Web scraping

Contents

History

Techniques

Legal issues

Methods to prevent web scraping

See also

References

Human copy-and-paste

Text pattern matching

HTTP programming

HTML parsing

DOM parsing

Vertical aggregation

Semantic annotation recognizing

Computer vision web-page analysis

AI-powered document understanding

United States

European Union

Australia

India

Web scraping,web harvesting, orweb data extractionisdata scrapingused forextracting datafromwebsites.[1]Web scraping software may directly access theWorld Wide Webusing theHypertext Transfer Protocolor a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using abotorweb crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central localdatabaseorspreadsheet, for laterretrievaloranalysis.

Scraping a web page involves fetching it and then extracting data 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. Having fetched, extraction can take place. The content of a page may beparsed, searched and reformatted, and 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 finding and copying names and telephone numbers, companies and their URLs, or e-mail addresses to a list (contact scraping).

As well ascontact scraping, web scraping is used as a component of applications used forweb indexing,web mininganddata mining, online price change monitoring andprice 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, andweb data integration.

Web pagesare built using text-based mark-up languages (HTMLandXHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for humanend-usersand not for ease of automated use. As a result, specialized tools and software have been developed to facilitate the scraping of web pages. Web scraping applications includemarket research, price comparison, content monitoring, and more. Businesses rely on web scraping services to efficiently gather and utilize this data.

Newer forms of web scraping involve monitoringdata feedsfrom web servers. For example,JSONis commonly used as a transport 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, web scraping systems use techniques involvingDOMparsing,computer visionandnatural language processingto simulate human browsing to enable gathering web page content for offline parsing.

After thebirth of the World Wide Webin 1989, the first web robot,[2]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 fewer websites available on the web, search engines at that time used to rely on human administrators to collect and format links. In comparison, Jump Station was the first WWW search engine to rely on a web robot.

In 2000, the first Web API and API crawler were created. AnAPI(Application Programming Interface) is an interface that makes it much easier to develop a program by providing the building blocks. In 2000,SalesforceandeBaylaunched their own API, with which programmers could access and download some of the data available to the public.[3]Since then, many websites offer web APIs for people to access their public database.

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 thesemantic webvision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding, artificial intelligence andhuman-computer interactions.

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.

A simple yet powerful approach to extract information from web pages can be based on the UNIXgrepcommand orregular expression-matching facilities of programming languages (for instancePerlorPython).

Staticanddynamic web pagescan be retrieved by posting HTTP requests to the remote web server usingsocket programming.

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 awrapper. 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.[4]Moreover, somesemi-structured dataquery languages, such asXQueryand the HTQL, can be used to parse HTML pages and to retrieve and transform page content.

By using a program such asSeleniumorPlaywright, developers can control a web browser such asChromeorFirefoxwherein they can load, navigate, and retrieve data from websites. This method can be especially useful for scraping data from dynamic sites since a web browser will fully load each page. Once an entire page is loaded, you can access and parse theDOMusing an expression language such asXPath.

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 theLong Tailof sites that common aggregators find complicated or too labor-intensive to harvest content from.

The pages being scraped may embracemetadataor semantic markups and annotations, which can be used to locate specific data snippets. If the annotations are embedded in the pages, asMicroformatdoes, this technique can be viewed as a special case of DOM parsing. In another case, the annotations, organized into a semantic layer,[5]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.

There are efforts usingmachine learningandcomputer visionthat attempt to identify and extract information from web pages by interpreting pages visually as a human being might.[6]

Uses advanced AI to interpret and process web page content contextually, extracting relevant information, transforming data, and customizing outputs based on the content's structure and meaning. This method enables more intelligent and flexible data extraction, accommodating complex and dynamic web content.

The legality of web scraping varies across the world. In general, web scraping may be against theterms of serviceof some websites, but the enforceability of these terms is unclear.[7]

In the United States, website owners can use three majorlegal claimsto prevent undesired web scraping: (1) copyright infringement (compilation), (2) violation of theComputer Fraud and Abuse Act("CFAA"), and (3)trespass to chattel.[8]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 inFeist Publications v. Rural Telephone Servicethat duplication of facts is allowable.

U.S. courts have acknowledged that users of "scrapers" or "robots" may be held liable for committingtrespass to chattels,[9][10]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 asauction sniping. However, in order to succeed on a claim of trespass tochattels, theplaintiffmust demonstrate that thedefendantintentionally 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.[11]

One of the first major tests ofscreen scrapinginvolvedAmerican Airlines(AA), and a firm called FareChase.[12]AA successfully obtained aninjunctionfrom 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.[13]

Southwest Airlineshas 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 beUS Copyright lawand that under copyright, the pieces of information being scraped would not be subject to copyright protection. Although the cases were never resolved in theSupreme Court of the United States, FareChase was eventually shuttered by parent companyYahoo!, and Outtask was purchased by travel expense company Concur.[14]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, inCraigslist 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 theComputer Fraud and Abuse Act(CFAA).

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.[15]

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. InCvent Inc.v.Eventbrite Inc.(2010), the United Statesdistrict court for the eastern district of Virginia, ruled that the terms of use should be brought to the users' attention in order for abrowsewrapcontract or license to be enforceable.[16]In a 2014 case, filed in theUnited States District Court for the Eastern District of Pennsylvania,[17]e-commerce siteQVCobjected 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.[18]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.[19]

InFacebook, 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 theElectronic Frontier Foundationfiled a brief in 2015 asking that it be overturned.[20][21]InAssociated 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.

TheNinth Circuitruled in 2019 that web scraping did not violate the CFAA inhiQ Labs v. LinkedIn. The case was appealed to theUnited States Supreme Court, which returned the case to the Ninth Circuit to reconsider the case in light of the 2021 Supreme Court decision inVan Buren v. United Stateswhich narrowed the applicability of the CFAA.[22]On this review, the Ninth Circuit upheld their prior decision.[23]

Internet Archivecollects and distributes a significant number of publicly available web pages without being considered to be in violation of copyright laws.[citation needed]

In February 2006, theDanish Maritime and Commercial Court(Copenhagen) ruled that systematic crawling, indexing, and deep linking by portal site ofir.dk of real estate site Home.dk does not conflict with Danish law or the database directive of the European Union.[24]

In a February 2010 case complicated by matters of jurisdiction, Ireland's High Court delivered a verdict that illustrates theinchoatestate of developing case law. In the case ofRyanair Ltd v Billigfluege.de GmbH, Ireland's High Court ruledRyanair'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, JusticeMichael Hannaruled 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.[25]The decision is under appeal in Ireland's Supreme Court.[26]

On April 30, 2020, the French Data Protection Authority (CNIL) released new guidelines on web scraping.[27]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.[28]

In Australia, theSpam Act 2003outlaws some forms of web harvesting, although this only applies to email addresses.[29][30]

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 theInformation Technology Act, 2000, which penalizes unauthorized access to a computer resource or extracting data from a computer resource.

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