News Letter Q2, 2014 - Web Scraping

Web Scraping (also termed as Screen Scraping, Web Data Extraction, Web Harvesting etc..) is a technique employed to extract large amounts of data from websites.

Data from third party websites in the Internet can normally be viewed only using a web browser. Examples are data listings at stock exchange sites, social networks, industrial inventory, online shopping sites, contact databases etc. Most websites do not offer the functionality to save a copy of the data which they display to your local storage. The only option then is to manually copy and paste the data displayed by the website in your browser to a local file in your computer - a very tedious job which can take many hours or sometimes days to complete.

Web Scraping Techniques

Depending on various data needs like volume of data, data security, interval of getting the data, website to be scraped etc. different Web Scraping Techniques are used.

What's in your Mind 

Web scraping is the process of automatically collecting information from the World Wide Web. Web scraping solutions are based on different technologies that are often entirely ad hoc. Therefore, there are different levels of automation that existing web-scraping technologies can provide:

> 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 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.
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.

There are few other technologies like Vertical aggregation platforms and Semantic annotation recognizing that are also used for Web Scraping.

Web Scraping for Financial Data Aggregation and Market Research

A Web Scraping software will interact with Financial services websites like Bloomberg and others in the same way as your web browser. But instead of displaying the data served by the website on screen, the Web Scraping software saves the required data from the web page to a local file or database.

Automated Web Scraping enables financial data providers to continually track bank and custodial sites, consumer financial services sites and disparate news, corporate governmental and media sources world-wide. An Automated solution can:-

Facilitate near real-time reporting by detecting market moving events using build in change detection and targeted keyword searches.
Improve Improve accuracy by de-duplicating collected data and eliminating manual intervention in data uploading.
Broaden coverage by collecting accurate data from thousands of websites worldwide in any language and transforming that data into a structured format

Scraping market data from public website is a legally questionable method that involves creating automated process to extract the data from public resource which generally violate the data issues for

different websites. As we know that time to market is very much important for asset managers, but Scraping is a time intensive process which again does not provide Real time data availability. The data totally depends on the source from which it is scraped. Additionally there is always a risk of formatting changes that make scraping not a very reliable automation method.

Conclusion

In the end, there is no one size fits all solution that can be envisaged but one has to take in to factor the maturity of their organization like number of clients, employees and their IT skill set, associated vendor eco system they've in place to chart their own.

We recommend checking out the "Terms and Conditions" of the data source to be scraped for legality and scraping may not be the solution always recommended.

About ConvergeSol

ConvergeSol is a premier industry focused technology consulting company that specializes in providing outsourcing services for financial services industry and Small / Medium Enterprises. Our offerings include:

Custom Software Development

Automation Solutions - Data Warehousing, Excel Automation, Automated testing, Custom software development, CRM

Reporting Solutions

Portal Solutions - Microsoft SharePoint Solutions, Web Portals.

Integration and Customization Solutions – Order Management systems, portfolio accounting systems, CRM, Work flow management etc.. for 3rd party products like Backstop, Charles River, Palladyne, Advent and Relativity.

Product Development

Product Management

Software Development

Quality Assurance Solutions.

Managed Services

Application Lifecylce Management Offerings, a one stop end-to-end solution perfect for organizations looking to use IT to their advantage, low touch and lower costs.

Operations and Support offering.

We bring deep expertise in understanding your needs be it order management system, enterprise content management, portals, account opening process or reconciliation process just to name few areas. We complement it by our proven execution expertise across entire software life cycle: project management, business requirement, software development, quality assurance, deployment and supSport.

For more information, please contact us at:

Phone: +1. 212.899.5148 (USA)
+1. 732.485.9207 (USA)
+91. 823.899.0929 (INDIA)
+91. 890.546.6278 (INDIA)
Email: info@convergesolution.com
Web: www.convergesolution.com