WHITEPAPER

SHIELDSQUARE

COMBATING WEB SCRAPING IN ONLINE BUSINESSES

Overview

More than half of the Internet traffic is bot traffic. With the number of Internet users increasing exponentially, there's a significant increase in the number of online businesses, ranging from e-commerce and online content generation, to ticketing and job portals. If majority of the Internet traffic is going to be from non-humans (bots), how can online businesses make sense out of their Web traffic? Most importantly, how can they retain their competitive edge when bots are created with malicious intents, to do Web scraping? To accomplish these, online businesses must understand how vulnerable their websites are to scraping, and how easily data can be extracted. That will set the fundamentals right to opt for the right anti-scraping solution that will give them the flexibility to deal with bad bots efficiently.

What's web scraping?

Web scraping, in general, refers to the extraction of data or information from websites. Price scraping and content scraping are two of the primary forms of Web scraping affecting several online businesses, such as, e-commerce, online media/ publishing, job portals, education content portals, real estate, travel, financial information sites, and so on. In short, online businesses that produce rich, unique, proprietary and time sensitive, content are always under threat from the competition.

Price scraping

Competitors use automated bot programs to scrape the prices off e-commerce and marketplace websites. Competitors resort to this malicious activity to steal the real-time dynamic pricing data, so that they can strategically undercut the product prices and attract those price-sensitive buyers. Bot programs can be written to automatically scrape pricing information from multiple competitor sites compare them and update the portal with the best prices to draw more customers. When potential buyers search for the products, these competitors get highlighted on Google, displaying lesser prices alongside their peers. This creates a psychological 'low price' perception in the minds of the potential buyers. Thus, competitors gain an unfair advantage! In addition to the pricing information, other data such as product listings, related products and SKUs are also scraped from the target website.

Content scraping

Content scraping bots harvest unique content from online media and publishing websites, and post them elsewhere. When a news portal, or any online marketplace, listings and classifieds website for that matter, publishes fresh content, and if the same content is scraped and posted immediately on multiple other places without proper attributions, the website's SEO is negatively impacted. Website owners invest a great deal on getting visibility across search engines, and this fundamental requirement is jeopardized due to content theft. Also, this reduces the brand competitiveness of the portal. Apart from losing unique content, when a large number of bots scrape the website, precious server and bandwidth resources are used up, impacting genuine user experience on the portal

Businesses vulnerable to scraping

Technically, any online business that generates unique content is prone to Web scraping. The site's vulnerability to scraping depends on one or all of these four important factors:

1. Upcoming competition from similar businesses

2. Popularity of the website in terms of traffic or user engagement

3. Uniqueness or criticality of the content that is created/changed

4. Existing website security loopholes

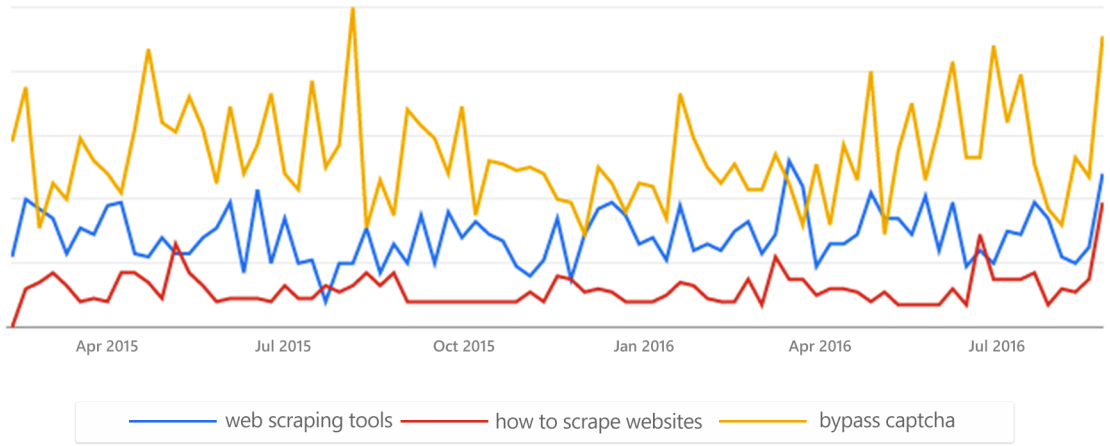
It's important to mention that there are scraper bots created to extract data for research purposes. For example, a bot program can be executed to scrape weather data from multiple websites, structure them and use it for a seemly useful analytical task, like comparing weather information over a period of time from various countries. Or, scrape Fantasy Football projections and analyze risk levels, determine player values, and so on. Having said that, as a business owner, one can't be too sure if you're losing data to the competition or just for educational or research purposes.

Scraping trends and tools

When thousands of bots attack a website (eCommerce, media portals, and so on) to harvest data, they cause multiple other issues like increase in server loads, network bandwidth wastage and most importantly, result in bad user experience.

The most common technologies used for scraping are cURL, Wget, HTTrack, Scrapy, Selenium, Node.js and PhantomJS. Of course, these are just a few, and there are hundreds of Web scraping tools and services to anyone who wants to illegally scrape data from popular websites.

There's a never ending demand for scraping tools, services as seen in this Google Trends depiction.



The search for Web scraping tools, how to scrape websites and bypassing captchas seems to be steady, and even show an upward trend. A search on Twitter for the term scrape websites returns a number of tweets, listing scraping services and tools, along with requirements like what type of website to scrape and in which country. Even Google Sheets[1] can be slyly used by scrapers to steal data from your website. For example, to scrape all the href references contained in the URL: *https://en.wikipedia.org/wiki/Moon_landing*, all you need to do is enter *"=IMPORTXML("https://en.wikipedia.org/wiki/Moon_landing", "//a/@href")"* in a cell within the Google spreadsheet.

1: http://www.cso.com.au/article/605136/manually-detecting-content-theft-hard-when-google-docs-used-scraping/

These types of attacks can be difficult to detect without the right tools, or the know-how. This is because; the data extraction request appears to be coming from a genuine Google IP address. This is just the tip of the iceberg. There may be other genuine tools that may be exploited by the scrapers to get to your data.

Why do scrapers get an unfair advantage?

Establishing an online business involves a lot of costs. Costs for running servers, network infrastructure services, and for marketing, sales, software development, and content teams. There are also a lot of resources spent on developing the business strategy and go-to-market plans. In short, there's a cost for everything when running a genuine online business. Business owners spend anywhere between thousands of dollars to billions in setting up Web hosting services, Web design, IT infrastructure and employees. They spend thousands of dollars to billions via marketing spends to take their business to the market.

On the other hand, for the scraper, it takes a mere $300 to set up virtual machines and proxies to start with data extraction from target websites. It's even quicker and cheaper to use one of the aforementioned scraping tools, alongside cloud-based data center services.

A competitor in the e-commerce space can hire scraping services for just a couple of thousand dollars, scrape the product and pricing information from target websites, and analyze the prices. They can undercut the competition with lesser prices, drawing the maximum number of customers from the target websites. So, at such low costs, the return on investment for the competitor, that is scraping other websites, is fairly huge. Also, a competitor can send bots to add items to the carts, only to be abandoned[2] later. Well, this unfair advantage is not restricted to e-commerce or price scraping; it affects all online industries across verticals.

It's a fundamental requirement that every online business needs the visibility provided by Google search engine rankings. A series of tests done by Intelligent Positioning using their Pi Datametrics software reveal an unsettling truth - the original website content is often outranked[3] on Google, by those sites that have scraped the content from.

2: https://www.shieldsquare.com/this-is-what-happens-when-bots-influence-cart-abandonment-in-ecommerce-infographic/
3: https://searchenginewatch.com/sew/how-to/2426960/is-your-content-working-better-for-someone-else

Types of bots

Here is what you should know about the different types of bots, as classified by ShieldSquare.

Data Center Bots: As the name suggests, these are malicious bots that operate out of a data center (usually, using services from Amazon AWS, Google, etc).

Bad User Agent: When a scraper uses tools like Wget, Scrappy, cURL, and when they identify themselves using the tool name as User Agent, they are classified as Bad User Agents.

Integrity Check Failed: These are the bots that come from normal ISPs and don't use standard scraping tools.

Legitimate Bots: These are friendly bots that crawl your site for news aggregation, social media and market intelligence purposes, and generally help bring traffic.

What are legitimate Bots?

Legitimate bots crawl your website for aggregation, backlink checking, SEO, market intelligence, and more. In short, these bots generally help bring traffic to your site. Some websites do not want certain categories of these legitimate-bots to crawl their portal. It depends on the industry and the use-case to allow/deny select categories of these bots.

Categories:

Monitoring Bots (e.g. Pingdom)- Bots that are used to monitor the system health of the websites

Backlink checker bots (e.g. UASlinkChecker) - Bots that check the backlinks of URLs

Social Network Bots (e.g. Facebook Bot)- Bots that are run by social networking websites

Partner bots (e.g. PayPal IPN) - Partner bots that are useful to the website

Aggregator bots (e.g. WikioFeedBot) - Bots that collate other sites information

So, how do you combat scraping?

Most online businesses do recognize the need for securing website content, and also understand that website security is an integral part of the success of the business. Having said that, the efforts to combat scraping must be headed in the right direction - one that ensures complete protection.

Robots.txt - nothing about protecting content, really

If you're going to start with your robots.txt to stop scraping, stop right there[4].

This robots.txt, an unprotected text file defining search engine crawl, will not protect your content from scrapers. Most search engine spiders/bots (or crawlers) visiting your website will look for your robots.txt file. Based on the configurations mentioned in the file, the crawlers decide what they should and should not index. The Allow setting instructs the search engine bots to crawl your website.

There's a general misconception that specifying Disallow on certain sections of the website will protect the content from scrapers. However, you are just specifying that robots should not visit, and in turn, index those specific URLs. This is detrimental to your SEO, and website visibility on prominent search engines.

In-house bot prevention teams - shooting moving targets

Traditional IP blocking and in-house manual bot detection techniques work for a while, but puts immense pressure on the team to identify and stop scrapers, at the risk of penalizing genuine users. A company keen on stopping bots needs a dedicated in-house team keeping a 24x7 vigil. However, this is easier said than done. The team should be equipped to detect bot patterns, behaviors, bot sources and devices, and then make an informed decision to block the bots.

The challenge here is it's really difficult to ensure zero false positives. Also, the in-house team will have varying and unpredictable workloads, especially when they had to deal with a firefight. The scraper writing bot programs, once they know their bots are being blocked come up with sophisticated methods to bypass the blocking mechanism in place. This is an additional headache to the team that is already overworked, endlessly aiming at moving targets.

Intelligent scrapers, inefficient tools

Unmonitored and uncontrolled load on business critical servers and network resources is an IT engineer's nightmare. For most of the known threats, there are tools like Web Application Firewalls (WAF), to protect the website from SQL Injection, DDoS attacks, cross-site scripting (XSS) and other Web application vulnerabilities.

4: https://www.techinasia.com/talk/robotstxt-secure-website-content

So, WAFs can stop an incoming security attack based on the firewalls rules defined in their Access Control List (ACL). However, WAF solutions lack the adaptability to stop emerging bot threats. They're ineffective in protecting website content from scrapers that try new techniques to evade detection and steal content.

Tip: Bot Detection vs traditional Web Application Firewalls

Automated bot prevention - the way forward

Let's face it. It's a constant battle against the bots, and only a robust bot prevention solution can provide an always-ON automated protection for your website. An automated solution will significantly reduce the in-house team's workload, save thousands of man hours, and be supremely effective in taking the right action against the different types of bots.

When you implement an automated bot prevention solution, make sure you consider the following five parameters:

1. Never block any of your genuine users - It's OK to wait for a few tens of requests before blocking the bad bots, but ensure zero false positives!

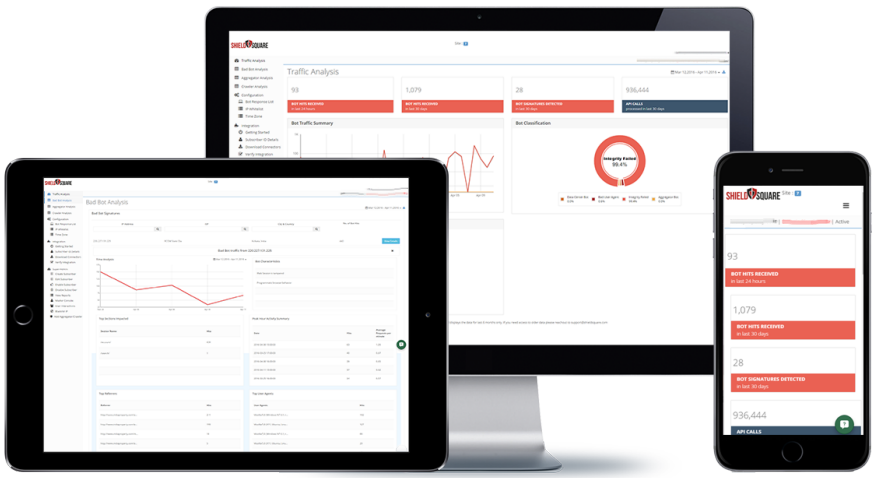
2. Ascertain if the solution can detect and take action on sophisticated bots that mimic human behavior

3. If your website has thousands of pages, but only a few of them are exposed to bot threats, implement bot prevention only where required, not for the entire website

4. See if the anti-bot solution is scalable, and available across geographies, to help with easier global expansion of your business

5. Make sure only a few important parameters (like IP addresses, user agents, etc.,) are passed on to the bot solution provider to identify, categorize and take action on bots. Nothing more!

About ShieldSquare



ShieldSquare is a cloud-based bot prevention solution that helps online businesses differentiate between human & non-human traffic (bots) on websites & mobile applications. ShieldSquare uses cutting-edge technologies to detect, categorize and prevent malicious bots in real-time, and ensuring zero false positives. Processing billions of page requests, ShieldSquare protects businesses spread across 68 countries against the threats of Web scraping, form spam, skewed traffic analytics and many more.

Sign Up for a free, fully functional 15-day trial
Contact Us: +1-646-770-2947 | sales@shieldsquare.com | www.shieldsquare.com

SHIELDSQUARE