IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

WEB SCRAPING OF LINKEDIN

[1]Satish Kumar B, [2]Dr. Mohan Aradhya
[1]Student, [2]Assistant Professor
[1]Department of Master of Computer Application,
[1]RV College of Engineering, Bengaluru, India

Abstract: There is a growing need to collect data from the website. In attempting to execute the projects related to data such as price tracking, market analysis, the data from the website will also have to be registered. However, it is out of fashion to copy and paste the data line by line. The Web Scraping project utilizes automation and computing; it uses custom selection criteria to sort Linkedin profiles beyond what is currently available, including position and department. By adding the ability to generate personalized indices of social media data, automated filtration provides companies with the opportunity to boost the use of rich prospective data for more effective customer analysis and targeting.[2]

Index Terms - Web Scraping, Automation

I. INTRODUCTION

Web Scraping is a guided, automated technique for extracting information from websites. Extraction can also be performed manually, but to automate the task, it is quicker, more effective, and less prone to error.[1] Web Scraping enables us to acquire non-tabular or poorly organized website data and convert it into a functional standardized format such as a.CSV file or spreadsheet. Scraping is about collecting data plainly: it can also help you archive data and track changes to online data.[2] To perform web scraping, the following steps are necessary. They are as follows:

- Monitoring Linkedin patterns by scraping data from client websites
- Collection for study of individual data and other correspondence (e.g. use of text mining)
- Gathering electronic organizations membership and activity data
- Archive collection of reports from multiple sites[4]

At the heart of the problem that web scraping addresses are that the web is built for users. Web pages are most often designed to view organized content. Yet during a way that loads quickly, is beneficial to someone with a mouse or touchscreen, they tend to supply content, and it appears good. The organized content is formatted using templates, surrounded by contents such as headers, making sections of it visible or hidden by a mouse click. A presentation such as this is also called unstructured.[4]

In other instances, the data displayed on a website are manually collated and do not show any structured database underlying it.[6]

Web Scraping aims at reworking similar content in a structured manner in an online platform: a database, a spreadsheet, an XML representation, etc.

Web Designers expect readers to interpret the content using prior knowledge of what a header looks like, what a menu looks like, what a next page connection looks like, what the name of a person, location, and email address is. Computers have the intuition not.[3]

II. EXISTING SYSTEM AND PROPOSED SYSTEM

In the existing system, the data extraction was performed entirely through the manual process. Manual human inspection and the process of copying-pasting may often prove irreplaceable. This technique can often be the only feasible method to be used, especially when websites are set up with barriers and computer automation cannot be allowed.

The data can be collected in the proposed system by automating the website. Client-side scripts translate the contents of the web page into a DOM tree to dynamically alter or inspect a web page. One can then extract the information from the tree by embedding a program into the web browser.

III. RELATED WORKS

The paper [1] discusses the commonly used measure which is tree edit distance to calculate the similarity of pattern matched in a tree. The only obstacle for this approach is it is time complexity, so one has to consider faster algorithms so the size of the tree can be reduced.

The paper [2] discusses extracting the data by not considering the time efficiency using string methods. The string strategies comprise of the accompanying successive advances: looking for a given example, at that point ascertaining the quantity of shutting HTML components for this example, lastly separating content for the example.

The paper [3] discusses the DEiXTo, which is a web extraction suite that provides arsenal features that aim at designing and deploying the extraction tasks. The DEiXTo focuses on core pattern matching algorithm and architecture which allows programming for custom-made solutions for extraction tasks.

The paper [4] discusses the extraction procedure called EXCTVS which considers tag and value likeness. EXCTVS extracts data from result pages by recognizing and segmenting the query result records based on its tag and value similarities.

The paper [5] discusses the survey of the Web data scraper process, Web data scraper tools to provide qualitative analysis of the web data.

The paper [6] addresses data extraction from news sources as well as steps in data extraction without manually copying and pasting. There are 3 steps for the method i.e., examining website layout, building a regex pattern, and applying the regex pattern as a set of rules in scraping.

The paper [7] discusses the use of surveillance tools to detect malicious web scraping activity.

IV. PROPOSED METHODOLOGIES

The working process of the application is simple and straightforward. As soon as the script is being executed, the system logs in to the LinkedIn website and applies the following filters i.e., Location: USA and Services: Analysis. The proposed methodology extracts the member's data from LinkedIn using selenium and is stored in the form of HTML tags. The contents on the webpage can be parsed into the DOM tree. By embedding a program into the web browser, one can retrieve the information from the DOM tree. The HTML tags can then be removed using Beautiful Soup. Then the extracted data is saved into the CSV format using python. The manually extracted data and the automated data are compared to check the accuracy of the extracted data.

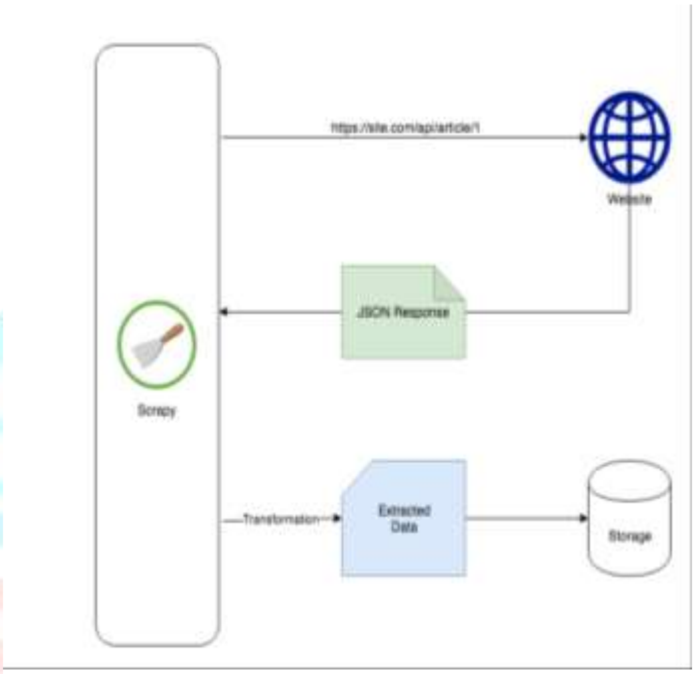


fig 1 architecture diagram of the proposed system

Figure 1 shows the entire flow of the proposed system from logging in into the system than applying the filters then storing the data in the CSV format and also displaying the result in the tool.

There are two web scraping extraction methods they are: DOM-based and String-based. DOM-based methods build a DOM tree for a web page and search in the tree for a particular element.

A web engineer can include, change and erase all components on a webpage in an internet browser by using JavaScript dependent strategies on the DOM tree. This structure is normally based on parsers of programming dialects in that errand. For the most part, extraction methods used to scrape the Internet by crossing the DOM tree.

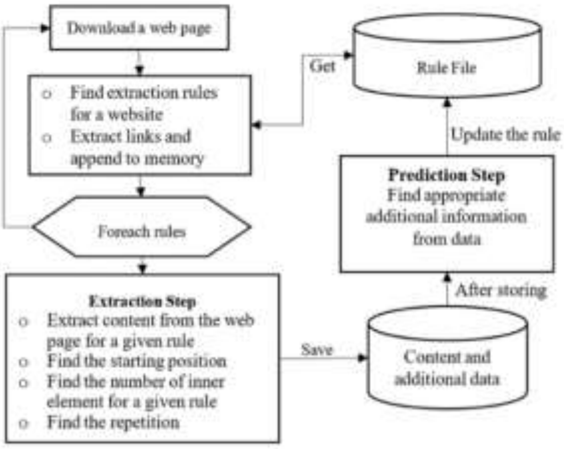

Download a web page
Find extraction rules for a website
Extract links and append to memory
Foreach rules
Extraction Step
Extract content from the web page for a given rule
Find the starting position
Find the number of inner element for a given rule
Find the repetition
Get
Rule File
Update the rule
Prediction Step
Find appropriate additional information from data
After storing
Save
Content and additional data

fig 2 flowchart of the process

The three methods described above are based on the DOM tree that is convenient to manipulate all elements. For web scraping very few elements are required. Regular expressions which are a well-known text processing technique can be considered as a solution for obtaining the content of certain elements.

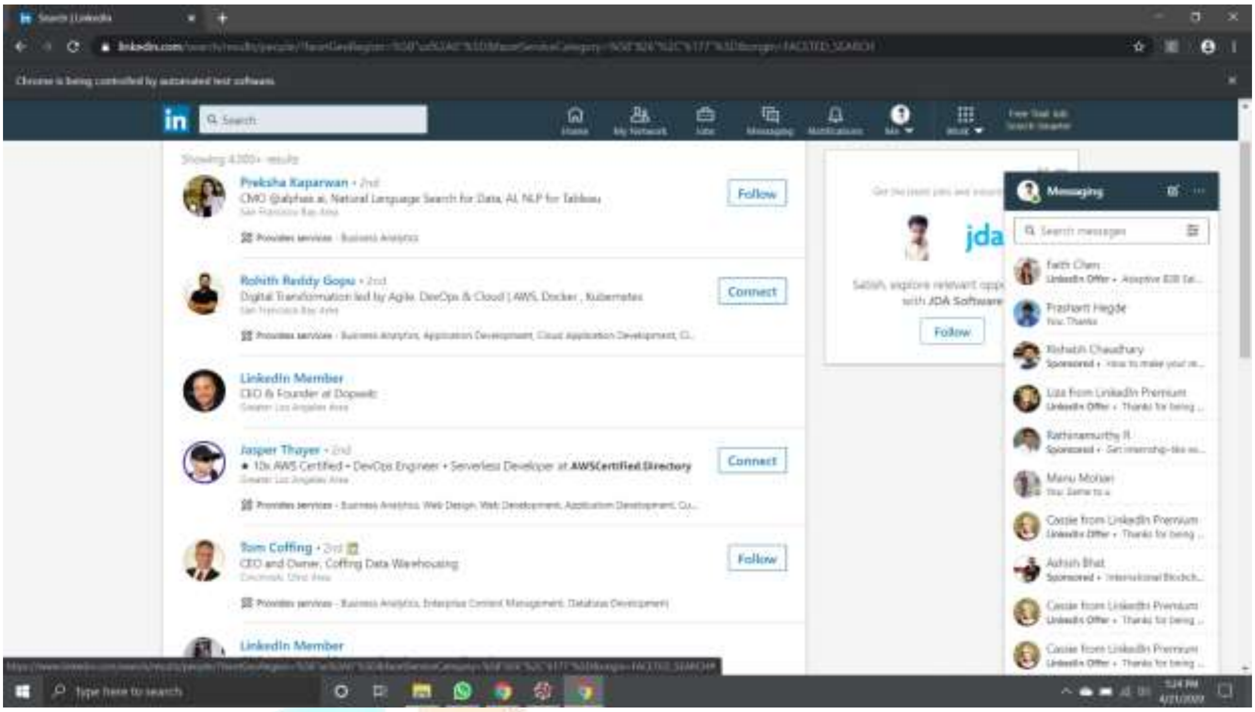
The proposed system gathers data of the various members who are in the highest position of the analytics department and also gathers those companies where jobs are open for the analytics in US location.

V. CHALLENGES

- This section discusses possible challenges or blockers that can occur while developing the application
- Extracting relevant information of various organization can be challenging, hence having a clear idea of which data is necessary
- Extracting more than 1000 people data would be challenging because Linkedin blocks the user who is trying to automate it
- There should be enough knowledge of the technology which is required for executing the script. Learning and implementing would be a challenging task for the new-comer

VI. RESULT

This section displays a Linkedin profile of the people from the various organizations in the US location. After this the data is being converted into a DOM tree with HTML tags. With the help of Beautiful Soup language, the data can be extracted by eliminating the HTML tags.



Chrome is being controlled by automated test software.

Showing 4,300+ results

Preksha Kaparwan • 2nd
CMO @alphaa.ai, Natural Language Search for Data, AI, NLP for Tableau
San Francisco Bay Area
Provides services · Business Analytics
Follow

Rohith Reddy Gopu • 2nd
Digital Transformation led by Agile, DevOps & Cloud | AWS, Docker , Kubernetes
San Francisco Bay Area
Provides services · Business Analytics, Application Development, Cloud Application Development, Cl...
Connect

LinkedIn Member
CEO & Founder at Dopwdc
Greater Los Angeles Area

Jasper Thayer • 2nd
★ 10x AWS Certified • DevOps Engineer • Serverless Developer at AWSCertified.Directory
Greater Los Angeles Area
Provides services · Business Analytics, Web Design, Web Development, Application Development, Cu...
Connect

Tom Coffing • 2nd
CEO and Owner, Coffing Data Warehousing
Cincinnati, Ohio Area
Provides services · Business Analytics, Enterprise Content Management, Database Development
Follow

LinkedIn Member

Satish, explore relevant opp... with JDA Software
Follow

Messaging

Faith Chen
LinkedIn Offer • Adaptive B2B Sal...

Prashant Hegde
You: Thanks

Rishabh Chaudhary
Sponsored • How to make your re...

Liza from LinkedIn Premium
LinkedIn Offer • Thanks for being ...

Rathinamurthy R
Sponsored • Get internship-like ex...

Manu Mohan
You: Same to u

Cassie from LinkedIn Premium
LinkedIn Offer • Thanks for being ...

Ashish Bhat
Sponsored • International Blockch...

Cassie from LinkedIn Premium
LinkedIn Offer • Thanks for being ...

Cassie from LinkedIn Premium
LinkedIn Offer • Thanks for being ...

Fig 3 Linkedin profiles

Figure 3 shows the member's data who are in the highest position in the analytics department of the various organizations in the US location.

VII. CONCLUSION

Contingent upon the main role, distinctive Web Scraping strategies can be utilized, taken measure of information, periodicity, and required result into thought. Web scrubbers have an expansive determination of instruments to choose from. The venture doesn't comprise just from the specialized arrangement and execution. Information hosts ought to consistently assess the advantage scrubbers can give and adopt a commonsense strategy to the individuals who scratch their information. Web Scrapers should hold the association with the Data has and permit recognizable proof of the Data Host as a wellspring of introduced data.

REFERENCES

[1] Vasani Krunal A, “Content Evocation Using Web Scraping and Semantic Illustration” in 2014 IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IX

[2] ERDINÇ UZUN, “A novel web scraping approach using the additional information obtained from web pages” in 2016 IEEE

[3] Fotios Kokkoras, Konstantinos Ntonas, Nick Bassiliades “DEiXTo: A Web Data Extraction Suite” in 2013 September Thessaloniki, Greece

[4] Vijay R. Thombare*, Shailesh Patil, “Web Scraping Based on Tag and Value Similarities” in 2019 April Ignited Minds Journals E-ISSN: 2230-7540 Volume: 16 / Issue: 5

[5] Sneh Nain, Bhumika Lall “Web Data Scraper Tools: Survey” in 2014 International Journal of Computer Science and Engineering

[6] Achmad Maududie, Windi Eka Yulia Retnani, Muhamat Abdul Rohim “An Approach of Web Scraping on News Website based on Regular Expression” in 2018 The 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)

[7] Pedro Marques, Zayani Dabbabi, Miruna-Mihaela Mironescu, Olivier Thonnard, Alysson Bessani, Frances Buontempo, Illir Gashi “Detecting Malicious Web Scraping Activity: a Study with Diverse Detector” in 2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC)