International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 6, Issue 12, 2019

Web Scrapping as Pre-Processing Method to Identify Current Careers and Trends in Information and Communication Technologies

Eka Angga Laksana[a], [a]Informatics Engineering,Widyatama University Bandung, Indonesia, Email: Eka.angga@widyatama.ac.id

Rapid growth of the internet has generated change in information and technology industries. Informatics departments as part of higher education studies have responsibility to create curriculum which adapts to industrial needs. This paper shows the potential information hidden in the internet which, by using web scrapping method, has enabled academics to automatically collect careers information in ICT industries. The experiment has successfully collected data regarding 800 jobs related to ICT industries. This study is important as pre-processing of the data mining process as it would make it easier for the next research to find the current trends in information and Communication Technologies.

Key words: *component, formatting, style, styling, insert.*

Introduction

Higher education is the next level of education after graduation from high school. It takes place at university or another education college and commonly includes undergraduate and post graduate pathways. Currently, employability of graduates is an issue for higher education institutions (Kumar and Khurana, 2017). Employability depends on subject and curricula chosen by the student. Higher education gives students a chance to choose subjects appropriate with their passion to increase career prospect and reveal their potential. A lot of companies post job listings to under graduate student and higher education institutions prepare students to fulfil these requirements. Informatics engineering is a part of a University program which combines two disciplines, Information Technologies and engineering. This study emphasized information processing combined with computer technology which is called Information Technology and aims to achieve improved quality of human life in

economic, social and medical areas. Currently, Informatics engineering programs are quite successful at attracting high school graduates who seek to develop their skill in Information Technology.

Careers in informatics industries become important in the last decade (Century, 2017). Employment in computer and information technology jobs is projected to increase 13% from 2016 to 2026, faster than the average for all industries. This means that the occupation is expected to add about 557,100 new jobs according to bls.gov. These demands will in particular, effect some informatics areas such as cloud computing, big data, information security and web development. Bls.gov also reported that the median annual wage for informatics or ICT occupation was $86,320 in May 2018. It was higher than all occupation other jobs which reach $38,640.

In contrast with the above fact, informatics graduates still found it hard to find jobs(Shadbolt, 2016)[3]. According to theregister.co.uk, computer science graduates placed on the top list of unemployability rankings for the past several years as for some reason, there is a skill gap in in the informatics discipline between higher education and industry with graduates told they are not fit for the market as employees. According to Shadbolt, there are three categories in the informatics discipline: computer science, software engineering and IT.

Higher education students are advised to seek the right career opportunity by examination of what kind of person they are, what kind of life they want and determine their real life goals. Career research is one step before making any career decision – the following are some suggested research areas:

1. Individual skill and employee requirement
2. Future choice with student's subject/qualification
3. Job requirement that applicants hard to fill
4. Job advertisement location
5. Required skill for now and likely to be in the future
6. What kind of work and life after graduation?
7. How career centre in higher education could help you

Creating best possible course material for higher education is a key career decision path for students to make and is often challenging. There is common competition between popular courses and later for future graduation jobs. There are some courses in informatics engineering which could be tough in different ways. Therefore, choosing the right material to make better skill to the student is important task, helps students increase their career opportunity in informatics. This study aims to construct careers in informatics technology through scrapping the job vacancy website. It may provide a dataset for further research to

find the patterns and solve the problem of employability in informatics industries. For this purpose, this study addresses the following question:

1. What are the best tools to collect job's data from the target web?
2. What is the configuration needed to make web crawling works?
3. How should a job's dataset be constructed based on the crawling result for further research?

The internet gives many possible tools and choices to find a job in the informatics area. Higher education can identify job requirements all over the world, find and learn files of worker literature, exchange information with professionals in informatics, share ideas among groups and get some advice on how to write resume and face interview etc (cite). In modern times, many companies use the internet recruiting tools. The internet also provides job vacancies which are freely accessed by people rather than paying for information. Below the benefits of using the internet to find a job are listed:

1. Real time job information
2. Access to specific career information
3. Communication with professional in informatics area
4. More information about companies and organization
5. Capacity to advertise skill to fulfil career position.

Some related research has using questioning technique to detect the trend of job or career information. However, according to surveyanyplace.com there are several disadvantages to the use of these responses to make decisions, which are:

1. Sometimes respondents give dishonest answers
2. Some questions will be ignored or unanswered
3. People may have different interpretations
4. A question cannot fully capture answers related to feeling or emotional
5. Participants may have hidden agendas
6. Lack of personalization
7. Unconscientious responses
8. Accessibility problems
9. Loss of concentration while facing questionnaire

The weakness described in previous paragraph shows that finding careers information by using questionnaires is not effective to define current job trends in the informatics department. One purpose of this research is to solve this problem, as the internet contains a large volume of information and by using web scrapping methods, it can help higher education institutions to understand the current condition of careers in informatics.

Web Scrapping

Web scrapping or web data extraction is a process to extract data from websites (cite). Web scrapping uses a tool or software to access any web site using HTTP protocols. Web scrapping is controlled manually by a user or programmer and can be fully automated in implementation and known as a bot or web crawler. The purpose of crawling is to retrieve specific information from raw html code. Humans can't read large html code while web crawlers extract data or information specifically required by a user. Typically, the extracted data is stored in a database or another format, for later process or analysis. Generally, web scrapping helps people to:

1. Retrieve HTML specific domain name
2. Parsing html to select contained information
3. Store found information
4. Optionally, walk through another page and repeat the process

Finding career information through internet as a human task is often difficult due to retrieval of whole information page to page and from one website to another website and then storage in a database. This research proposes a tool to help higher education institutions, especially in informatics departments, to collect recent current, global career information. This information will create the global picture of what companies needs and then generate appropriate subject and learning material to the students. Web scrapping also helps student to do some research on their future career path and motivates them to improve particular skill in the informatics area. Moreover, with jobs information in databases, data mining method can be applied to find patterns and current trends in informatics careers.

Data pre-processing is a data mining technique which involves transforming raw data into feasible format for further processing (Witten, 2011). Real world data sometime consist of incomplete, inconsistent behaviour and lack of behaviour where many errors occur. Data pre-processing is a method to solve such issues. Data pre-processing prepares raw data for further processing through data mining. In order to facilitate finding career information, raw career data has some issues. This is where data pre-processing technique can be taken into account. Changing raw data related to job information into a readable format for further processes such as the use of the data mining technique.

Information extraction (IE) is the automated retrieval of certain information correlated to a particular topic from a body or bodies of text. There are information extraction tools in python programming language which make it possible to retrieve information from text documents, databases, websites, or other sources. Information retrieval may extract

information from unstructured and semi-structured structured text into machine-readable text. Information Extraction is commonly used in Natural Language Processing (NLP)

Methodology

The mining research domain has a sufficient amount of data which can be processed to find certain patterns and relevant information. Acquiring the data is only the first phase. The data is often collected in unstructured form and needs to be transferred into structured format for feasible processing. In this paper, we use crawling to collect the data from two websites, they are: Id.Jobsdb.com, Jobstreet.co.id. Information stored from crawling processes is useful as a pre-processing phase in data mining. Further data mining process to reveal knowledge to help informatics department close the gap with industrial will be discussed the next paper. In this paper, we divide the process into several step below:

a. Select URLs to crawl
b. Fetch and parse each page
c. Save important content into python list: job description content
d. Save content list into flat file for further process.

BeatifulSoup Library

In this research, python programming language is used with Beautifulsoup Library to do web crawling. Beautifulsoup library is useful to convert the messy HTML code into friendly-traversable python object format represented by XML structure. There are three pillars of web design: HTML, CSS and javascript. Most of modern websites today contain those elements. HTML code and others element are made readable by web browsers, not by human. However, web scrapper tools rely on CSS styling to differentiate the HTML element in order to retrieve information in human friendly form. For example, some html tag might look like this:

 or

Beautifulsoup can easily be separated into two different tags based on their id. The program can choose only 'top-left' text but none of the 'bottom'. CSS depends on these attributes to apply styling on HTML elements appropriately and is in place in most modern websites.

This is the example code to scrapping a web site

```
Import request
webdriver.get('http://ekaangga.net')
page = requests.get(url)
```

```
soup = BeautifulSoup(page.text, 'lxml')

page1 = []
for i in soup.find_all('div', attrs={'class':'post content'}):
    for j in i.find_all('a', attrs={'class':'post-title'}):
        print j['href']
        page1 += [j['href']]
```

In the above example, the web scrapper targets a website with address: ekaangga.net. First, the analyst must identify which part of the information will be gathered. The web scrapper targets title of each post and then extracts the link behind it. Website links are crucial because they can reveal full content when opened by a web browser. Then page1 variable is created as an array to store each founded link and links or urls are stored in page1 and used for further analysis to retrieve post content in each of them.

Results and analysis

For research purpose, two job vacancy websites as jobs list description were targeted. These websites were popular and reputable within jobs seekers in Indonesia. There are many other job vacancy websites in Indonesia from which a potential amount of data could be gathered, and it is suggested this be a recommendation for further research. In this research, we make sure the method could work properly to gather enough job description data. Once the model has been built and proven, it can be used for further research.

The experiment successfully obtained a total 800 jobs description crawled during April 2019. Both python library, and Beautifullsoup took jobs content from the websites smoothly as shown in Table 1 below.

Table 1: Experiment result

No.	site	Jobs crawled
1	Id.jobsdb.com	600
3	Jobsstreet.co.id	200

After the crawling process was finished, the data was saved into flat text file. Generally, Python uses pickle library to save array or lists into a file. A saved file can be read by the system anytime and stored into original array or as per list value. These processes will be useful for further processing. Below is part of the job content taken from the array:

- Minimum hold a Bachelor Degree in Computer Science or IT.
- Minimum 2 years experience.
- Maximum age 30 years old.

- Having knowledge of Java C/C++ / PHP.
- Having Knowledge of Oracle database/SQL
- Server/Postgresql/Mysql.
- Having Knowledge of HTML (AJAX, JSON,
- XML) and Javascript/jQuery/Bootstrap/Phyton.
- Familiar with Dreamweaver, eclipse/NetBeans.
- Familiar with CSS, Javascript, and web design.
- Have a good knowlegde in PHP framework such as
- Codelgniter or Zend.
- Experience with Android Studio and material Design.
- Be able to work as an individual and as part of a team.
- Be able to work to tight deadlines.
- Good analytical and problem solving skills.
- Ability to work extra hours as needed.
- Ability to be self-motivated and a self-learner.

From the above example, some analysis needs to be done to break each requirement into the more understandable term. Job skills are divided into two well-known terms: soft skills and hard skills. Soft skills are related to interpersonal skills which is harder to define and evaluate. Some skills belong to soft skill and include: communication, listening and empathy. According to the previous example, humans can easily interpret soft skills required by a company, such as: work as individual or part a team, good analytical problem-solving skills, ability to work extra hours and to be self-motivated/self-learner. However, further research is recommended to be conducted to find trends in soft skills specifically in the informatics industry.

In the other hand, hard skill also need to be considered as job specific skill and knowledge needed to perform a job. From the previous example, some hard skills can easily be identified by people. Holding a bachelor degree, 2 years experience, maximum 30 years old, having knowledge of java/C++/PHP are some hard skills required by companies. However, identifying trends in informatics careers industries by analyzing hundreds of record is a potential topic in further research.

In this paper, experiment results are proposed from which to build the dataset or model that id needed as preparation for the next research. Information from about 800 job has been scrapped as described in point III, each contain: job title, description, specification, company name and location. Informatics engineering students in higher education can use this basic information to find out the most recent trends in ICT technology. Moreover, by using data mining technique patterns from this career information data will be revealed.

Conclusion

Information and Communication technologies in industry has grown rapidly in recent decades (Yu et al., 2019; Var, 2018). Informatics departments are one of the higher education faculties that play an important role to ensure curriculum matches industrial needs. Web scrapping helps informatics department course writers gather information across the internet to provide the students with the recently researched skills which will help them to face their career after graduation. The experiment has successfully crawled information from about 800 jobs, each of them consist with provision of: job title, company name, job specification, requirement and location. However, to study the current trends in job skills and requirement is a topic for further research. A research plan involving some data mining techniques created to find patterns in ICT careers will be the subject of a further paper.

REFERENCES

Kumar, R. and Khurana, K. (2017). Sciencedirect employability skills among information technology professionals : A literature review. Procedia Comput. Sci., 122: 63–70.

Shadbolt, S. N. (2016). Shadbolt Review of Computer Sciences Degree Accreditation and Graduate Employability, no. April, 2016.

Witten, I. H. (2011). Data mining : Practical machine learning tools and techniques. 3rd ed. Burlington: Morgan Kaufmann.

Yu, D., Ebadi, A.G., Jermsittiparsert, K., Jabarullah, N., Vasiljeva, M.V. & Nojavan, S. (2019). Risk-constrained stochastic optimization of a concentrating solar power plant. IEEE Transactions on Sustainable Energy, https://doi.org/ 10.1109/TSTE.2019.2927735.

Var, L. (2018). The analysis of teacher candidates' self-sufficiency about their teaching abilities at different departments. Asian Journal of Education and Training, 4(3): 246-249.

Century, T. (2017). Oxford Handbooks Online, no. May 2018, pp. 1–24.