International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Prevent Web Scrapping

[1]*Yash Shelar*, [2]*Ajay Gupta*, [3]*Asst. Prof. Sonali Patil*

[1,2] PG Student, [3]Guide
Keraleeya Samajam's Model College, Dombivli East, Mumbai, Maharashtra, India, Yashshelar333@gmail.com

ABSTRACT

Web Scraping is an automated bot threat where cybercriminals collect data from your website for malicious purposes, such as content reselling, price undercutting, etc.

In this article, we look at how scraping attacks are used to take advantage of online retailers, who is carrying out web scraping attacks and why, how scraping attacks unfold, what web scraping tools are used, common protection tactics against web scraping, and in what ways datadome protects against content scraping and all other automated OWASP threats.

Content scraping is when bots and sometimes humans steal content for legit research and even nefarious reasons. The content is "scraped" using a bot, but in small cases, it may be taken using a manual method. The reason is to produce a duplicate copy of your website that can hurt your reputation as the duplicate website has your content, but has malware or sells bogus products.

That being said, there are data scientists and researchers that will scrape content without malicious intent. However, content scraping has been used a lot for bad reasons.

Content scraping is also known as data scraping. It is the same, as you use a form that tells a bot what content to search for and steal. In fact, some people don't even need to know code to do content or data scraping. They can go to places like Octoparse.com to mine data like stats from sports. It's as simple as telling the bot what site you want to scrape, and then specifying the particular data you wish to gather.

Main types of data scraping

Businesses face various types of data extraction, each with its own unique characteristics and potential risks:

Content scraping copies website content without permission, including text, images, videos, and other types of digital material. This type of scraping can include email addresses, phone numbers, and social media handles. This form of fraud is often used for email marketing, lead generation, and customer outreach. However, it can also be used for malicious purposes, such as identity theft, fraud, and spamming.

Businesses may also face the theft of personal data, where personal information such as names, addresses, and financial data is stolen from websites. This type of data scraping can result in a significant data breach, which can lead to reputational damage, regulatory fines, and legal action.

Website scraping uses the html code for the main article for web scraping. This language is designed for human end-users and not for ease of automated use. A scraper bot sends an http GET request to a specific website. The website responds and the scraper parses the html document for a specific pattern of data. Once the data is extracted, it is converted into whatever format the scraper bot's author designed.

Javascript is a powerful and fast method for extracting text and links from html pages. It is also used to target a linear or nested html page. Website scraping involves the use of javascript to retrieve content from a web page and process it into a data file.

Price scraping involves a data gathering technique that extracts pricing information from various e-commerce websites. This technique is commonly used by competitors to gain a competitive edge by analyzing market trends and pricing strategies. Additionally, customers can also use price scraping tools to find the best deals and compare prices across different platforms.

Even so, businesses may face negative consequences from price scraping as it can lead to reduced revenue and profit margins. This is because price scraping allows competitors to easily monitor the pricing strategies of their rivals, leading to intense price competition. Further, the excessive use of price scraping can result in server overloads and other technical issues that can negatively impact website performance.

Screen scraping refers to the technique of extracting data from a visual output generated by a software application. This method involves using specialized software that can interpret the graphical interface of an application and extract the relevant data. Screen scraping software has become increasingly popular due to its ability to automate data extraction from a wide range of applications, including web browsers, desktop applications, and mobile apps.

With screen scraping tools, businesses can quickly and easily extract data from various sources, which can be used for analysis, reporting, and decision-making purposes. Additionally, screen scraping can help organizations to reduce manual data entry tasks and improve data accuracy by eliminating human errors. However, it is important to note that this form of extraction may raise legal and ethical concerns, particularly when it involves extracting data from third-party websites without permission.

The dark side of data scraping

Data scraping or web scraping is an essential tool for organizations to access and collect data from the internet—but it also has a dark side that poses serious risks to businesses and their customers. While data scraping is used for a variety of purposes, from market research to price monitoring to lead generation, it can also be used to commit all sorts of cybercrime, from identity theft, fraud, and other criminal activities. For example, scraped personal information can be used to create fake accounts, apply for loans or credit cards, or launch phishing attacks.

Scraped intellectual property, such as trade secrets and copyrighted materials, is often used to gain a competitive advantage or generate illicit profits. To make matters worse, scrapers can use advanced tools to access confidential data stored on company servers, giving them access to sensitive information such as financial records and client data.

Regulations such as the EU General Data Protection Regulation and the California Consumer Privacy Act provide organizations with guidance on how to properly access and use data from the internet. Additionally, organizations should ensure that any data scraped is used responsibly and does not violate any privacy laws.

How to protect your business from malicious website scraping :

With the state of the digital environment and cybercrime in recent years, unfortunately, 100% prevention of malicious web scraping from targeting your website is virtually impossible.

Yet, we can still be proactive to make it as difficult as possible for bad actors to perform web scraping on our website by strengthening three key aspects:

Establishing a solid foundation on the website by optimizing structure, ensuring frequent updates, etc.

Since web scraping is typically performed by malicious bots, a strong malicious bot detection and management solution should be implemented.

Establishing automated monitoring for duplicated content and brand infringement and ensuring these pages get reported and taken down quickly.

Below, we’ll discuss them one by one.

Establishing a strong security foundation

A key objective when aiming to prevent web scraping is to make it as challenging as possible for the web scraper bot to access and extract your data, letting the bot waste its resources and discourage the perpetrator from targeting your site. Here are some tips you can use:

Don’t expose your dataset: don’t provide a way for these bots to access all of your datasets at once. For example, don’t publish a page listing all your blog posts, so users would need to use the search feature to find any post. Also, make sure *not* to expose any endpoints and APIs. Obfuscate and encrypt all your endpoints at all times.

- **Honeypotting:** another effective technique to protect your content is to add ‘honeypots’ to the HTML codes of your page’s content to trick the web scraper bots. A basic honeypotting technique is to add codes or content that would be invisible to human users but can be parsed by bots; then, you can make it so that when the bot clicks on this code, it will be redirected to a fake page serving thin/fake content to poison its data.
- **Frequent update and modification:** an especially effective technique to protect your website from HTML parsers and scrapers is to frequently and intentionally change your HTML codes and patterns. This approach is especially important if your site has a collection of similar content that may naturally form HTML patterns (i.e., regularly updated blog posts.)

Monitoring and mitigating web scraper bot activities

The second foundation to preventing web scraping attacks from affecting your business is to monitor and mitigate activities from malicious web scraper bots.

You can either check your traffic logs manually (i.e., via Google Analytics) and try to identify signs of malicious bot activities, including:

Users who are very fast when filling and submitting forms

Robotic mouse-click and keyboard stroke patterns

Linear mouse movements

Repeated similar requests from the same IP address (or a group of IP addresses)

Inconsistent JavaScript signatures like different time zones, inconsistent screen resolutions, etc.

Alternatively, you can invest in advanced bot detection and mitigation solutions that will automatically detect the presence of web scraper bots in real-time and on autopilot (won't need human supervision and intervention.)

Once you've identified the presence of web scraper bots, there are several options you can try to mitigate their activities:

- **Challenge:** challenge the bot with a CAPTCHA or other challenge-based mitigation approaches. However, keep in mind that using too many CAPTCHAs may affect your site's overall user experience. With the presence of CAPTCHA farm services in recent years, challenge-based bot mitigation techniques have also been rendered relatively ineffective.
- **Rate-limiting:** limiting resources served to these identified bots, for example, only allowing a limited number of searches per second from a single IP address. Bots run on resources, and the idea is that slowing them down and letting them waste their resources may discourage the bot operator from continuing to target your website.
- **Blocking:** In cases where you are absolutely certain that the traffic originates from a malicious bot, you can consider blocking its access to your website. However, it's crucial to acknowledge that this might not always be the most effective approach. In some instances, determined attackers can adapt by modifying the bot, potentially circumventing your existing defenses like PerimeterX, which employs a sophisticated anti-bot system to block website access. Persistent attackers may find ways to bypass PerimeterX, resulting in a bot that returns even more sophisticated than before. The choice of the most suitable mitigation strategy should be based on a thorough assessment of your specific circumstances, considering factors such as the potential impact on user experience and the adaptive nature of bot attacks.

Monitoring, reporting, and taking down fake websites publishing scraped content

The third foundation is to monitor the internet for the existence of your scraped content being published on another URL, then you can take the necessary actions to report and take down this content.

A more effective approach both in terms of accuracy and cost-efficiency is to use a dedicated website scraping and Copyright Infringement Monitoring solution like Red Points'.

Red Points leverages state-of-the-art technology to conduct real-time domain research and monitoring, so it will automatically detect any malicious web scraping attempt, notify you, and automatically take the necessary steps to take down the fake website so you can use your time to focus on your core business tasks instead.

When needed, Red Points' Investigation Services can also collect data that might be used as evidence if you are taking legal action against the individuals or organizations performing the malicious website scraping attempt.

Acknowledgment:

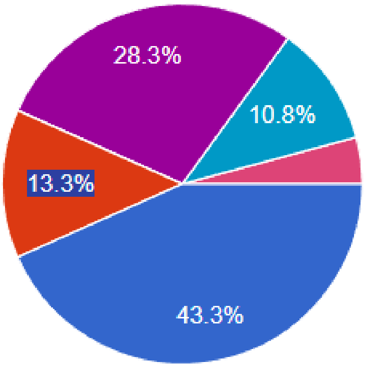
By using this website, you acknowledge and agree that the unauthorized collection of data through automated means can disrupt the normal functioning of the site, compromise user privacy, and violate our terms of service.Website reserves the right to take legal action against individuals or entities involved in such activities.

Thank you for your cooperation and understanding. If you have any questions regarding these terms, please contact us.

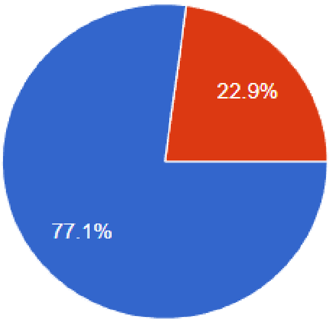
It's important to note that while including such statements in your terms of use may deter some users from attempting to scrape your website, it may not prevent all instances of web scraping. Implementing technical measures such as rate limiting, CAPTCHAs, or other anti-scraping tools can complement these acknowledgments and help protect your website from unwanted data extraction.

Figures and survey result

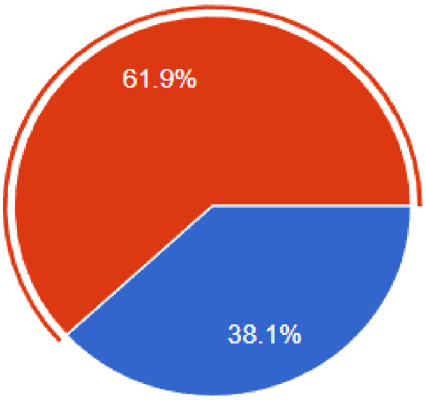
1.Select your age group



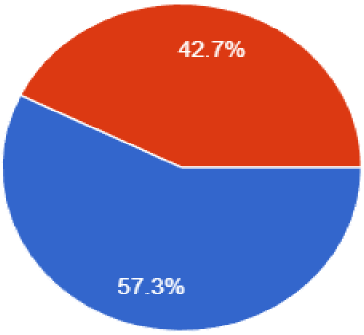
2. how familiar are you with the concept of web scrapping?



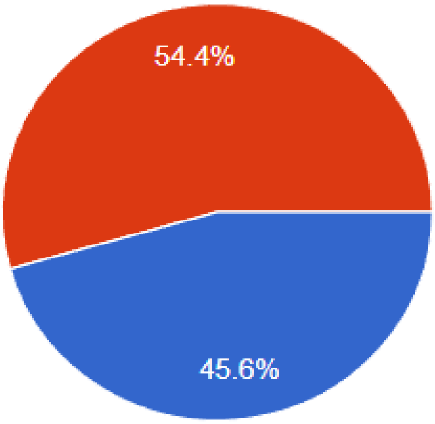
3. have you ever encountered issues related to web scraping while using or managing a website?



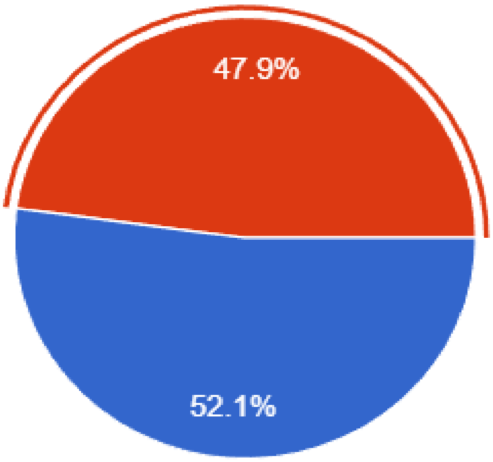
4. do you believe there should be stricter regulation on web scraping activities?



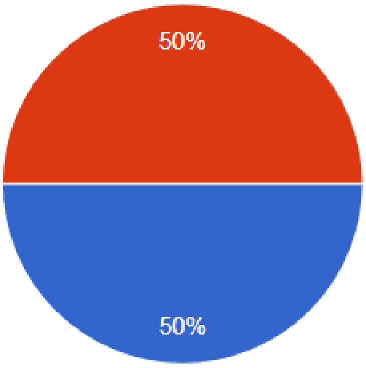
5. how do you feel about the impact of anti-scaraping measures on your user experience when interacting with websites?



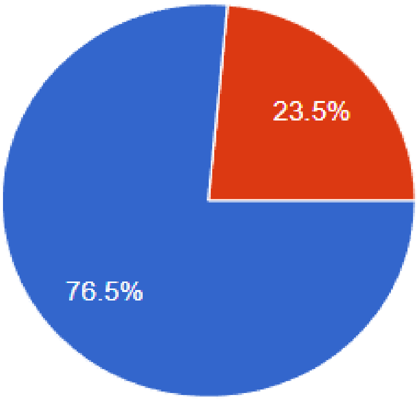
6. are you aware of any emerging technologies or strategies for preventing web scraping?



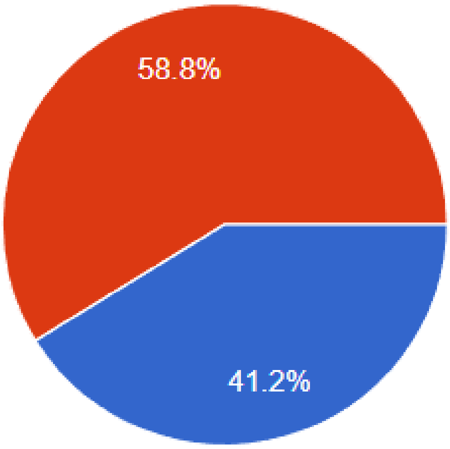
7. In your opinion, what are the primary concerns associated with web scraping?



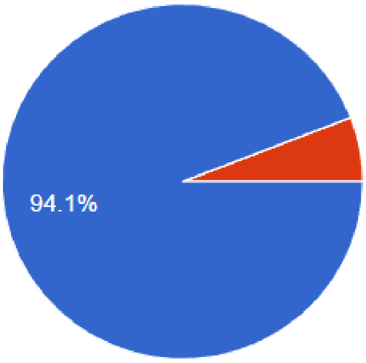
8.Are there specific instances where you found anti-scraping measures intrusive or hindering?



9.**How do you envision the future landscape of web scraping prevention?**



10.From an ethical standpoint, do you believe there are situations where web scraping is justified?



CONCLUSION :

Preventing web scraping is a multifaceted challenge that involves a combination of technological, legal, and ethical considerations. In conclusion, effective measures to prevent web scraping require a thoughtful approach that balances the protection of data, user experience, and adherence to legal and ethical standards. Here are some key points to consider:

- Technological Measures:

The use of technical tools and methods, such as rate limiting, CAPTCHAs, and IP blocking, can contribute to the prevention of web scraping.

Continuous monitoring and updating of anti-scraping measures are essential to stay ahead of evolving scraping techniques.

- User Experience Considerations:

Implementing anti-scraping measures should be done with careful consideration of their impact on the user experience. Striking a balance between data protection and user accessibility is crucial.

Transparent communication with users about the reasons behind anti-scraping measures can help build understanding and trust.

- Legal and Ethical Dimensions:

Legal frameworks around web scraping vary, and it's important for website owners to be aware of the regulations in their jurisdiction.

Ethical considerations should guide the decision-making process, with a focus on respecting user privacy and ensuring fair use of data.

- Challenges and Solutions:

Website owners face challenges in preventing sophisticated scraping attempts, and the landscape of web scraping is constantly evolving.

Collaboration within the industry and sharing best practices can help address challenges and enhance the overall resilience against web scraping.

- User Awareness and Education:

Educating users about the risks and benefits of web scraping, as well as the measures in place to protect their data, can foster a more informed online community.

- Adaptation to Future Trends:

As technology advances, the landscape of web scraping prevention will continue to evolve. Website owners should remain vigilant and adapt their strategies to emerging trends and technologies.

REFERENCES

1.Web Scraping Prevention Tools and Techniques:

- OWASP Anti-Scraping Guide: The Open Web Application Security Project (OWASP) provides a guide on anti-scraping measures and techniques.
- Scrapy: An open-source and collaborative web crawling framework for Python, Scrapy can be used for both scraping and preventing scraping.

- Distil Networks - The Ultimate Guide to Preventing Web Scraping: Distil Networks offers insights and strategies for preventing web scraping in their comprehensive guide.

2.Legal and Ethical Considerations:

Legal and Ethical Implications of Web Scraping: A publication from the Berkman Klein Center for Internet & Society at Harvard University that explores the legal and ethical aspects of web scraping.

Web Scraping and Crawling Are Perfectly Legal, Right?: An article on Lexology that delves into the legal considerations surrounding web scraping.

3.Industry Practices and Insights:

- Preventing Web Scraping: Best Practices for Keeping Your Content Safe: A blog post that discusses best practices for preventing web scraping and safeguarding content.
- Web Scraping: A Practical Guide: A practical guide on web scraping and prevention strategies from Datahut.