IOP Conf. Series: Materials Science and Engineering **288** (2018) 012038 doi:10.1088/1757-899X/288/1/012038

Web Scraping and Naïve Bayes Classification for Job Search Engine

C Slamet[1*], R Andrian[1], D S Maylawati[1], Suhendar[1], W Darmalaksana[2] and M A Ramdhani[1]

[1]Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati, Bandung, Indonesia
[2] Fakultas Usuludin, UIN Sunan Gunung Djati, Bandung, Indonesia

*cepy_lucky@uinsgd.ac.id

Abstract. Many organisations (government of non-government) use websites to share information of new recruitment for the workers. This information overflows on thousands of sites with various attributes and criteria. However, this availability forms a complex puzzle in the selection process and lead to inefficient runtime. This study proposes a simple method for job searching simplification through a construction and collaboration of web scraping technique and classification using Naïve Bayes on search engine. This study is resulting an effective and efficient application for users to seek a potential job that fit in with their interests.

1. Introduction

Competition and growth of Indonesian citizen keep increasing rapidly. To cope with this, there needs to be more job vacancies available [1] Nowadays, in Indonesia, information of job vacancies is presented online using traditional administration system and also job fair. In addition, some of the vacancies are on certain websites that can help job seekers find information they need [2]. However, to acquire the information, job seekers usually need to browse several websites which is not effective and efficient [3].

To give easy access for job seekers in having information related to recruitment through one recourse, there needs to be a search engine which gives accurate and specific information [3]. Thus, job vacancy classification can be one of the solutions so that job seekers focus on vacancies suitable with their qualification, interests, and talents [1].

In this study, web scrapping technique and *naïve bayes* algorithm are adopted to solve the problems of random job vacancy availability so that they no longer have to browse several websites. Web scrapping will automatically collect information of the job vacancies from several websites. In the meantime, *naïve bayes* classifies those vacancies based on categories determined in advance.

2. Literature Review

2.1 Web Scraping

Web scrapping is a process of automatic data and information collection from the internet, commonly in website pages using markup languages such as HTML or XHTML whose data analyzed for certain needs and purposes [4]–[8]. Figure 1 illustrates the scheme of web scrapping technique which can be elaborated into four main processes 1) creating scrapping template: inserting scrapping template by defining HTML documents from website whose information is collected; 2) exploring site navigation:

making website navigation exploration system from websites whose information is collected; 3) automating navigation and extraction: from the processes number 1 and 2, automation from the data and information acquired from the websites is conducted; and 4) extracting data and package history: the information acquired from process number 3 is saved in tables and database.

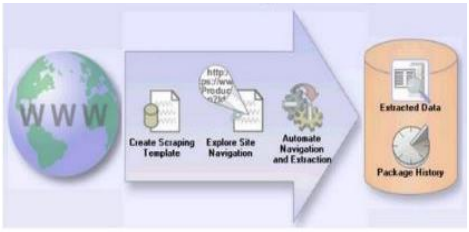

Figure 1. Scheme of Web Scrapping [6]

2.2 Naïve Bayes Algorithm

Naïve bayes is a simple probability-based prediction referring to *bayes* theorem having strong independence assumption (naïve) [3]. Previous studies on algorithm of classification have proven that *naïve bayes* is the best algorithm in comparison with the other ones such of Decision Tree, Naïve Bayes, and K-NN [9]–[11]. It has also been found that accuracy and speed are the most supporting and helpful features of the algorithm in classifying data.

The basic formula of naïve bayes algorithm, which is based on bayes theorem, is as follows.

$$P(V_i) = \frac{P(A|B)P(B)}{P(A)} \qquad (1)$$

QUOTE $\frac{\mathrm{P(A|B)P(B)}}{\mathrm{P(A)}}$ *Naive Bayes Classifier* is an algorithm with simplification model suitable with classification of data and documents. The equation is.

$$Vmap = \underset{vj\varepsilon V}{args\ max}(Vj|a1, a2, \dots, an) \qquad (2)$$

Based on the equation, the bayes formula is.

$$Vmap = \underset{vj\varepsilon V}{args\ max}\frac{P(Vj|a1, a2, \dots, |Vj)P(Vj)}{P(a1,a2,\dots,an)} \qquad (3)$$

Since the value of *P(a1,a2,…..,an)* for each Vj is similar, the value can be ommited so that the equation becomes.

$$Vmap = \underset{vj\varepsilon V}{args\ max}(Vj|a1, a2, \dots, an)P(Vj) \qquad (4)$$

Assuming that each word in <a1, a2, an> is independent, so *(Vj|a1, a2, an) P(Vj)* in equation is written as follows.

$$(Vj|a1, a2, \dots, an)P(Vj) = \prod P(ai|Vj) \qquad (5)$$

Thus, the equation is.

$$Vmap = \underset{vj\varepsilon V}{args\ max}\ P(Vj)\prod P\ (ai|Vj) \qquad (6)$$

Meanwhle, the value of *P(vj)* is determined in the training, whose value is approached by.

$$\mathrm{P(Vj)} = \frac{|\mathrm{Doc}}{|\mathrm{D}|} \qquad \mathrm{P(Wk|Vj)}\frac{\mathrm{Nk+1}}{\mathrm{n+|Vocabulary|}} \quad (7)$$

3. Methods

To give easy access for job seekers to get information related to one recourse recruitment, there needs to be a search engine that gives accurate and specific information [3]. Web scrapping technique and naïve bayes classification are believed to be one of the solutions in search engine giving easy access

collected from several websites. Figure 2 explains the sequence of information collection of the job vacancies utilizing web scrapping technique.

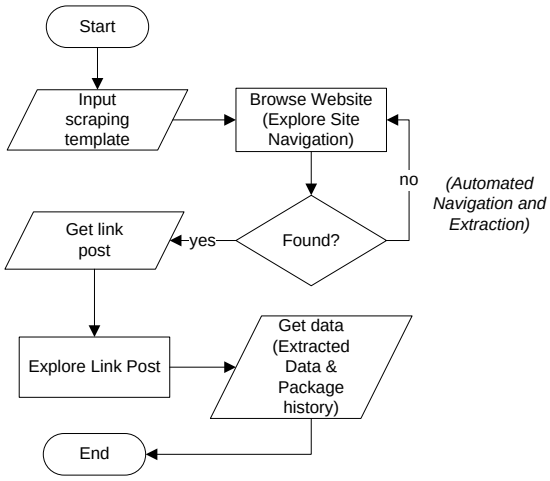

Figure 2. Proposed Sequence of Information Collection on Job Vacancies

Sequencing proposed in Figure 2 includes processes of scrapping template input, website exploration which goes through looping process of data that have not been found; if the data have already been found, the next process in linking post to be explored and extracted on the website page's HTML. This whole process is also called website page capture.

Naïve bayes algorithm is used to classify information collected. In this algorithm process, there are training and testing, in which data are used to calculate the probability of information found and categorized, while classification is done to information that is not categorized yet.
Below is a series of activities in the training session.

1. Determining documents that are categorized;
2. Forming vocabulary by sampling unique words;
3. Calculating the probability of data training using naïve bayes formula that is P*(Vj)* and *P (Wk | Vj)*.

$$P\frac{|Doc\ j|}{|D|} \qquad (8)$$

In which |*Doc j*| is the number of documents having *j* categories in the training, while |D| is the number of documents used in the training session.

$$P(Wk|Vj) = \frac{Nk+1}{n+|Vocabulary|} \qquad (9)$$

In which *Nk* is the frequency of *Wk* words in documents categorized into *Vj*, while *n* is the number of words in *Vj* documents, and |*Vocabulary*| is the number of words in documents in the training session.

Below is the series of activities in the testing session.

1. Sorting out documents that are not categorized yet;
2. Analyzing unique words correlated to the training data;
3. Calculating $\mathrm{Vmap} = \underset{\mathrm{vj}\varepsilon\mathrm{V}}{\arg\max}\, \mathrm{P(Vj)} \prod(\mathrm{ai}|\mathrm{Vj})$

In which Π *P(ai | Vj)* is a series of multiplied values of *P(Vj)* and *P (Wk | Vj)* in each *P(Vj)* category, and *arg max P(Vj)* is comparing the results of *P(Vj)* in each category, dan taking the

maximum values as *Vmap* category. Generally, the proposed working sequence of information classification collected is explain in Figure 3.

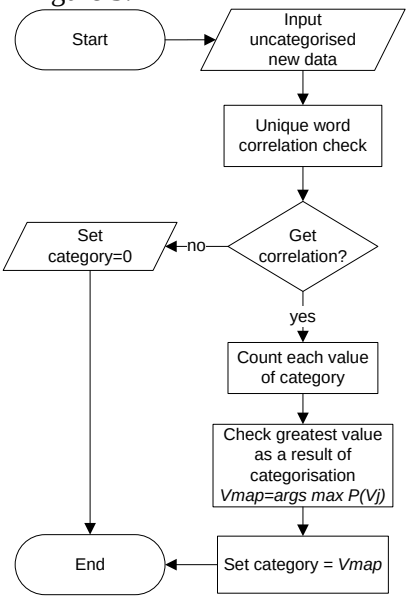

Figure 3. Proposed Working Sequence of Classification

To meet the need of job vacancy search engine that can give information from several websites classified in terms of the categories of the job, the following architectural design system is proposed.

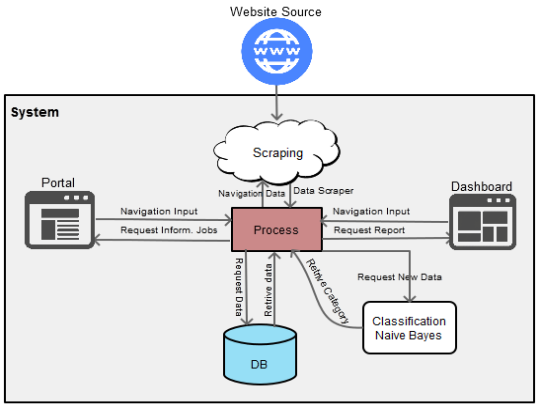

Figure 4. Proposed Architecture of System

In the proposed architecture of system, there are two sides of application namely portal and dashboard. When portal and dashboard request for navigation input, the system is processing the request. Once the results are required, the system is returning the results of the request.

In scrapping process, the system is doing automated system towards several websites to collect new information. The searching addresses are acquired through data retrieval saved in the database. After the scrapping succeeds, the system will send the data scrapper to be saved in the database. When doing classification process, the system is gaining new data from the data scrapper on the database which are not clarified. Furthermore, the system processes the data using naïve bayes classification. After the results are gained, the Data Retrieve is returned back to the database.

4. Implementation

Figure 5 is a use-case diagram explaining the functions of system built based on the proposed architecture of system in Figure 4 that involves three actors (User, Web Sources, and Admin System) who's each case is closely related to each other.

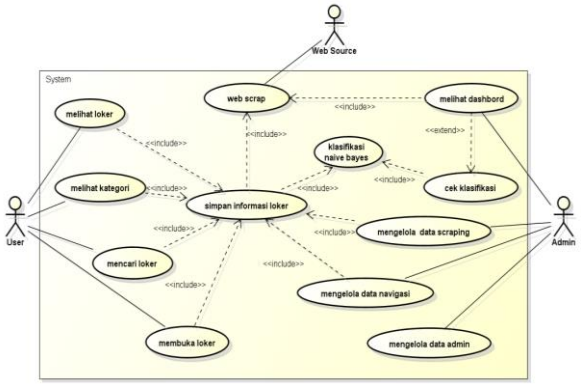

Figure 5. *Use case* diagram

User in this context is job seekers who can see and look for information about job vacancy, filter the information based on certain categories, and access the details of the information. *Admin* is a system manager who is authorized to access the information clarification results as well as manage data scrapping, navigation data, and admin data. Web Sources, in the meantime, includes other systems/ data sources and information accessed by the search engine.

5. Testing

The system setting is carried out using black-box testing method after the search engine is developed. Below are the results of the testing using black-box method or the so-called system functionality testing.

5.1 Testing Scrapping and Classification Results

The results of testing of web scrapping technique and naïve bayes classification algorithm can be seen in Table 1.

Table 1 Testing Scrapping and Classification Results \

Testing	Scrapping Results			Classification Results		
	Time	Number	Total	Number	Total	Accuracy
1	2017-01-14 23:49:42	4.343	4.343	3.105	3.105	71,5%
2	2017-01-29 23:30:05	44	4.387	43	3.148	71,75%
3	2017-01-30 04:46:16	75	4.462	74	3.222	72,2%
4	2017-01-30 15:29:48	249	4.711	186	3.408	71,5%
5	2017-01-30 16:15:40	55	4.766	41	3.449	72,4%

In general, Table 1 reveals that the implementation of web scrapping technique has been optimum in automatic update of data and information required. On the other hand, naïve bayes algorithm used in five sessions of the testing has been doing classification with consistent accuracy (>70%). The average of the accuracy itself reaches 71.87%.

5.2 User Interface Testing

The results of interface testing can be seen in Table 2 and Table 3.

Table 2. Client-Side Interface Testing (Portal)

No	Scenario	Input	Results		Status
			Expected Output	Actual Output	
1	Seeing job vacancies	Accessing websites	Showing information of newest vacancies	Showing information of newest vacancies	Successful
2	Seeing categories	Choosing categories	Showing vacancies based on categories	Showing vacancies based on categories	Successful
3	Seeing categories if vacant	Choosing categories	Showing information of the vacancies based on the “not found” category	Showing information of the vacancies based on the “not found” category	Successful
4	Looking for jobs	Inserting keywords	Showing information of job vacancies	Showing information of job vacancies	Successful
5	Looking for jobs if vacant	Inserting keywords	Showing information job vacancies “not found”	Showing information job vacancies “not found”	Successful
6	Opening job vacancies	Typing titles of wesbsites on job vacancies	Showing websites about job vacancies	Showing websites about job vacancies	Successful

Table 3. Server-Side Interface Testing (Dashboard)

No	Scenario	Input	Results		Status
			Expected Output	Actual Output	
1	Seeing dashboard	Clicking dashboard menu	Showing dashboard page	Showing dashboard page	Successful
2	Web scrapping	Clicking dashboard menu	Doing scrapping process	Doing scrapping process	Successful
3	Classification check	Pushing classification button	Sending classification instruction	Sending classification instruction	Successful
4	Naïve Bayes classification	Sending classification instruction	Giving classification results	Giving classification results	Successful
5	Saving job vacancies information	Saving the results of scrapping/ classification	Giving validation of ”sucessful upate of scrapping/ classification”	Giving validation of ”sucessful upate of scrapping/ classification”	Successful
6	Managing data scrapping	Choosing scraping menu	Showing scrapping page	Showing scrapping page	Successful
		Searching data to be deleted	Searching the data	Searching the data	Successful
		Choosing the data to be deleted and pusing ’delete’ button	Deleting the data and validating the deleted data	Deleting the data and validating the deleted data	Successful
		Pushing ’delete’ button and not yet choosing the data to be deleted.	Giving validation of not choosing the data to be deleted	Giving validation of not choosing the data to be deleted	Successful
7	Managing navigation data	Choosing maintain menu	Showing maintain page	Showing maintain page	Successful
		Choosing add new, edit, or delete buttons	Executing the action	Executing the action	Successful
		Confirming add new, edit, or delete	Validating the upgrade of the navigation	Validating the upgrade of the navigation	Successful
8	Managing admin data	Choosing admin menu	Showing admin page	Showing admin page	Successful
		Chooing add new, edit, or delete buttons	Executing the action	Executing the action	Successful

Table 2 and table 3 prove that after having tested using black-box method, all the functions/ features of the system of the search engine works as expected.

6. Conclusion

1) Web scrapping technique and naïve bayes classification algorithm implemented in a search engine of job vacancies in Indonesia work well; and

2) Naïve bayes classification algorithm shows optimum performance of classification of job vacancies information. The results of the testing of five-time classification on eight categories show that the algorithm performs consistent accuracy above 70% (the average is 71.87%).

References

[1] R Horne, S Katiwada O Paray 2016 *Indonesia labor market Outlook* Jakarta: International Labor Organzation
[2] E Allen and K Kim 2015 *Indonesia Labour market information systems and services* Jakarta: International Labour Organization
[3] A Pranav and S Chauhan 2015 Efficient Focused Web Crawling Approach for Search Engine," *Int. J. Kalol Inst. Technol. Res. Canter* **4** 5
[4] M. Thurland 2010 *Architect's Guide to Web Scraping With PHP* Canada
[5] S Borrate 2013 *A Practical Guide to Web Scraping* Pune: Lean Publication
[6] V Krunal 2014 Content Evocation using Web Scraping and Semantic Illustration *IOSR J. Comput. Eng.*
[7] G Sandi, S H Supangkat and C Slamet 2016 *Health Risk Prediction for Treatment of Hypertension*
[8] D S Maylawati and G A P Saptawati 2016 Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang in *International Conference on Computing and Applied Informatics* 1–6
[9] A Ashari, I Paryudi and M Tjoa 2013 Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool *Int. J. Adv. Comput. Sci. Appl.* **4** 11 33–39
[10] L Dey, C Sanjay, A Biswas, B Beepa and S Tiwari 2016 *Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier*
[11] L P Rajeswari, K Juliet and Aradhana 2017 Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier *Int. J. Comput. Trends Technol.* **2** 4