A Comparative Study on Web Scraping

SCM de S Sirisuriya
Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka
mihiri@kdu.ac.lk

***Abstract*— *The World Wide Web contains all kinds of information of different origins; some of those are social, financial, security and academic. Most people access information through internet for educational purposes. Information on the web is available in different formats and through different access interfaces. Therefore, indexing or semantic processing of the data through websites could be cumbersome. Web Scraping is the technique which aims to address this issue. Web scraping is used to transform unstructured data on the web into structured data that can be stored and analysed in a central local database or spreadsheet. There are various web scraping techniques including Traditional copy-and-paste, Text grapping and regular expression matching, HTTP programming, HTML parsing, DOM parsing, Web-scraping software, Vertical aggregation platforms, Semantic annotation recognizing and Computer vision web-page analysers. Traditional copy and paste is the basic and tiresome web scraping technique where people need to scrap lots of datasets. Web scraping software is the easiest scraping technique since all the other techniques except traditional copy and paste require some form of technical expertise. There are hundreds of web scraping software available today, most of them designed by using Java, Python and Ruby. There are also some open source web scraping software and as well as commercial software. Web scraping software such as YahooPipes, Google Web Scrapers and Outwit Firefox extensions are the best tools for beginners in web scraping. This study focused on giving comparative clarification about web scraping techniques and famous web scraping software. To accomplish this, we compare and contrast several web scraping techniques and some famous web scraping software. The outcome of this study offers a review on web scraping techniques and software which can be used to extract data from educational web sites.*

Keywords*— *Web Scraping, Information Extraction

I. INTRODUCTION

Data is an essential part of any research, either it can be academic, marketing or scientific (SysNucleus, n.d.). People might want to collect and analyse data from multiple websites. The different websites which belongs to the specific category displays information in different formats. Even with a single website you may not be able to see all the data at once. The data may be spanned across multiple pages under various sections. Most websites do not allow to save a copy of the data, displayed in their web sites to your local storage (Penman et al., 2009). The only option is to manually copy and paste the data shown by the website to a local file in your computer. This is a very tedious job which can take lot of time. Web Scraping is the technique which people can extract data from multiple websites to a single spreadsheet or database so that it becomes easy to analyse or even visualize the data. The aim of this study is to offers a review on web scraping techniques and software which can be used to extract data from web sites.

Rest of the paper arranged as follows. Section II describes the overview of web scraping. Section III describes the practical usage of web scraping. Section IV describes some web scraping techniques. Section V gives the detail description about web scraping software. Finally Section VI gives the discussion, comparing the several web scraping techniques and some famous web scraping software.

II. OVERVIEW OF WEB SCRAPING

Web Scraping is a great technique of extracting unstructured data from the websites and transforming that data into structured data that can be stored and analysed in a database. Web Scraping is also known as web data extraction, web data scraping, web harvesting or screen scraping. Web scraping is a form of data mining. The overall goal of the web scraping process is to extract information from a websites and transform it into an understandable structure like spreadsheets, database or a comma-separated values (CSV) file as shown in Figure 1. Data like item pricing, stock pricing, different reports, market pricing and product details, can be gathered through web scraping. Extracting targeted

information from websites assists you to take effective decisions in your business.

Figure 2. Structure of the Web Scraping

III. PRACTICES OF WEB SCRAPING

- Online price comparison
- Contact scraping
- Weather data monitoring
- Website change detection
- Research
- Web mash up — integrate data from multiple sources
- Extract offers and discounts
- Scrape job postings information from job portals
- Collect properties lists from real estate websites
- Brand monitoring
- Extract business details from business directory websites like Yelp and Yellow pages
- Collect government data
- Market Analysis

IV. WEB SCRAPING TECHNIQUES

A. Traditional copy and paste

Occasionally the human's manual examination and copy-and-paste method is the best and the workable web-scraping technology. But this is an error-prone, boring and tiresome technique when people need to scrap lots of datasets ("Web scraping," 2015a).

B. Text grapping and regular expression

This is the simple and powerful approach to extract information from web pages. This technique based on the UNIX command or regular expression-matching facilities of programming language ("Web scraping," 2015b).

C. Hypertext Transfer Protocol (HTTP) Programming

This technique used to extract data from static and dynamic web pages. Data can be retrieved by posting HTTP requests to the remote web server using socket programming ("Web scraping," 2015b).

D. Hyper Text Markup Language (HTML) Parsing

Semi-structured data query languages, like XQuery and the Hyper Text Query Language (HTQL), can be used to parse HTML pages and to retrieve and transform page content ("Web scraping," 2015b).

E. Document Object Model (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," 2015b).

F. 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("Web scraping," 2015b).

G. 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 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 labour-intensive to harvest content from ("Web scraping," 2015b).

H. Semantic annotation recognizing

The pages being scraped may embrace metadata or semantic mark-ups 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, 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 ("Web scraping," 2015b).

I. Computer vision web-page analysers

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 ("Web scraping," 2015b).

V. WEB SCRAPING SOFTWARE

Web Scraping Software are the tools that are used to automate the manual copy paste work to gather large amount of data from websites like directory sites, real estate sites, classified websites and job boards. Suppose you want to scrape real estate property details of UK then you need to appoint few guys to copy and paste details from websites to excel by visiting each property page. This way it will take days and even months to get your property data ready to use. So web scraping can automate the manual work programmatically by visiting each page and extract data from pages and parsing the html pages. There are number of Web Scraping Software that available in market that can help you to scrape data from any website you want. Following are the list of some scraping tools.

The Price of Web Scraping Software varies based on features it provide, support and upgrade period. You can always get the trial version and check whether it has all the scraping features that you need ("List of Web Harvester, Data Scraper,Web Scraping Software and Tools," n.d.).

A. Visual Web Ripper

Visual Web Ripper is one of the most advance web scraping software, created by Sequentum group in 2006 that provides functionality that allows you to scrape data from any websites like Business Directories, Simple Web Pages, Classified Sites, Forums and e-commerce site scraping (eBay, amazon, magento sites). Once data scraping finish, data can be exported to structured CSV, Excel, or XML format("List of Web Harvester, Data Scraper,Web Scraping Software and Tools," n.d.).

B. Web Content Extractor

Web Content Extractor (WCE) is a simple user-oriented application developed by Newprosoft. It has good wizard that guide user to setup scraper. You can scrape data from website with few clicks and Web Content Extractor is excellent for putting data into different formats like Excel, text, HTML formats, Microsoft Access database, Structured Query Language(SQL) Script File, MySQL Script File, Extensible Markup Language (XML) file, HTTP submit form and Open Database Connectivity (ODBC) Data source. ("List of Web Harvester, Data Scraper, Web Scraping Software and Tools," n.d.) ("Software for Web Scraping," n.d.).

C. Mozanda Web Scraper

Mozanda Web Scraper is powerful web data extraction service. It can extract data from websites as well as PDFs. It has simple Point and selection interface so non-technical can also make simple scrape. Mozenda runs your scraping project (agent) on their cloud environment which is the main difference of Mozanda from other scrapers. ("List of Web Harvester, Data Scraper, Web Scraping Software and Tools," n.d.).

D. UiPath – Robotic Process Automation

UiPath can automatically log in to a web site, extract data spanning multiple webpages, filter and transform it into the format of user choice, before integrating it into another application or web service. UiPath resembles a real browser with a real user, so it can extract data that most automation tools cannot even see (Savinkin, n.d.).
No programming is needed to create intelligent web agents using its drag-and-drop graphical designer-but the .NET hacker inside you has complete control over the data ("List of Web Harvester, Data Scraper,Web Scraping Software and Tools," n.d.).

E. Out Wit Hub

The OutWit Hub is a powerful Firefox extension Tool for Everyone. The contents extracted from a Web page are presented in an easy and visual way, without requiring any programming skills or advanced technical knowledge. Users can easily extract links, images, email addresses, data tables, etc. from series of pages without ever seeing the source code. Extracted data can be exported to CSV, HTML, Excel or SQL databases, while images and documents, are directly saved to your hard disk. The OutWit Hub is best to use for beginners in web scraping ("Software for Web Scraping," n.d.).

F. Screen Scraper

Screen Scraper is advance web scraping application that comes in three flavor Enterprise, Professional and Basic. Basic version is free to download and use with basic scraping features ("List of Web Harvester, Data Scraper, Web Scraping Software and Tools," n.d.). Other versions take much time for an inexperienced user to master the techniques. The important mechanism is that Screen Scraper can integrate with other systems, with Java support allowing you to write serious scripts for a large scale program (Savinkin, n.d.).

G. WebHarvy

WebHarvy is a lightweight, visual, point-to-click scrape tool. It takes minimum time to master and to extract data. WebHarvy is best suited for quick scraping of text, URLs and images from web pages. Extracted data can be saved into common formats (CSV, Tab Separated Values(TSV), XML) and also SQL for database input" (SysNucleus, n.d.). It is best known for tabular data

extraction, it can extract data that has well-structured HTML. It can't extract data by doing deep crawling and Ajax based data scraping ("List of Web Harvester, Data Scraper, Web Scraping Software and Tools," n.d.).

H. Easy Web Extract

Easy Web Extract by Web2Mine Founded in 2009 is designed for simple and quick data extraction. This scrape tool is written using .NET technology and allows you to apply data transforming built-in scripts (C#, VB, JS). Easy Web Extract is excellent for exporting data into Excel (CSV), text, XML file, HTML formats, MS Access DB, SQL Script File, MySQL Script File, HTTP submit form and ODBC Data source. One shortcoming is that while making a scrape project, loading the URL sometimes takes a long time(Savinkin, n.d.) .

I. WebSunDew

WebSundew is as easy to use web scraping software that allows point-and-click user interface to define fields that you want to scrape from webpages. This screen scraper is designed for high productivity and speed data ripping. The Enterprise edition allows the scrape to run at a remote Server and publish extracted data through FTP (Savinkin, n.d.). It also supports images and file extraction. It can perform multilevel web extraction by doing deep crawling ("List of Web Harvester, Data Scraper, Web Scraping Software and Tools," n.d.).

J. Web Data Extarctor

Web Data Extractor by Automation Anywhere United States founded in 2003 is a web scraping tool specifically designed for Link Extraction, Meta Tag, Body Text, Emails, Phones, Faxes number scraping. It is not good for rule based web scraping. ("List of Web Harvester, Data Scraper, Web Scraping Software and Tools," n.d.).

K. Helium Scraper

Helium is one of the powerful web scraping software that has all the features that one need to scrape data from any web pages. It has point-and-click user interface to define scraping fields. It has support of Ajax based scraping, CAPTCHA based scraping and proxy supports ("List of Web Harvester, Data Scraper,Web Scraping Software and Tools," n.d.).

L. WebExtractor 360

WebExtractor 360 is an open source web scraper. It uses Regular Expression to scrape data from web pages. You need to have good knowledge of Regular Expressions to work with this regular expression based scraping tool. This scrape is completely free and also provides source code.

M. FMiner

Fminer is one of the best Visual Web Scraping tool built in Python. It has nice diagrammatic representation of scraping flow and actions. It also allows to run custom python code ("List of Web Harvester, Data Scraper,Web Scraping Software and Tools," n.d.).

N. Scrapy

An open source and collaborative framework for extracting the data you need from websites. Scrapy written in Python and runs on Linux, Windows, and Mac.

O. import io

import io is a free online web scraper founded in March 2012, which allows you to scrape various types of information and then organize the extracted information into data sets. import io is a cloud-based platform so you don't need to run the scraper on your machine, and all your data is kept somewhere in the cloud. import io is usable for all kinds of people, regardless of their technical ability ("Software for Web Scraping," n.d.).

P. Web Scraper

Web Scraper offer two great options for users. Those are free Google Chrome Extension and Enterprise Data Extraction Service. In Google Chrome Extension user can create a plan (sitemap) how a web site should be traversed and what should be extracted. Using these sitemaps the Web Scraper will navigate the site accordingly and extract all data. Scraped data later can be exported as CSV. In Enterprise Data Extraction Service offers top quality results driven at the level you require. This option allows you to extract large amounts of data, run multiple scrapings at once, and even run them on a set schedule .

VI. DISCUSSION

Visual Web Ripper, Helium Scraper, Screen Scraper, OutWit Hub, Mozenda, WebSundew, Web Content Extractor, Easy Web Extract are commercial web scraping tools. Screen Scraper has free basic edition and OutWit Hub has free Light version and all the others have free trial version. WebExtractor 360 and Scrapy are open source web scraping tools. import io is a free online web scraper. The main difference of the Mozenda screen scraper software from other scrapers is that it runs your scraping projects in clouds.

Table 3. Comparison of Web Scraping Software

Web Scraping Software	Operating System	Data Export formats
Visual Web Ripper	Win	CSV, Excel, XML, SQL Server, MySQL, SQLite, Oracle and OleDB, Customized C# or VB script file output
Helium Scraper	Win	CSV, XML, MS Access database, MySQL script file
Screen Scraper	Win, Mac, Unix/Linux	Text. HTML, SQL Script File, MySQL Script File, XML file, HTTP submit form
OutWit Hub	Win, Mac OS-X, Linux,	CSV (TSV), HTML, Excel or SQL script
Mozenda	Win	CSV, TSV, or XML only.
WebSundew	Win	Text, CSV, Excel, XML; SQL Server, MySQL, Oracle and JDBC compatible DB (Pro and Enterprise edition)
Web Content Extractor	Win	Excel, text, HTML, MS Access DB, SQL Script File, MySQL Script File, XML file, HTTP submit form, ODBC Data source
Easy Web Extract	Win	Excel (CSV, TSV), text, HTML, MS Access DB, SQL Script File, MySQL Script File, XML file, HTTP submit form, ODBC Data source

According to the Table 1 most of the web scraping software supports Windows operating system except Screen Scraper and OutWit Hub. Excel, CSV and XML file are most common data export formats.

Disparate Data Collection, Email Address Extraction, Image Extraction, IP Address Extraction, Phone Number Extraction and Web Data Extraction are common features to import io, Visual Web Ripper, Easy Web Extract and FMiner.

Disparate Data Collection, Document Extraction, Email Address Extraction, Image Extraction, Phone Number Extraction, Pricing Extraction and Web Data Extraction are key features to Helium Scraper.

Email Address Extraction, Image Extraction and Web Data Extraction are the main features of Web Data Extractor.

OutWit Hub and Visual Web Ripper is two scrapes which can table and listed HTML table data.

According to this comparative study, we identified most of the web scrapers are often quite generic and mostly designed to perform common, simple tasks. In other words, they may appear not to be as flexible and universal as you would expect. All the web scraper developers try to make their products scrape all kinds of web pages, but we realized some web scraping software are better suited for one type of task and some are suited for another.

ACKNOWLEDGMENT

The author would like to thank Dr. L. Ranathunga, Prof. S.P. Karunanayaka and Prof. N.A. Abdullah for their support.

REFERENCES

List of Web Harvester, Data Scraper,Web Scraping Software and Tools [WWW Document], n.d. WebData Scraping. URL http://webdata-scraping.com/web-scraping-software/ (accessed 6.9.15).

Penman, R.B., Baldwin, T., Martinez, D., 2009. Web scraping made simple with site scraper. Text.

Savinkin, I., n.d. UiPath – Robotic Process Automation Software. Web Scraping.

Savinkin, I., n.d. Screen Scraper Review. Web Scraping.

Savinkin, I., n.d. Easy Web Extract Review. Web Scraping.

Savinkin, I., n.d. WebSundew Data Extractor Review. Web Scraping.

Software for Web Scraping, n.d. Web Scraping.

SysNucleus, n.d. WebHarvy Web Scraper [WWW Document]. URL https://www.webharvy.com/articles/what-is-web-scraping.html (accessed 6.3.15).

Web scraping, 2015a. . Wikipedia Free Encycl.

Web scraping, 2015b. . Wikipedia Free Encycl.

BIOGRAPHY OF AUTHORS

S.C.M. de S Sirisuriya is a lecturer at General Sir John Kotelawala Defence University. She received her B.Sc. Degree in Computer Science from the University of Sri Jayewardenepura. She completed her Master of Computer Science degree from University of Colombo School of Computing. She is interested in the field of E-Learning content evaluation and reads her MPhil degree on that area.