The 2016 Economics of Web Scraping Report

The Ubiquity and Danger of Web Scraping
What is Web Scraping

Web scraping is responsible for millions of dollars in lost annual revenue. It is the wholesale theft of website content. On average, 2% of online revenue is lost as a result of web scraping.

Web scraping, or scraping, is a computer software technique of extracting information from the internet, usually transforming unstructured data on the web into structured data that can be stored and analyzed in a central database. The key culprits behind web scraping are bots, a software application that runs automated tasks, or scripts, over the Internet.

Bots typically perform tasks at a much higher and faster rate than humans alone. They make up 61% of web traffic. Because information on the Internet is public domain, web scraping is legal. However, it has a negative effect on content owners.

Today’s cyber security landscape is crawling with sophisticated bots doing the dirty work for hackers, unsavory competitors, and fraudsters. Any content that can be viewed on a webpage can be scraped. There is an entire economy built around web scraping that only looks to grow in the future if malicious bots are not blocked and threats are not mitigated. The number of existing web scraping bots is innumerable because new bots are constantly being created. It is important for website and content owners to have a rough understanding of the economy behind web scraping and how to disrupt the economy in order to protect their content and revenue.

Key Findings

- Content scraping is the leading use for web scraping
- Web scraping services run as low as $3.33 per hour
- The average web scraper makes $58,000 annually
- Real estate sites are the #1 web scraping victims

Uses for Web Scraping

There are 6 main use cases for web scraping: content scraping, research, contact scraping, price comparison, weather data monitoring, and website change detection. Content scraping is stealing original content from a legitimate website and posting it on another website without the knowledge or permission of the original content owner. Content scraping can come in the form of web mashups, using information from more than one source to create a new display of information, also known as web data integration. For example, users can build...
new aggregators, event aggregators, and centralized job portals by using data from other websites. Content scraping is the top use of web scraping by customers, with 38% of companies who web scrape using it for the purpose of content scraping.

The second main use for web scraping is research. 26% of companies that hire web scrapers use web scraping bots to gather research on listening services that monitor consumer opinions about products and companies. Companies also use web scraping bots for mass data collection for various projects. For example, users can get marketing intelligence by using bots to identify key market developments from various sources on the web.

19% of companies use web scraping for contact scraping. These companies wish to gain access to customers’ emails in order to obtain contact information for marketing purposes or for background reports. Bots help generate leads from business directories and social media sites like Twitter and LinkedIn.

Another use for web scraping is online price comparisons between competitors. The real estate and travel industries see a majority of bot activity that is based on price comparisons. From 2014 to 2015, the real estate industry saw a 300% increase in bad bot activity, copying the price comparison model of the travel industry. Simultaneously, travel industry leaders saw that 48% of their traffic in 2015 was bad bots.

Finally, web scraping is used for weather data monitoring and website change detection, which emails notifications to users about changes made to specific websites.

The statistics used in this report for the top uses of web scraping were created through analysis of the industry leaders: Scraping Hub, Diffbot, and Screen-Scraper. On the customer pages of the company websites, these three industry leaders list their clients and the use cases for web scraping by client. The data was sorted into an excel sheet by web scraping use category to create the statistics.

Industry Leaders, Customers, and Victims

The uses of web scraping are so broad in scale that it is not used by one specific industry.

Example of an original article vs a scraped article
In fact, a wide variety of industries use web scraping. Consulting and marketing research firms, like the Nielsen Company, used web scraping to monitor online buzz for clients and get insights for their consumers until it decided web scraping was bad PR. E-commerce sites in various industries, such as fashion and ticket retailers use web scraping to compare prices and inventory and gather content. These are just a few examples of industries that use web scraping for personal revenue gain.

With a wide variety of web scraping customers, comes a wide variety of web scraping victims. If a website contains content that drives revenue for a business, that business is at risk. The top web scraping victims by industry in 2015 were real estate, digital publishing, e-commerce, directories and classifieds, and airlines and travel. Real estate had the highest percentage of bad bots at 32%. Many of these industries are being targeted by an influx in startups that are scraping information from industry leaders in order to compete.

The impact of web scraping can have a devastating effect on victims’ sales and revenue. Loss of sales is caused by decreased traffic and visitor engagement, which is caused by a lower search engine optimization (SEO) ranking and deflated brand awareness. As a result, this decreases advertising revenue and creates a loss of readership and subscriber base by providing a poor user experience. Costs are then increased by a rise in network and bandwidth costs and new legal fees to handle duplicated content and copyright infringement lawsuits. This loss of revenue and surge in costs ultimately runs victims into debt and out of business.

**Scale of the Industry, Cost, and Size**

Web scraping is readily available in a variety of forms, enabling the average person to obtain scraped data and content. These forms include both web scraping services and do-it-yourself web scraping software, making web scraping easily accessible. Additionally, a number of websites host ads providing freelance and company web scraping services and ads seeking web scraping services, with new ads being posted every day. For example, Freelancer.com hosts a page where individuals can post ads of jobs for web scrapers. Interested web scrapers can then bid on the jobs. Another freelance website, Guru.com, lists 1,800 web scraping services as of August 15, 2016. The wide scale costs of web scraping products and services also contributes to web scraping’s accessibility.
Screen-Scraper

One of the web scraping industry leaders, Screen-Scraper, which was founded in 2001, provides web data extraction products and services, with downloadable software costs ranging from $1,680 to free. Screen-Scraper’s services range from $10,000-$1,500 per job. In contrast, Diffbot, a web scraping software provider, has prices ranging from $299 per month for 250,000 scraped URLs and $3999 per month for 3,000,000 scraped URLs. On the opposite end of the scale, freelance web scraping services posted on UpWork.com are priced as low as $3.33 per hour of work, with the average scraping project costing $135.

Screen-Scraper’s customers include some of the largest companies in the world, Microsoft, Amazon, Oracle, and FedEx. Screen-Scraper claims that its capabilities include scraping data from virtually any website. This positions Screen-Scraper as a large threat to web scraping victims and the competitors of its customers.

Other web scraping industry leaders include Mozenda, Diffbot, and Scrapinghub. Founded in 2007, Mozenda helps users convert unstructured web data into usable data sets and data feeds. Once data is obtained, on either a scheduled or manual basis, it can be viewed online, exported for offline use, or accessed through an application program interface (API).

Mozenda

Popular uses of Mozenda include web data mashups, competitive pricing, and analysis of customer sentiment on the web. Because of the nature of the tool, programmers and non-programmers alike can extract data from any website for a variety of purposes. Additionally, Mozenda keeps its customers confidential as to prevent any negative connotations from being attached to customers, as in the Nielsen Company case.

Diffbot

Unlike Screen-Scraper, Diffbot only provides web scraping software, not services. Diffbot’s web scraping product is a set of APIs that enables developers to easily use web data in their own applications. Diffbot analyzes documents much like a human would, using the visual properties to determine how the parts of the page fit together. Being a startup company that was founded in 2010, Diffbot’s technology is spreading fast and being used by some of the world’s largest content companies.

Scrapinghub

Also founded in 2010, Scrapinghub offers web scraping products and services, specializing in data extraction. Scrapinghub takes pride in the fact that their products empower everyone from programmers to CEOs to extract data quickly and effectively using open source technologies. Their platform is used to scrape over 3 billion web pages a month.
On average, web scrapers at the leading web scraping companies make $58,000, within the range of $20,000-128,000\(^1\). However, for freelance web scrapers who do not have the backing of a large company, the hourly wage for a web scraping service ranges from $5-$30.

The accessibility of web scraping software provides someone looking to make money the opportunity to learn the skill of web scraping and contribute to the growth of the web scraping economy. The most common programs used for web scraping are Python, Ruby, Java, and Perl.

Some web pages with small amounts of content and little to no barriers to web scraping, like web application firewalls (WAFs) and CAPTCHA, can be scraped in less than an hour\(^2\). Other web pages, that contain larger amounts of content and greater barriers can take weeks to scrape. The amount of time that it takes to scrape a web page depends on the amount of data that is being scraped, the level of the bot (advanced persistent bots are able to scrape faster than other malicious bots), and the complexity of barriers that are trying to protect against web scraping.

---

\(^1\) [http://www.indeed.com/jobs?q=web+scraping&l=San+Francisco%2C+CA](http://www.indeed.com/jobs?q=web+scraping&l=San+Francisco%2C+CA)

Conclusion

The world wide web is already seeing the rise of web scrapers as a service and product and an increase in technology sophistication. The web is also seeing the democratization of scraping, where now the average person can download an app for Windows that will allow them to scrape an entire website and save it to a computer. These services are allowing more people with low levels of technical sophistication, and less money, to get involved in the web scraping game. A likely future trend is an increase in the number of people who use web scraping. As web scraping services advance, it can be predicted that there will be continued advancements in web scraping and how it works. Luckily for content owners, Distil Networks is disrupting the web scraping economy by providing businesses with a way to prevent web scraping, eventually bringing this lucrative line of malicious content theft to a very quick end.

For more information on Distil Networks, visit us at http://www.distilnetworks.com or follow @DISTIL on Twitter.