IOP Conf. Series: Materials Science and Engineering **732** (2020) 012071 doi:10.1088/1757-899X/732/1/012071

Automatically collect alumni data on social media

R Ariyanto, C Rachmad and A R Syulistyo*

State Polytechnic of Malang, Malang, Indonesia

*arie.rachmad.s@polinema.ac.id

Abstract. Tracer studies are useful for universities, institutes, and academies, as well as scholarship providers such as government, companies, and institution. Tracer studies are used to track alumni which can be used to measure the relevance of objectives, educational processes and curriculum in accordance with current alumni conditions. However, the tracer study has several obstacles in the distribution process to reach out and make alumni willing to spare their time for filling the tracer studies. This study aims to overcome this problem by gathering the information needed from social media such as workplaces and professions. The results of this study is alumni data which scrap from social media.

1. Introduction

Tracer studies are carried out to track alumni to measure the relevance of educational goals and processes in accordance with the current condition of alumni [1]. In general, tracer studies are carried out by universities, institutes, and academies, as well as scholarship providers such as government, companies, and foundations.

Tracer Study is useful for educational entities measuring the relevance of educational curriculum to the work environment of alumni [2]. As for the scholarship provider, it is useful to measure the relevance of the policies applied to the graduates of the program. For the government, tracer study is useful to evaluate existing processes in tertiary institutions so that it is expected to be able to improve existing standards in tertiary institutions. This paper will focus on tracer studies applied to educational entities, especially the Information Technology Department in State Polytechnic of Malang, for alumni of the Information Management program and the Information Technology program.

However, tracer studies have several obstacles in the distribution process to reach out and make alumni have willingness to take the time to fill out tracer studies [3]. There are several ways to improve alumni response to tracer studies by sending SMS (Sort Message Service), contacting alumni via telephone, and using technology [4]. The staff of the Information Technology Department has made an effort to improve response rate of alumni by creating an online tracer study so make it easier for JTI alumni to fill in their latest personal data.

This study aims to overcome this problem by making a tracer study using social media analysis so that the workplace of the alumni of JTI is known. By using several methods including scraping methods, text mining, sentiment analysis to get data from social media and process the data that has been obtained. In the end, by using existing methods in the information system, a system will be created using the methods that have been developed. After the information system is complete, the results of this research will be implemented at the Information Technology Department in Malang State Polytechnic.

2. Background study

2.1. Scraping

Scrapping is collecting or extracting data from social media and certain website in the form of unstructured text [5,6]. One example of web scraping is copying contact lists from a web directory. This can be done manually by copying and pasting the data to Excel. However, if the data is processed a lot, then the manual process becomes ineffective. Therefore, we need an automatic process that can help the web scraping process.

Web scraping can be done using web scraper, bot, web spider, or web crawler. The way the web scraper works is by going to a web page, then downloading the content, then extracting expected data from the content, and saving the data to one file or database in accordance with the configuration of the system.

2.2. Selenium

Selenium is automation test software on web browser and open source tool. Although selenium used to test software however the function does not limit to that which can be used to scrap the content of the website. The scraping process can be seen on Figure 1 which take from published paper Cloud Based Web Scraping for Big Data Applications [7], illustrated the scraping process using selenium and request library where user give URLs to scrape with certain configuration that located on Scrapehub on Figure 1. This model has capability to scrap one site at one time which will be used to scrap alumni Facebook group and LinkedIn in this paper.

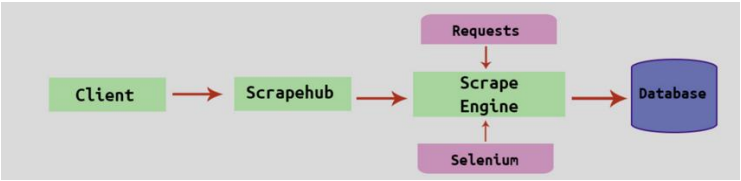

Figure 1. The illustration scraping process.

Selenium does not charge the user which is the major advantages and another reason Selenium gets its popularity because support more than one programming language such as Java, Python, C#, PHP, Ruby, Perl and .Net [8]. Furthermore, selenium supports many platforms such as Windows, Mac and Linux and supports many browsers such as Mozilla Firefox, Internet Explorer, Google Chrome, Safari and Opera.

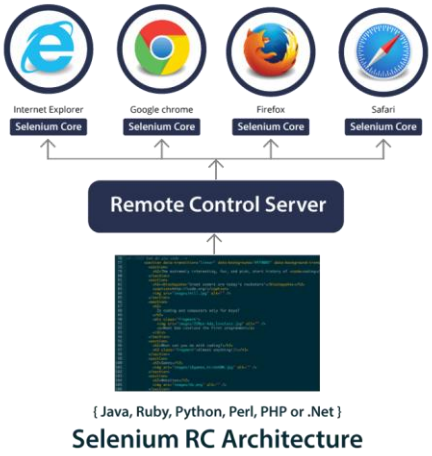

Figure 2. The Selenium illustration works on many browser.

3. Proposed method

This research is part of the big system of smart tracer study, the position of this research is at the beginning of smart tracer study which can be seen on Figure 3.

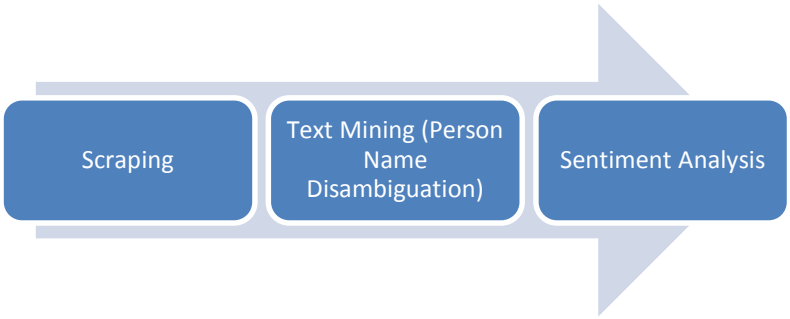

Figure 3. Position of this research based the whole system that will develop.

The scraping process is integral part for the next process because by scraping process this system get expected data. This scraping process will focus on two social media such as LinkedIn and Facebook. The scraping process will get the employee history records of each user whom alumni of department of information technology, state polytechnic of Malang. The future research will focus on text mining which used to make sure no double alumni data based on LinkedIn and Facebook scraping by using the person name disambiguation technique [9]. The final step of this research will focus on sentiment analysis [10] used to find out the status or comment meaning from alumni in social media that represent the alumni changed work position or not.

3.1. Scraping implementation

Based on explanation at subchapter 2 this research use selenium library, parsel library and python as the main programming language. The flow of scraping process can be seen on Figure 4, the first step is preparing the data which is needed by selenium and parsel library such as destination URLs, username and password. After that the system works as illustrate in flowchart on Figure 4.

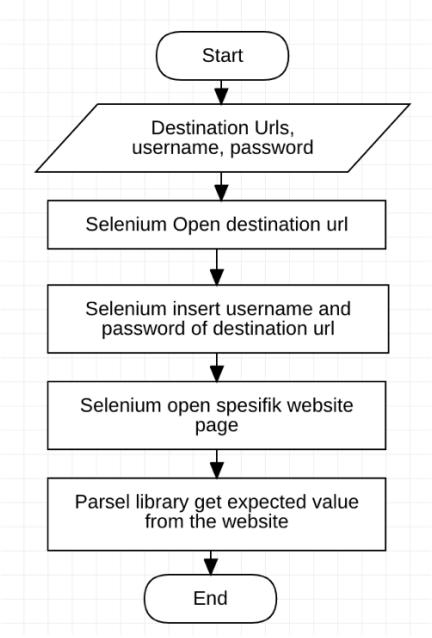

Figure 4. Flowchart scraping process.

4. Implementation and analysis

Th one of implementation of the scraping process is Facebook scraping can be seen in Table 1, code in line number 1 intend to import the parameter for setting username and password of destination URLs. Then selenium library will open the destination URLs, and login by inserting password, username and click the login button this action based on code that written respectively in line number 13,17 and 21. After that parsel library will get value from page that selenium has opened by using code in line 31, 34 and 35. The final result of this process will save on csv file based on code in line 45.

Table 1. Source code facescrap.py.

facescrap.py

```
1   import faceparam
2   from time import sleep
3   from selenium import webdriver
4   from selenium.webdriver.common.keys import Keys
5   from parsel import Selector
6   import csv
7
8   driver = webdriver.Chrome('/Users/arie/Downloads/chromedriver')
9
10  driver.get('https://www.facebook.com/')
11  sleep(0.5)
12
13  username = driver.find_element_by_name('email')
14  username.send_keys(faceparam.facebook_username)
15  sleep(0.5)
16
17  password = driver.find_element_by_name('pass')
18  password.send_keys(faceparam.facebook_password)
19  sleep(0.5)
20
21  sign_in_button = driver.find_element_by_id("u_0_b")
22  # driver.find_element_by_xpath('//*[@type="submit"]')
23  sign_in_button.click()
24  sleep(5)
25
26  driver.get('https://www.facebook.com/groups/1719628528289629/members/')
27  sleep(10)
28
29  #get Name and description
30  sel = Selector(text=driver.page_source)
31  profile = sel.xpath('//*[starts-with(@class, "clearfix _60rh _gse")]')
32  description = profile.xpath('//div[starts-with(@class,
33  "_60rj")]/text()').getall()
34  namePath = profile.xpath('//*[starts-with(@class, "_60ri")]')
35  name = namePath.xpath('//a[starts-with(@data-hovercard-prefer-more-content-show,
36  "1")]/text()').getall()
37
38  print(name)
39  print(description)
40
41  myFile = open('/Users/arie/Downloads/results_file.csv','w')
42
43  with myFile:
44      writer = csv.writer(myFile)
45      writer.writerows([['Name', 'description'], [name, description]])
46
47
48  driver.quit()
```

The result of this process of scraping process on Facebook and LinkedIn can be seen Figure 5 and 6. By using selenium and parsel, this process is success to get the expected result However, the results of Facebook scraping are still needed preprocess to clean data because it contains unexpected data from users such as those who add them to Facebook groups or descriptions when they join Facebook groups.

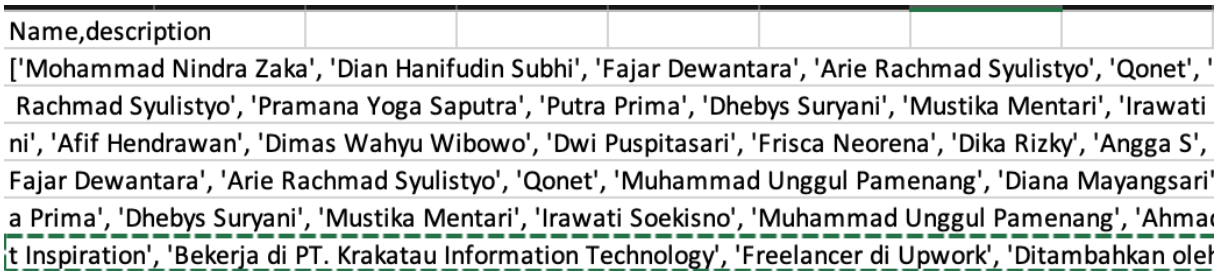

Figure 5. Screenshot of Facebook scraping.

Figure 6. Screenshot of LinkedIn scraping.

5. Conclusion

Web scraping is one of the solutions to collect alumni profiles which were successfully carried out by using selenium and parsel library and make this research easier to do. The result of this research is data which save on excel in csv format. The future research will focus to process raw data from scraping result.

References

[1] Schomburg H 2010 Concept and Methodology of Tracer Studies–International Experiences *Presentation at Workshop in Sinala* 2-4

[2] Wibisono A, Ulama B S S, Asmoro W A and SAC S A C 2012 Tracer Study at Institut Teknologi Sepuluh Nopember (ITS), Promoting Localization and Multiple Touch Points to Capture Alumni *International Conference on Experience with Link and Match in Higher Education: Result of tracer studies world wide, Bali, Indonesia*

[3] Toba H and Wijaya E A 2017 Enhanced unsupervised person name disambiguation to support alumni tracer study *Global Journal of Engineering Education* **19**(1) 42-48

[4] Mwizerwa J R W N C 2017 Improving Response Rates to an Alumni Survey in East Africa *Advances in Social Sciences Research Journal* **4** 120-127

[5] Batrinca B 2015 Social media analytics: a survey of techniques, tools and platforms *AI & SOCIETY* **30** 89-116

[6] Raulamo-Jurvanen P, Kakkonen K and Mäntylä M 2016 Using Surveys and Web-Scraping to Select Tools for Software Testing Consultancy *International Conference on Product-Focused Software Process Improvement* 285-300

[7] Chaulagain R S, Pandey S, Basnet S R and Shakya S 2017 Cloud based web scraping for big data applications *2017 IEEE International Conference on Smart Cloud (SmartCloud)* 138-143
[8] Vardhan 2019 Edureka [Online] Retrieved from: https://www.edureka.co/blog/what-is-selenium/ [Accessed 24 September 2019]
[9] Yoshida M, Ikeda M, Ono S, Sato I and Nakagawa H 2010 Person name disambiguation by bootstrapping *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* 10-17
[10] Jain T I and Nemade D 2010 Recognizing contextual polarity in phrase-level sentiment analysis *International Journal of Computer Applications* **7**(5) 12-21