DATA COLLECTION USING WEB SCRAPPPING WITH ROBOTIC PROCESS AUTOMATION

Komal Tathe*1, Sanjot Sharma*2, Prof. Dr. Jyoti Kharade*3

*1,2Student, Master Of Computer Application, Bharati Vidyapeeth’s Institute Of Management And Information Technology, Navi Mumbai, Maharashtra, India.

*3Associate Professor, Bharati vidyapeeth’s Institute of Management and Information Technology, Navi Mumbai, Maharashtra, India.

ABSTRACT

In the today's digitalized world, data is the most essential asset and may cause significant impact. The challenge is to complete appetite of data. Data can be access from many different sources. The most usable source of data is website. Businesses tend to look for a competitor's website or specific website for valuable information. Web scraping use cases can be applied to see the industry trends, customer feedback about a particular product or service web scraping proves to be your smart decision-making tool. But it requires careful attention to choosing the right tools, languages and programs [1]. There is a lot of research conducted on the data collection methods. This research presents the effective use of robotic process automation in data collection. The research methodology used is assists with time saving, efficient and quick way of gathering information. This research incorporates an illustration of data extraction from website for work opportunities and an approach to automatically save produced data in the CSV design information document utilizing UiPath automation tool. This method will be valuable to small scale industries because of its simple execution strategy and simplicity of exportation of information in required design.

Keywords: Robotic Process Automation, UiPath, Data Collection, Web Scrapping, Information, Extraction.

I. INTRODUCTION

According to the study of “Robotic Process Automation Market in India, 2025” RPA market in India will grow at above 20.0% during the forecast period 2019-2025. As mentioned in “The social economy: Unlocking value and productivity through social technologies” report publish by McKinsey, just 39% of worker’s time is spent for job explicit assignments, 28% is spent perusing and noting email, 19% for searching and gathering data and the excess 14% being utilized for conveying and teaming up inside. [2] Total of these tasks covers the 61% of total working hours, this can be make self-regulating with the help of robotic process automation.

A comparable circumstance is experienced when a specialist searching for information on sites the problem of downloading the data, and can have only solution to manually note down data the time spent doing this can be very lengthy, from hours to days. There comes the need of web scrapping. Web scrapping is the method created by RPA to extract the any type of data automatically. With the available features of RPA one can save the collected data according to their choice in any format.

II. METHODOLOGY

UiPath studio is open source software (community version) used for developing the RPA bots. It provides advanced automation tools including recording, web scrapping etc.

Steps

1.1 Open UiPath studio create new project. Give proper name to the project and click on create.

1.2 Create new sequence and start creating the bot. Open Naukri.com (here used as a platform to collect the data related to job vacancies) in the browser and search for a job.

1.3 Now, start the data scraping process. Go to the UiPath studio Design tab and click on “Data Scraping”. One pop up will be open click next.
1.4 Select the job title, as the first column of extracted data is job title. "Select second element" pop up will open, click on Next button and select the second job title. This is requiring to create information pattern in UiPath studio.

1.5 Give proper column header name and click on Next. You will see the extracted data next. In this section, as a result maximum rows present on the website can be extracted. To extract more information such as company name, location, salary etc. click on extract correlated data and follow the step 2.4. After completing click on finish.
1.6 Now, the sequence is ready and project is completed. To save the extracted data in CSV format from the activities take the "Write range" activity on the wok space of the sequence and use it.

1.7 To run the bot go to the design tab click on debug file and run the project. The CSV file is stored under the specified location and can be checked later.
Figure 6: The data set extracted from the jobs posted on the Naukri.com platform in excel format (with Job Title, Company Name, Package and City columns)

III. MODELING AND ANALYSIS

The term Robotic Process Automation (RPA) was first used in 2012 (Hindle, 2018) [7]. It began to gain popularity in 2014 and 2015 when companies started to announce considerable savings due to automation. According to the global RPA market which includes both RPA services and RPA software increased by about 64% from 2016 to 2017 and reported a 42% increase in the market from 2017 to 2018 and an expected increase around 94% from 2018 to 2021 (Fersht, 2017) [9]. Some scholars and consultancies claim that RPA is just one step on the way to more intelligent and cognitive automation (Hull and Motahari-Nezhad 2016; KPMG 2016; Lacity and Willcocks 2016b; Van der Aalst et al. 2018). [10] [11] [12] [13][14] conducted a study on how Robotic automation process will help to save manual exertion and time. on completion of deployment of the robot they have monitored the time taken by the robot to perform the particular task. As per their observation the total time taken by the robot was only 43 secs to read a mail and parse the appropriate text and store it in the database without any human intervention whereas if manual effort is utilized for the same then time can consume minimum 5mins to complete the same job. Now they tried to multiply the mail volume by 100 times it would save a lot of time, effort and money. [14] concludes that Robots are good at extracting, processing and integrating information but they do not really understand the information. As per their result, it cannot read or parse the text which is in an unstructured manner. Hence, there is a considerable measure of research work are occurring in cognitive document automation.[15] suggested that RPA should be used strategically while focusing on its long term influence in the process. [16] introduce the research and application field, some common techniques used in RPA Data Scrapping and their preliminary results and conclusions. [16] described general business processes for documents flow and current approaches to automating them & also presented our baseline solutions for a specific business process. They want to take action on more complicated use-cases, like responding to a client via e-mails based on collected information from Basic Processing and Information Extraction modules.[17] hints that RPA technology can be used with minimal programming skills. The implementation of RPA is a rapid process and it does not require re-engineering of existing business processes.

IV. RESULTS AND DISCUSSION

RPA technology shows remarkable speed and accuracy doing repetitive tasks. RPA technology also shows notable results completing complex tasks while maintaining ease of use and simplicity for users. The bot developed in this study is proven effective in data collection as compare to the humans. With assistance of devices like UiPath, RPA can be developed and utilized by standard clients with no programming experience. RPA innovation is truly versatile and adaptable because of its simple development process. RPA innovation can deal with complex assignments with appropriate execution.
V. CONCLUSION

Robotic process automation as a technology automates existing tasks. It’s just training software to perform repetitive tasks. The current system is unaffected by the introduction of RPA, so there is no need to modify the existing system.

The bot performed in this study showed the effective results. Data gathered by the RPA bot is almost precise. The work that takes 19% of employees time, the bot developed with RPA done that in much less time as compare to human. The activities like log messages, included directly into the sequence to improve the code readability. In RPA Simple workflow can handle in fast and precise way while more complex workflow requires more work to be done.

VI. REFERENCES