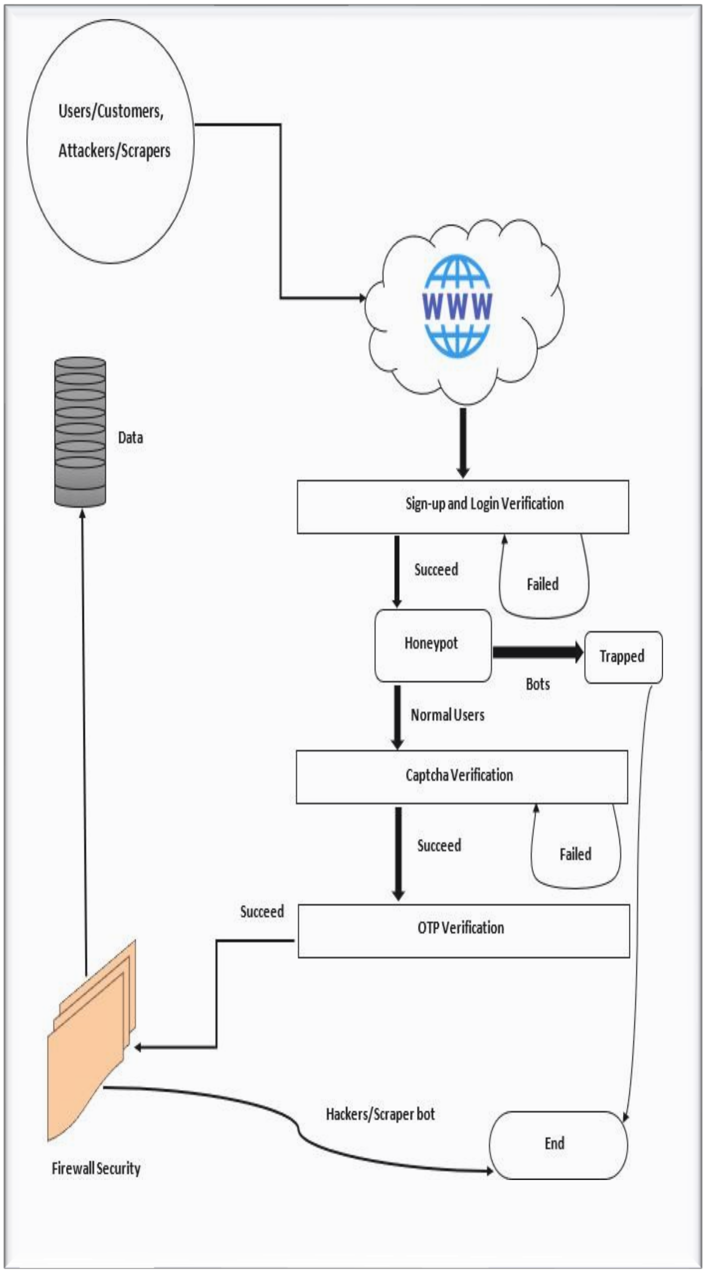


Fig 9 : Working Principle of the Anti-Scraping Technique

IV. FUTURE SCOPE

In the future, we will turn this anti-scraping technique into anti-scraping software. In that anti-scraping software, we will use better Firewalls, a Login system, Captcha, OTP verification, Email verification, honeypot, IP tracing, etc. And we will also try to stop different illegal requests of web servers.

V. CONCLUSION

In this era when everything is based on the internet, our important and personal data also remains on the internet so this data can be stolen by scrappers (hackers) and those scrappers (hackers) can harm us using those data. To get out of this problem we have to do something to block those scrappers (hackers) and save our important and personal data from those scrappers (hackers).

we have created a website that has an anti-web scraping technique. In this anti-web scraping technique, we have used some modules like- Sign-up and login verification, Captcha verification, OTP verification, Honeypot, and Web Application Firewall.

By using this anti-web scraping technique, we can block/avoid scrapers (hackers) and protect our important and personal data from those scrapers (hackers).

Since scraper will try to steal data from the website by developing new methods, again and again, we also need to develop new methods to prevent theft.

REFERENCES

[1] Haque, Afzalul, and Sanjay Singh. "Anti-scraping application development." 2015 international conference on advances in computing, communications and informatics (ICACCI). IEEE, 2015.

[2] Parikh, Kaushal, et al. "Detection of web scraping using machine learning." Open access international journal of Science and Engineering (2018): 114-118.

[3] Sirisuriya, De S. "A comparative study on web scraping." (2015).

[4] Lemon, Tatiana, et al. "Development of a Student Work-Hour Verification Database for the Pace University Center for Community Action and Research."

[5] Sun, San-Tsai, et al. "What makes users refuse websingle sign-on? An empirical investigation of OpenID." Proceedings of the seventh symposium onusable privacy and security. 2011.

[6] Gamboa, H., A. L. N. Fred, and A. K. Jain. "Webbiometrics: User verification via web interaction." 2007 Biometrics Symposium. IEEE, 2007.

[7] Mehra, Mahendra, et al. "Mitigating denial of service attack using CAPTCHA mechanism." Proceedings ofthe International Conference & Workshop on Emerging Trends in Technology. 2011.

[8] Cui, Jing-Song, et al. "A CAPTCHA implementation based on moving objects recognition problem." 2010 International Conference on E-Business and E-Government. IEEE, 2010.

[9] Kurniawan, Dwi Ely, et al. "Login Security Using One Time Password (OTP) Application with Encryption Algorithm Performance." Journal of Physics: Conference Series. Vol. 1783. No. 1. IOP Publishing,2021.

[10] Singh, B., K. Sh Ranjan, and D. Aggarwal. "Smart voting web-based application using face recognition,Aadhar and OTP verification." International Journal of Research in Industrial Engineering 9.3 (2020): 260- 270.

[11] MarchMairh, Abhishek, et al. "Honeypot in network security: a survey." Proceedings of the 2011 international conference on communication, computing & security. 2011.

[12] Muzammil, Akanksha Chaudhary, and Rohit Nandan. "Comparative Analysis of Packet Filtering Firewall."

[13] Bhamra, Satnam Singh. "The 2010 Personal Firewall Robustness Evaluation." (2010).

978-1-6654-7512-9/23/$31.00 ©2023 IEEE