Web scraping

Web scraping, **web harvesting**, or **web data extraction** is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet or loaded into a database. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and telephone numbers, or companies and their URLs, or e-mail addresses to a list (contact scraping).

Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), gathering real estate listings, weather data monitoring, website change detection, research, tracking online presence and reputation, web mashup, and web data integration.

Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. As a result, specialized tools and software have been developed to facilitate the scraping of web pages.

Newer forms of web scraping involve monitoring data feeds from web servers. For example, JSON is commonly used as a transport storage mechanism between the client and the web server.

There are methods that some websites use to prevent web scraping, such as detecting and disallowing bots from crawling (viewing) their pages. In response, there are web scraping systems that rely on using techniques in DOM parsing, computer vision and natural language processing to simulate human browsing to enable gathering web page content for offline parsing.

Contents

History

The history of the web scraping dates back nearly to the time when the World Wide Web was born.

- After the birth of **World Wide Web** in 1989, the first web robot,[1] **World Wide Web Wanderer**, was created in June 1993, which was intended only to measure the size of the web.
- In December 1993, the first **crawler-based web search engine**, JumpStation, was launched. As there were not so many websites available on the web, search engines at that time used to rely on their human website administrators to collect and edit the links into a particular format. In comparison, JumpStation brought a new leap, being the first WWW search engine that relied on a web robot.
- In 2000, the **first Web API and API crawler** came. API stands for **Application Programming Interface**. It is an interface that makes it much easier to develop a program by providing the building blocks. In 2000, Salesforce and eBay launched their own API, with which programmers were enabled to access and download some of the data available to the public. Since then, many websites offer web APIs for people to access their public database.

Techniques

Web scraping is the process of automatically mining data or collecting information from the World Wide Web. It is a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding, artificial intelligence and human-computer interactions. Current web scraping solutions range from the ad-hoc, requiring human effort, to fully automated systems that are able to convert entire web sites into structured information, with limitations.

Human copy-and-paste

The simplest form of web scraping is manually copying and pasting data from a web page into a text file or spreadsheet. Sometimes even the best web-scraping technology cannot replace a human's manual examination and copy-and-paste, and sometimes this may be the only workable solution when the websites for scraping explicitly set up barriers to prevent machine automation.

Text pattern matching

A simple yet powerful approach to extract information from web pages can be based on the UNIX grep command or regular expression-matching facilities of programming languages (for instance Perl or Python).

HTTP programming

Static and dynamic web pages can be retrieved by posting HTTP requests to the remote web server using socket programming.

HTML parsing

Many websites have large collections of pages generated dynamically from an underlying structured source like a database. Data of the same category are typically encoded into similar pages by a common script or template. In data mining, a program that detects such templates in a particular information source, extracts its content and translates it into a relational form, is called a wrapper. Wrapper generation algorithms assume that input pages of a wrapper induction system conform to a common template and that they can be easily identified in terms of a URL common scheme.[2] Moreover, some semi-structured data query languages, such as XQuery and the HTQL, can be used to parse HTML pages and to retrieve and transform page content.

DOM parsing

By embedding a full-fledged web browser, such as the Internet Explorer or the Mozilla browser control, programs can retrieve the dynamic content generated by client-side scripts. These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages. Languages such as Xpath can be used to parse the resulting DOM tree.

Vertical aggregation

There are several companies that have developed vertical specific harvesting platforms. These platforms create and monitor a multitude of "bots" for specific verticals with no "man in the loop" (no direct human involvement), and no work related to a specific target site. The preparation involves establishing the knowledge base for the entire vertical and then the platform creates the bots automatically. The platform's robustness is measured by the quality of the information it retrieves (usually number of fields) and its scalability (how quick it can scale up to hundreds or thousands of sites). This scalability is mostly used to target the Long Tail of sites that common aggregators find complicated or too labor-intensive to harvest content from.

Semantic annotation recognizing

The pages being scraped may embrace metadata or semantic markups and annotations, which can be used to locate specific data snippets. If the annotations are embedded in the pages, as Microformat does, this technique can be viewed as a special case of DOM parsing. In another case, the annotations, organized into a semantic layer,[3] are stored and managed separately from the web pages, so the scrapers can retrieve data schema and instructions from this layer before scraping the pages.

Computer vision web-page analysis

There are efforts using machine learning and computer vision that attempt to identify and extract information from web pages by interpreting pages visually as a human being might.[4]

Software

There are many software tools available that can be used to customize web-scraping solutions. This software may attempt to automatically recognize the data structure of a page or provide a recording interface that removes the necessity to manually write web-scraping code, or some scripting functions that can be used to extract and transform content, and database interfaces that can store the scraped data in local databases. Some web scraping software can also be used to extract data from an API directly.

Legal issues

The legality of web scraping varies across the world. In general, web scraping may be against the terms of use of some websites, but the enforceability of these terms is unclear.[5]

United States

In the United States, website owners can use three major legal claims to prevent undesired web scraping: (1) copyright infringement (compilation), (2) violation of the Computer Fraud and Abuse Act ("CFAA"), and (3) trespass to chattel.[6] However, the effectiveness of these claims relies upon meeting various criteria, and the case law is still evolving. For example, with regard to copyright, while outright duplication of original expression will in many cases be illegal, in the United States the courts ruled in *Feist Publications v. Rural Telephone Service* that duplication of facts is allowable.

U.S. courts have acknowledged that users of "scrapers" or "robots" may be held liable for committing trespass to chattels,[7][8] which involves a computer system itself being considered personal property upon which the user of a scraper is trespassing. The best known of these cases, *eBay v. Bidder's Edge*, resulted in an injunction ordering Bidder's Edge to stop accessing, collecting, and indexing auctions from the eBay web site. This case involved automatic placing of bids, known as auction sniping. However, in order to succeed on a claim of trespass to chattels, the plaintiff must demonstrate that the defendant intentionally and without authorization interfered with the plaintiff's possessory interest in the computer system and that the defendant's unauthorized use caused damage to the plaintiff. Not all cases of web spidering brought before the courts have been considered trespass to chattels.[9]

One of the first major tests of screen scraping involved American Airlines (AA), and a firm called FareChase.[10] AA successfully obtained an injunction from a Texas trial court, stopping FareChase from selling software that enables users to compare online fares if the software also searches AA's website. The airline argued that FareChase's websearch software trespassed on AA's servers when it collected the publicly available data. FareChase filed an appeal in March 2003. By June, FareChase and AA agreed to settle and the appeal was dropped.[11]

Southwest Airlines has also challenged screen-scraping practices, and has involved both FareChase and another firm, Outtask, in a legal claim. Southwest Airlines charged that the screen-scraping is Illegal since it is an example of "Computer Fraud and Abuse" and has led to "Damage and Loss" and "Unauthorized Access" of Southwest's site. It also constitutes "Interference with Business Relations", "Trespass", and "Harmful Access by Computer". They also claimed that screen-scraping constitutes what is legally known as "Misappropriation and Unjust Enrichment", as well as being a breach of the web site's user agreement. Outtask denied all these claims, claiming that the prevailing law, in this case, should be US Copyright law and that under copyright, the pieces of information being scraped would not be subject to copyright

protection. Although the cases were never resolved in the Supreme Court of the United States, FareChase was eventually shuttered by parent company Yahoo!, and Outtask was purchased by travel expense company Concur.[12] In 2012, a startup called 3Taps scraped classified housing ads from Craigslist. Craigslist sent 3Taps a cease-and-desist letter and blocked their IP addresses and later sued, in *Craigslist v. 3Taps*. The court held that the cease-and-desist letter and IP blocking was sufficient for Craigslist to properly claim that 3Taps had violated the Computer Fraud and Abuse Act.

Although these are early scraping decisions, and the theories of liability are not uniform, it is difficult to ignore a pattern emerging that the courts are prepared to protect proprietary content on commercial sites from uses which are undesirable to the owners of such sites. However, the degree of protection for such content is not settled and will depend on the type of access made by the scraper, the amount of information accessed and copied, the degree to which the access adversely affects the site owner's system and the types and manner of prohibitions on such conduct.[13]

While the law in this area becomes more settled, entities contemplating using scraping programs to access a public web site should also consider whether such action is authorized by reviewing the terms of use and other terms or notices posted on or made available through the site. In a 2010 ruling in the *Cvent, Inc. v. Eventbrite, Inc.* In the United States district court for the eastern district of Virginia, the court ruled that the terms of use should be brought to the users' attention In order for a browse wrap contract or license to be enforced.[14] In a 2014 case, filed in the United States District Court for the Eastern District of Pennsylvania,[15] e-commerce site QVC objected to the Pinterest-like shopping aggregator Resultly's 'scraping of QVC's site for real-time pricing data. QVC alleges that Resultly "excessively crawled" QVC's retail site (allegedly sending 200-300 search requests to QVC's website per minute, sometimes to up to 36,000 requests per minute) which caused QVC's site to crash for two days, resulting in lost sales for QVC.[16] QVC's complaint alleges that the defendant disguised its web crawler to mask its source IP address and thus prevented QVC from quickly repairing the problem. This is a particularly interesting scraping case because QVC is seeking damages for the unavailability of their website, which QVC claims was caused by Resultly.

In the plaintiff's web site during the period of this trial, the terms of use link are displayed among all the links of the site, at the bottom of the page as most sites on the internet. This ruling contradicts the Irish ruling described below. The court also rejected the plaintiff's argument that the browse-wrap restrictions were enforceable in view of Virginia's adoption of the Uniform Computer Information Transactions Act (UCITA)—a uniform law that many believed was in favor on common browse-wrap contracting practices.[17]

In *Facebook, Inc. v. Power Ventures, Inc.*, a district court ruled in 2012 that Power Ventures could not scrape Facebook pages on behalf of a Facebook user. The case is on appeal, and the Electronic Frontier Foundation filed a brief in 2015 asking that it be overturned.[18][19] In *Associated Press v. Meltwater U.S. Holdings, Inc.*, a court in the US held Meltwater liable for scraping and republishing news information from the Associated Press, but a court in the United Kingdom held in favor of Meltwater.

Internet Archive collects and distributes a significant number of publicly available web pages without being considered to be in violation of copyright laws.

European Union

In February 2006, the Danish Maritime and Commercial Court (Copenhagen) ruled that systematic crawling, indexing, and deep linking by portal site ofir.dk of estate site Home.dk does not conflict with Danish law or the database directive of the European Union.[20]

In a February 2010 case complicated by matters of jurisdiction, Ireland's High Court delivered a verdict that illustrates the inchoate state of developing case law. In the case of *Ryanair Ltd v Billigfluege.de GmbH*, Ireland's High Court ruled Ryanair's "click-wrap" agreement to be legally binding. In contrast to the findings of the United States District Court Eastern District of Virginia and those of the Danish Maritime and Commercial Court, Justice Michael Hanna ruled that the hyperlink to Ryanair's terms and conditions was plainly visible, and that placing the onus on the user to agree to terms and conditions in order to gain access to online services is sufficient to comprise a contractual relationship.[21] The decision is under appeal in Ireland's Supreme Court.[22]

On April 30, 2020, the French Data Protection Authority (CNIL) released new guidelines on web scraping.[23] The CNIL guidelines made it clear that publicly available data is still personal data and cannot be repurposed without the knowledge of the person to whom that data belongs.[24]

Australia

In Australia, the Spam Act 2003 outlaws some forms of web harvesting, although this only applies to email addresses.[25][26]

India

Leaving a few cases dealing with IPR infringement, Indian courts have not expressly ruled on the legality of web scraping. However, since all common forms of electronic contracts are enforceable in India, violating the terms of use prohibiting data scraping will be a violation of the contract law. It will also violate the Information Technology Act, 2000, which penalizes unauthorized access to a computer resource or extracting data from a computer resource.

Methods to prevent web scraping

The administrator of a website can use various measures to stop or slow a bot. Some techniques include:

- Blocking an IP address either manually or based on criteria such as geolocation and DNSRBL. This will also block all browsing from that address.
- Disabling any web service API that the website's system might expose.
- Bots sometimes declare who they are (using user agent strings) and can be blocked on that basis using robots.txt; 'googlebot' is an example. Other bots make no distinction between themselves and a human using a browser.
- Bots can be blocked by monitoring excess traffic
- Bots can sometimes be blocked with tools to verify that it is a real person accessing the site, like a CAPTCHA. Bots are sometimes coded to explicitly break specific CAPTCHA patterns or may employ third-party services that utilize human labor to read and respond in real-time to CAPTCHA challenges.
- Commercial anti-bot services: Companies offer anti-bot and anti-scraping services for websites. A few web application firewalls have limited bot detection capabilities as well. However, many such solutions are not very effective.[27]
- Locating bots with a honeypot or other method to identify the IP addresses of automated crawlers.
- Obfuscation using CSS sprites to display such data as telephone numbers or email addresses, at the cost of accessibility to screen reader users.

- Because bots rely on consistency in the front-end code of a target website, adding small variations to the HTML/CSS surrounding important data and navigation elements would require more human involvement in the initial set up of a bot and if done effectively may render the target website too difficult to scrape due to the diminished ability to automate the scraping process.
- Websites can declare if crawling is allowed or not in the robots.txt file and allow partial access, limit the crawl rate, specify the optimal time to crawl and more.
- Load database data straight into the HTML DOM via AJAX, and use DOM methods to display it. No visible data in the source document means that it can't be scraped.

See also

- Archive.today
- Comparison of feed aggregators
- Data scraping
- Data wrangling
- Importer
- Job wrapping
- Knowledge extraction
- OpenSocial
- Scraper site
- Fake news website
- Blog scraping
- Spamdexing
- Domain name drop list
- Text corpus
- Web archiving
- Web crawler
- Offline reader
- Link farm (blog network)
- Search engine scraping
- Web crawlers

References

1. "Search Engine History.com" (http://www.searchenginehistory.com/). *Search Engine History*. Retrieved November 26, 2019.
2. Song, Ruihua; Microsoft Research (Sep 14, 2007). "Joint Optimization of Wrapper Generation and Template Detection" (https://web.archive.org/web/20161011080619/https://pdfs.semanticscholar.org/4fb4/3c5a212df751e84c3b2f8d29fabfe56c3616.pdf) (PDF). *The 13th International Conference on Knowledge Discovery and Data Mining*: 894. doi:10.1145/1281192.1281287 (https://doi.org/10.1145%2F1281192.1281287). ISBN 9781595936097. S2CID 833565 (https://api.semanticscholar.org/CorpusID:833565). Archived from the original (https://pdfs.semanticscholar.org/4fb4/3c5a212df751e84c3b2f8d29fabfe56c3616.pdf) (PDF) on October 11, 2016.
3. Semantic annotation based web scraping (http://www.gooseeker.com/en/node/knowledgebase/freeformat)

4. Roush, Wade (2012-07-25). "Diffbot Is Using Computer Vision to Reinvent the Semantic Web" (http://www.xconomy.com/san-francisco/2012/07/25/diffbot-is-using-computer-vision-to-reinvent-the-semantic-web/). www.xconomy.com. Retrieved 2013-03-15.
5. "FAQ about linking – Are website terms of use binding contracts?" (https://web.archive.org/web/20020308222536/http://www.chillingeffects.org/linking/faq.cgi#QID596). www.chillingeffects.org. 2007-08-20. Archived from the original (http://www.chillingeffects.org/linking/faq.cgi#QID596) on 2002-03-08. Retrieved 2007-08-20.
6. Kenneth, Hirschey, Jeffrey (2014-01-01). "Symbiotic Relationships: Pragmatic Acceptance of Data Scraping" (http://scholarship.law.berkeley.edu/btlj/vol29/iss4/16/). *Berkeley Technology Law Journal*. **29** (4). doi:10.15779/Z38B39B (https://doi.org/10.15779%2FZ38B39B). ISSN 1086-3818 (https://www.worldcat.org/issn/1086-3818).
7. "Internet Law, Ch. 06: Trespass to Chattels" (http://www.tomwbell.com/NetLaw/Ch06.html). www.tomwbell.com. 2007-08-20. Retrieved 2007-08-20.
8. "What are the "trespass to chattels" claims some companies or website owners have brought?" (https://web.archive.org/web/20020308222536/http://www.chillingeffects.org/linking/faq.cgi#QID460). www.chillingeffects.org. 2007-08-20. Archived from the original (http://www.chillingeffects.org/linking/faq.cgi#QID460) on 2002-03-08. Retrieved 2007-08-20.
9. "Ticketmaster Corp. v. Tickets.com, Inc" (http://www.tomwbell.com/NetLaw/Ch07/Ticketmaster.html). 2007-08-20. Retrieved 2007-08-20.
10. "American Airlines v. FareChase" (https://web.archive.org/web/20110723131832/http://www.fornova.net/documents/AAFareChase.pdf) (PDF). 2007-08-20. Archived from the original (http://www.fornova.net/documents/AAFareChase.pdf) (PDF) on 2011-07-23. Retrieved 2007-08-20.
11. "American Airlines, FareChase Settle Suit" (http://www.thefreelibrary.com/American+Airlines,+FareChase+Settle+Suit.-a0103213546). The Free Library. 2003-06-13. Retrieved 2012-02-26.
12. Imperva (2011). Detecting and Blocking Site Scraping Attacks (http://www.imperva.com/docs/WP_Detecting_and_Blocking_Site_Scraping_Attacks.pdf). Imperva white paper..
13. Adler, Kenneth A. (2003-07-29). "Controversy Surrounds 'Screen Scrapers': Software Helps Users Access Web Sites But Activity by Competitors Comes Under Scrutiny" (https://web.archive.org/web/20110211123854/http://library.findlaw.com/2003/Jul/29/132944.html). Archived from the original (http://library.findlaw.com/2003/Jul/29/132944.html) on 2011-02-11. Retrieved 2010-10-27.
14. "QVC Inc. v. Resultly LLC, No. 14-06714 (E.D. Pa. filed Nov. 24, 2014)" (http://www.fornova.net/documents/Cvent.pdf) (PDF). 2014-11-24. Retrieved 2015-11-05.
15. "QVC Inc. v. Resultly LLC, No. 14-06714 (E.D. Pa. filed Nov. 24, 2014)" (https://www.scribd.com/doc/249068700/LinkedIn-v-Resultly-LLC-Complaint?secret_password=pEVKDbnvhQL52oKfdrmT). *United States District Court for the Eastern District of Pennsylvania*. Retrieved 5 November 2015.
16. Neuburger, Jeffrey D (5 December 2014). "QVC Sues Shopping App for Web Scraping That Allegedly Triggered Site Outage" (http://newmedialaw.proskauer.com/2014/12/05/qvc-sues-shopping-app-for-web-scraping-that-allegedly-triggered-site-outage/). *The National Law Review*. Proskauer Rose LLP. Retrieved 5 November 2015.
17. "Did Iqbal/Twombly Raise the Bar for Browsewrap Claims?" (http://www.fornova.net/documents/pblog-bna-com.pdf) (PDF). 2010-09-17. Retrieved 2010-10-27.
18. "Can Scraping Non-Infringing Content Become Copyright Infringement... Because Of How Scrapers Work? | Techdirt" (https://www.techdirt.com/articles/20090605/2228205147.shtml). *Techdirt*. 2009-06-10. Retrieved 2016-05-24.
19. "Facebook v. Power Ventures" (https://www.eff.org/cases/facebook-v-power-ventures). *Electronic Frontier Foundation*. Retrieved 2016-05-24.

20. "UDSKRIFT AF SØ- & HANDELSRETTENS DOMBOG" (https://web.archive.org/web/20071012005033/http://www.bvhd.dk/uploads/tx_mocarticles/S_-_og_Handelsrettens_afg_relse_i_Ofir-sagen.pdf) (PDF) (in Danish). bvhd.dk. 2006-02-24. Archived from the original (http://www.bvhd.dk/uploads/tx_mocarticles/S_-_og_Handelsrettens_afg_relse_i_Ofir-sagen.pdf) (PDF) on 2007-10-12. Retrieved 2007-05-30.
21. "High Court of Ireland Decisions >> Ryanair Ltd -v- Billigfluege.de GMBH 2010 IEHC 47 (26 February 2010)" (http://www.bailii.org/ie/cases/IEHC/2010/H47.html). British and Irish Legal Information Institute. 2010-02-26. Retrieved 2012-04-19.
22. Matthews, Áine (June 2010). "Intellectual Property: Website Terms of Use" (http://www.lkshields.ie/htmdocs/publications/newsletters/update26/update26_03.htm). *Issue 26: June 2010*. LK Shields Solicitors Update. p. 03. Retrieved 2012-04-19.
23. "La réutilisation des données publiquement accessibles en ligne à des fins de démarchage commercial | CNIL" (https://www.cnil.fr/fr/la-reutilisation-des-donnees-publiquement-accessibles-en-ligne-des-fins-de-demarchage-commercial). *www.cnil.fr* (in French). Retrieved 2020-07-05.
24. FindDataLab.com (2020-06-09). "Can You Still Perform Web Scraping With The New CNIL Guidelines?" (https://medium.com/@finddatalab/can-you-still-perform-web-scraping-with-the-new-cnil-guidelines-bf3e20d0edc2). *Medium*. Retrieved 2020-07-05.
25. National Office for the Information Economy (February 2004). "Spam Act 2003: An overview for business" (https://www.lloyds.com/~/media/5880dae185914b2487bed7bd63b96286.ashx). Australian Communications Authority. p. 6. Retrieved 2017-12-07.
26. National Office for the Information Economy (February 2004). "Spam Act 2003: A practical guide for business" (http://www.webstartdesign.com.au/spam_business_practical_guide.pdf) (PDF). Australian Communications Authority. p. 20. Retrieved 2017-12-07.
27. Mayank Dhiman Breaking Fraud & Bot Detection Solutions (https://s3.us-west-2.amazonaws.com/research-papers-mynk/Breaking-Fraud-And-Bot-Detection-Solutions.pdf) *OWASP AppSec Cali' 2018* Retrieved February 10, 2018.
