DATA-XTRACT

Scrapers, Crawlers, Bots and the Future of Capturing Important Data

Executive Summary

This white paper is intended for marketers, data executives, and sales professionals who are in the process of developing solutions that best leverage the opportunities offered by big data. The evolving and intermingling relationship between AI and big data alters the landscape of real-time data acquisition online. But how is this acquisition best accomplished? Two methods currently dominate the world of online data collection, 'data scrapers' and 'data crawlers'. Specifically, this paper compares the advantages and disadvantages of data scraping and data crawling.

Unlike crawling, data scraping can be done at any point in a data exchange, such within the browser, which is where Data-Xtract collects its data, or from any server along the way that houses data. As AI capabilities become ever more woven into big data analytics, more and more crawlers, meaning 'bots' that crawl around the web collecting information from web pages, will be dispatched with ever more sophisticated missions. The job of a crawler is to move through every link on every page it visits.

In the example of search engines, this information is sent to a home server where it is processed through algorithms that will determine the relevance of each page that had been crawled.

However, a problem is already arising in the web crawler space that is likely to grow worse. As crawlers take up more of a website's resources for the purpose of using its data without compensation, sites are blocking crawlers and the practice is likely to grow.

So then how best can a company collect and use data relevant to its own operations?

Scraping vs. Crawling

'Scraping' and 'crawling' are often used interchangeably, even by coders. But the colloquialism ignores an issue of scale that makes them functionally two separate things.

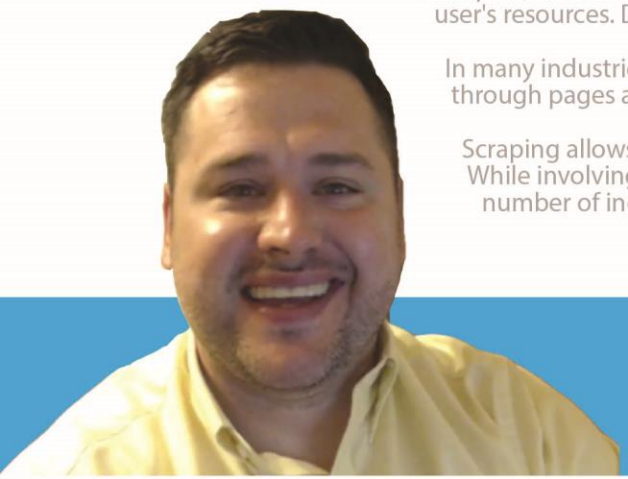
It's important to keep in mind that scraping is usually a part of crawling. So while one can scrape while crawling, one can't crawl while scraping. For search engines, the scraping that is done during a crawl usually amounts to capturing images, keywords, and meta-descriptions or a few hundred characters from a page's first paragraph.

Crawling involves the mapping of large segments of the web and then calibrating the bots to extract all of the targeted data. The crawlers must follow the rules of the road, ignoring pages it's told to skip. This includes pages that you can see on your browser. A server can be programmed to instruct a crawler to ignore a page while still publicly serving the page to browsers. Simply stated, you can see the data on the page, but your crawler can't use it.

Scrapers, on the other hand, like Data-Xtract, can take data from inside the browser and feed it to the user's resources. Data loss from blocked pages is therefore eliminated.

In many industries, the value of human analysis puts many tech workers in front of screens searching through pages as a function of their jobs, one that takes up a good deal of their time.

Scraping allows for further offline analysis of data collected as the browser visits page after page. While involving less automation than crawling, scraping adequately serves professionals in a number of industries like sales, recruiting and real estate.



"Data mining changed the way I view sourcing,recruiting, or any other online search. My vision is to create a product where everyone can fully leverage data to their advantage"

- Shawn Petty - Founder / Product Owner - Data-Xtract

Scrapers, Crawlers, Bots and the Future of Capturing Important Data continued...

Use it, Don't Lose it

Let's take a real estate office and the opportunities for its professionals to collect, analyze, and better use Internet data. Every real estate agent needs to keep up with what is happening in their market, and conduct searches for clients looking for homes.

In other words, a real estate agent is going to be visiting a lot of web pages to find the data they need to sell or search for listings on mulitple platforms. In this circumstance, being able to collect current data on all pages visited, rather than depend on the outdated or incomplete data on crawler built databases, gives the office a competitive advantage.

Using a scraper, especially a browser-based scraper like Data-Xtract, allows for the extraction of data from all visited pages because it does not interfere with, or even communicate with, the server housing the content. A real estate office can be confident in its data and the resulting analytics, which will give them the best possible insights into the local market.

Get Your Data the Easy Way

Data scraping can be done in-house much easier than crawling. Crawling involves creating a bot program, defining its parameters, and then sending it loose on the web. Only high-experienced, and very sought after, data professionals are capable of building, launching, and managing an effective crawler.

In order to keep from being banned, crawlers must adhere to all the rules of the road, and that not only means obeying 'road closed' signs and leaving blocked pages alone, but making sure crawlers don't re-visit a page too often or conflict with other crawlers on the site.

Scraping requires some technical prowess, but basically, anyone who has ventured 'under the hood' of their digital world and come out alive can figure out how to use a scraper. They are painless to configure for specific tasks on any scale.

Plus, data management best practices prefer scraping over crawling as scraping requires much less deduping than crawling. Deduping is the process of eliminating duplicate content. For example, newspaper websites that subscribe to wire services run many stories as-is, meaning the story appears as written by the wire service on hundreds of websites. All this duplicate content must be dealt with, and that means computer resources which can slow down other aspects of a company's digital operations.

Scrapers can also retrieve information from other sources, such as spreadsheets or internal databases, which gives them more versatility than crawlers.

Conclusion

The future of crawlers looks like it will be filled with controversy. As big data demands escalate, crawlers will require more and more resources, behind which will be the understandable cries of 'foul' from website operators. If current trends prevail, the likely course will be a period of intense competition to build the most brutal crawlers, followed by a couple of royally stupid events that will bring a lot of negative publicity, like consumer data corruption caused by aggressive crawlers. Calls for laws and regulations will certainly follow.

So a good plan would be one which steered clear of the predictable obstacles that will revolve around big data. Scrapers operate on home resources only, so they are no threat to the operations of any website or website resource. Using them now, with an eye toward greater adoption of scrapers in data collection, will give businesses the best big data opportunity, one in which data will be relevant and reliable.