Web Scraping and Naïve Bayes Classification for Job Search Engine

To cite this article: C Slamet et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 288 012038

View the article online for updates and enhancements.
Web Scraping and Naïve Bayes Classification for Job Search Engine

C Slamet¹*, R Andrian¹, D S Maylawati¹, Suhendar¹, W Darmalaksana² and M A Ramdhani¹
¹Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati, Bandung, Indonesia
²Fakultas 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](image)

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.

\[
P(V_i) = \frac{P(A|B)P(B)}{P(A)}
\]

(1)

Naïve Bayes Classifier is an algorithm with simplification model suitable with classification of data and documents. The equation is.

\[
V_{map} = \underset{v_j \in V}{\text{arg max}} (V_j|a_1, a_2, \ldots, a_n)
\]

(2)

Based on the equation, the bayes formula is.

\[
V_{map} = \underset{v_j \in V}{\text{arg max}} \frac{P(V_j|a_1, a_2, \ldots, a_n)}{P(a_1, a_2, \ldots, a_n)}
\]

(3)

Since the value of \(P(a_1, a_2, \ldots, a_n)\) for each \(V_j\) is similar, the value can be omitted so that the equation becomes.

\[
V_{map} = \underset{v_j \in V}{\text{arg max}} P(V_j|a_1, a_2, \ldots, a_n)P(V_j)
\]

(4)

Assuming that each word in \(<a_1, a_2, a_n>\) is independent, so \(P(V_j|a_1, a_2, a_n)\) in equation is written as follows.

\[
(V_j|a_1, a_2, \ldots, a_n)P(V_j) = \prod P(a_i|V_j)
\]

(5)

Thus, the equation is.

\[
V_{map} = \underset{v_j \in V}{\text{arg max}} \prod P(a_i|V_j)
\]

(6)

Meanwhile, the value of \(P(v_j)\) is determined in the training, whose value is approached by.

\[
P(V_j) = \frac{\text{Doc}_{V_j}}{|\text{Doc}|} \prod P(W_k|V_j)\frac{n_k+1}{n+|\text{Vocabulary}|}
\]

(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 scraping 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 \( P(V_j) \) and \( P(W_k | V_j) \).

\[
\begin{align*}
P(V_j) &= \frac{|Doc_j|}{|D|} \\
P(W_k | V_j) &= \frac{N_k + 1}{n + |Vocabulary|}
\end{align*}
\]

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.

\[
P(W_k | V_j) = \frac{N_k + 1}{n + |Vocabulary|}
\]

In which \( N_k \) is the frequency of \( W_k \) words in documents categorized into \( V_j \), while \( n \) is the number of words in \( V_j \) 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 \( V_{map} = \max_{V_j \in V} P(V_j) \prod P(ai | V_j) \)

In which \( \prod P(ai | V_j) \) is a series of multiplied values of \( P(V_j) \) and \( P(W_k | V_j) \) in each \( P(V_j) \) category, and \( \max P(V_j) \) is comparing the results of \( P(V_j) \) in each category, dan taking the
maximum values as \( V_{map} \) category. Generally, the proposed working sequence of information classification collected is explain in Figure 3.

![Proposed Working Sequence of Classification](image)

**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.

![Proposed Architecture of System](image)

**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.
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>
<thead>
<tr>
<th>Testing</th>
<th>Scapping Results</th>
<th>Classification Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>Number</td>
</tr>
<tr>
<td>1</td>
<td>2017-01-14 23:49:42</td>
<td>4.343</td>
</tr>
<tr>
<td>2</td>
<td>2017-01-29 23:30:05</td>
<td>4.387</td>
</tr>
<tr>
<td>3</td>
<td>2017-01-30 04:46:16</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>2017-01-30 15:29:48</td>
<td>249</td>
</tr>
<tr>
<td>5</td>
<td>2017-01-30 16:15:40</td>
<td>55</td>
</tr>
</tbody>
</table>

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)
Table 3. Server-Side Interface Testing (Dashboard)

<table>
<thead>
<tr>
<th>No</th>
<th>Scenario</th>
<th>Input</th>
<th>Expected Output</th>
<th>Actual Output</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Seeing dashboard</td>
<td>Clicking dashboard menu</td>
<td>Showing dashboard page</td>
<td>Showing dashboard page</td>
<td>Successful</td>
</tr>
<tr>
<td>2</td>
<td>Web scrapping</td>
<td>Clicking dashboard menu</td>
<td>Doing scrapping process</td>
<td>Doing scrapping process</td>
<td>Successful</td>
</tr>
<tr>
<td>3</td>
<td>Classification</td>
<td>Pushing classification button</td>
<td>Sending classification instruction</td>
<td>Sending classification instruction</td>
<td>Successful</td>
</tr>
<tr>
<td>4</td>
<td>Naïve Bayes</td>
<td>Sending classification instruction</td>
<td>Giving classification results</td>
<td>Giving classification results</td>
<td>Successful</td>
</tr>
<tr>
<td>5</td>
<td>Saving job</td>
<td>Saving the results of scrapping/ classification</td>
<td>Giving validation of &quot;successful update of scrapping/ classification&quot;</td>
<td>Giving validation of &quot;successful update of scrapping/ classification&quot;</td>
<td>Successful</td>
</tr>
<tr>
<td>6</td>
<td>Vacancies</td>
<td>Choosing scraping menu</td>
<td>Showing scraping page</td>
<td>Showing scraping page</td>
<td>Successful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Searching data to be deleted</td>
<td>Searching the data</td>
<td>Searching the data</td>
<td>Successful</td>
</tr>
<tr>
<td>7</td>
<td>Managing data</td>
<td>Choosing the data to be deleted and pushing 'delete' button</td>
<td>Deleting the data and validating the deleted data</td>
<td>Deleting the data and validating the deleted data</td>
<td>Successful</td>
</tr>
<tr>
<td></td>
<td>Scrapping</td>
<td>Pushing 'delete' button and not yet choosing the data to be deleted.</td>
<td>Giving validation of not choosing the data to be deleted</td>
<td>Giving validation of not choosing the data to be deleted</td>
<td>Successful</td>
</tr>
<tr>
<td>8</td>
<td>Managing admin</td>
<td>Choosing maintain menu</td>
<td>Showing maintain page</td>
<td>Showing maintain page</td>
<td>Successful</td>
</tr>
<tr>
<td></td>
<td>Data</td>
<td>Choosing add new, edit, or delete buttons</td>
<td>Executing the action</td>
<td>Executing the action</td>
<td>Successful</td>
</tr>
</tbody>
</table>

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
[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