Web scraping

From Wikipedia, the free encyclopedia

Web scraping (**web harvesting** or **web data extraction**) is a computer software technique of extracting information from websites. Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser, such as Internet Explorer or Mozilla Firefox.

Web scraping is closely related to web indexing, which indexes information on the web using a bot or web crawler and is a universal technique adopted by most search engines. In contrast, web scraping focuses more on the transformation of unstructured data on the web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web scraping is also related to web automation, which simulates human browsing using computer software. Uses of web scraping include online price comparison, contact scraping, weather data monitoring, website change detection, research, web mashup and web data integration.

Web scraping related traffic has increased during recent years. In average 23% of all traffic was scraping-related in 2013.[1]

Contents

Techniques

Web scraping is the process of automatically 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:** 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 grepping and regular expression 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 parsers:** 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.

- **Web-scraping 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.
- **Vertical aggregation platforms:** 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, 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 analyzers:** 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]

Legal issues

Web scraping may be against the terms of use of some websites. The enforceability of these terms is unclear.[5] 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,[6][7] 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.[8]

One of the first major tests of screen scraping involved American Airlines (AA), and a firm called FareChase.[9] AA successfully obtained an injunction from a Texas trial court, stopping FareChase from selling software that enables users to compare online fares if it 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.[10]

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.[11] 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.[12]

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 the latest 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.[13]

In the plaintiff's web site during the period of this trial the terms of use link is 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.[14]

Outside of the United States, in February 2006, the Danish 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.[15]

In 2009 Facebook won one of the first copyright suits against a known web scraper. This laid the groundwork for numerous lawsuits that tie any web scraping with a direct copyright violation and very clear monetary damages. The most recent case being AP v Meltwater, where the courts stripped what is referred to as fair use on the internet.[16]

In a February 2010 case complicated by matters of jurisdiction, Ireland's *An Ard-Chúirt* 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, Mr. 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. [17] The decision is under appeal in Ireland's Supreme Court, the *Cúirt Uachtarach na hÉireann*.[18]

In Australia, the Spam Act 2003 outlaws some forms of web harvesting, although this only applies to email addresses.[19][20]

Notable tools

- Apache Camel
- Automation Anywhere
- Boilerpipe
- Convertigo
- cURL
- Data Toolbar
- Diffbot
- Firebug
- Goose (web development project)
- Greasemonkey
- HtmlUnit
- HTTrack
- iMacros
- Jaxer
- JSoup
- Node.js
- nokogiri
- ScraperWiki
- Scrapy
- SimpleTest
- watir
- Wget
- Wireshark
- WSO2 Mashup Server
- Yahoo! Pipes
- Yahoo! query language (yql)
- mechanize
- selenium
- phantomjs

See also

- 30 Digits
- Comparison of feed aggregators
- Job wrapping
- Importer
- OpenSocial
- Report mining
- Scraper site
- Spamdexing
- Text corpus
- Web crawlers

Technical measures to stop bots

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

- Blocking an IP address. 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. Some bots make no distinction between themselves and a human browser.
- Bots can be blocked by excess traffic monitoring.
- 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.
- 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.
- Locating bots with a honeypot or other method to identify the IP addresses of automated crawlers.
- Using CSS sprites to display such data as phone numbers or email addresses, at the cost of accessibility to screen reader users.

References

1. ^ Sentor Managed Security Services (April 2014). "ScrapeSentry Scraping Threat Report 2014" (http://www.scrapesentry.com/scrapesentry-scraping-threat-report-2014/). Retrieved 2014-06-19.
2. ^ Song, Ruihua; Microsoft Research (Sep 14, 2007). "Joint Optimization of Wrapper Generation and Template Detection". *The 13th International Conference on Knowledge Discovery and Data Mining*.
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?" (http://www.chillingeffects.org/linking/faq.cgi#QID596). www.chillingeffects.org. 2007-08-20. Retrieved 2007-08-20.
6. ^ "Internet Law, Ch. 06: Trespass to Chattels" (http://www.tomwbell.com/NetLaw/Ch06.html). www.tomwbell.com. 2007-08-20. Retrieved 2007-08-20.
7. ^ "What are the "trespass to chattels" claims some companies or website owners have brought?" (http://www.chillingeffects.org/linking/faq.cgi#QID460). www.chillingeffects.org. 2007-08-20. Retrieved 2007-08-20.
8. ^ "Ticketmaster Corp. v. Tickets.com, Inc." (http://www.tomwbell.com/NetLaw/Ch07/Ticketmaster.html). 2007-08-20. Retrieved 2007-08-20.
9. ^ "American Airlines v. FareChase" (http://www.fornova.net/documents/AAFareChase.pdf). 2007-08-20. Retrieved 2007-08-20.
10. ^ "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.
11. ^ Imperva (2011). Detecting and Blocking Site Scraping Attacks. Imperva white paper. Retrieved from http://www.imperva.com/docs/WP_Detecting_and_Blocking_Site_Scraping_Attacks.pdf.
12. ^ Adler, Kenneth A. (2003-07-29). "Controversy Surrounds 'Screen Scrapers': Software Helps Users Access Web Sites But Activity by Competitors Comes Under Scrutiny" (http://library.findlaw.com/2003/Jul/29/132944.html). Retrieved 2010-10-27.
13. ^ "IN THE UNITED STATES DISTRICT COURT FOR THE EASTERN DISTRICT OF VIRGINIA Alexandria Division" (http://www.fornova.net/documents/Cvent.pdf). 2010-09-15. Retrieved 2010-10-27.
14. ^ "Did Iqbal/Twombly Raise the Bar for Browsewrap Claims?" (http://www.fornova.net/documents/pblog-bna-com.pdf). 2010-09-17. Retrieved 2010-10-27.
15. ^ "UDSKRIFT AF SØ- & HANDELSRETTENS DOMBOG" (http://www.bvhd.dk/uploads/tx_mocarticles/S_-_og_Handelsrettens_afg_relse_i_Ofir-sagen.pdf). bvhd.dk. 2006-02-24. Retrieved 2007-05-30.
16. ^ "Is web scraping illegal depends on what the meaning of the word is is" (http://www.distilnetworks.com/is-web-scraping-illegal-depends-on-what-the-meaning-of-the-word-is-is/). Distil Networks Retrieved 2013-07-18
17. ^ "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.
18. ^ 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.
19. ^ National Office for the Information Economy (February 2004). "Spam Act 2003: An overview for business" (http://www.acma.gov.au/webwr/consumer_info/spam/spam_overview_for%20_business.pdf). Australian Communications Authority. p. 6. Retrieved 2009-03-09.
20. ^ National Office for the Information Economy (February 2004). "Spam Act 2003: A practical guide for business" (http://www.acma.gov.au/webwr/consumer_info/frequently_asked_questions/spam_business_practical_guide.pdf). Australian Communications Authority. p. 20. Retrieved 2009-03-09.

See also

- Data scraping
- Data wrangling
- knowledge discovery