Exploiting Web Scraping for Education News Analysis Using Depth-First Search Algorithm

Endah Ratna Arumi¹, Pristi Sukmasetya²
¹²Department of Informatics Engineering, Universitas Muhammadiyah Magelang, Indonesia

Article Info

Article history:
Received April 14, 2020
Revised June 24, 2020
Accepted June 27, 2020
Published July 15, 2020

Keywords:
Algorithm
Depth-first search
Education news
Online news
Web scraping

ABSTRACT

Online news is one source of data that is always up to date and provides information or factual data. The search engine is one of the features for users to be able to enter keywords based on the expected category quickly. The development of education in Indonesia makes it essential to discuss, in this study using unstructured data in online news with the keyword Education included as a parameter, and adding search methods in the field of Artificial Intelligence so that the data becomes more accurate. Data that used here was from online news, namely CNN Indonesia, Detikcom, and Liputan6. Using Python Programming with depth-first search method (DFS), when compared with the results data for relevant news. Web erosion using DFS will be very helpful in searching because this method can check the date data was sent and then track the destination URL. Of the three online media sites, Detikcom produces the highest monthly data yielding an average of 885 news about education. At the same time, Liputan6 has the least amount of data on average, 28 news per month, but the data obtained are very relevant compared to Detikcom and CNN Indonesia.

1. INTRODUCTION

The development of information technology, especially the web site, making social media and online news will continue to use in the era of big data, online news is one of the sources of the data used by the people of Indonesia in gathering information/news for the current. The wide range can be accessed using digital devices / mobile devices[1]–[4]. Online news provides information in the form of data or facts that can be obtained easily by providing a search engine feature by entering a keyword in accordance with the expected category [5]–[7]. Nowadays, information and news that spread to much online news, it is likely many lies will be accepted by the public [8], users should be more selective, and watch any news to know the truth of the news. News category of online news is very diverse, ranging from economic, political, cultural, sports, natural disasters, and the news is no less critical on education.

News of the use of information technology, especially the Internet in the field of education, by making use of online systems in developing students learning processes [9]. Educational development should be monitored to create more advanced countries and able to compete in the world. One effort to improve human resources in Indonesia in terms of education, the quality of teaching and learning process conducted by educators[10]. Therefore, the necessary information regarding the current state of education, educational
constraints, and related infrastructure in support of education. News of Education in several online news, but related news of education has not grouped in a data bank in the form of a web, so users should use the search feature to filter the news.

Analysis of text mining is a quick way to read and summarize textual information to collect sample data on the website or text documents, finding information that can be accessed via a web browser[11]–[13]. Extracting information from online news, web sites or web mining is a solution that can be used to collect information related to education issues [2]. Collecting data using surveys, interviews and questionnaires, often constrained by funding and validity of the data [14]. Web mining or scraping of the Web as an application of data mining techniques to the website to categorize the relevant information, the classification step, similar cases, and to detect web spam from several sources available on the Web, in the form of documents and web services[15]–[18]. The process in the same web mining with data mining, which only differentiates ways of collecting data. In data mining, processing data retrieved from the data that has been received and stored in the database, whereas, in web mining, data collection using a web search can automatically download a web page [19]. Automated data collection, extract information, and traversing a hyperlink on a web page called Web crawling or Robot [2] [7] [14] [20] [21].

Research about the web scraping has been done, one of the study is to analyze the market in Indonesia using IP manipulation so that the user agent used is not hindered by the destination site [22]. Scraping Web can be used to collect data on natural disasters, in a study conducted source data retrieved from online news, then enter a keyword as a parameter, and extracted into Excel format (.csv) [5]. Web scraping declared valid if it does not steal data, and manipulate the data, even with the scraping of the Web provides reciprocal symbiosis with increased traffic from the scraping of the source link [23]. Web crawling application may also be combined with in-depth search methods to find specific news because the results are more accurate and valid [6]. Crawler robots can be used in libraries, whose job is to find book content that users often look for before the library makes a book purchase [24]. Web crawlers can be integrated with REST-based web services, to support the use of the cross-platform system, and then serves as a library for Python that provides an API to access the data contained in Google Play Store [25]. Currently, web scraping/web mining can combine with several methods in Artificial Intelligence to speed up the search data is valid and relevant to the keywords.

Data collection in this study uses web scraping, and web crawlers with Python programming language, combined with one of the methods applied in Artificial Intelligence [16], to identify and select a reliable source of data from web sites and online news so that the data obtained is accurate and in accordance with the purpose of collecting information related to education [26]. Web scraping need to pay attention to the quality of the data, the accuracy of the data, especially when dealing with the World Wide Web, where a lot of unstructured data, so a lot of data cleansing and data necessary fail-safe program scraper made. The source of data was taken from the online news, i.e. www.cnindonesia.com, www.detik.com, www.liputan6.com. Based on data that are collected from web crawlers, there is a link from the search data and unstructured online news. In this research, a depth-first search method is used to be a searching method. Then, this study used unstructured data from online news with the keyword Education included as a parameter and added search methods in the field of Artificial Intelligence so that the data becomes more accurate. The addition of this method is expected to increase the accuracy and obtain the relevant data, and results will be stored in a NoSQL database.

2. METHOD

This study using a combination of Python Programming web scraping and depth-first search method. We can see the detail in Figure 1. In this study, the source of data collection was taken from the online news, i.e. www.cnindonesia.com, www.detik.com, www.liputan6.com. Web scraping process used to input a search using the keywords "Education", with data taken in the form of dates, headlines, links to photos, news links, categories, so that specific categories can group them.

Figure 1 is an explanation of the first and in-depth search method. The method used in the search process starting from the root node of the date to be traced, down to the news URL node, afterwards into categories, and finally to the final node or news content page that can be determined. This process will continue until the last node is executed; it will do a reverse search or reverse to search for the next node [6]. When the process of backtracking is not found, it will be repeated to other new branches to get to the endpoint search by categories that have been determined [27].

Exploiting Web Scraping for Education News Analysis Using Depth-First Search Algorithm

(Endah Ratna Arumi1, Priсти Sukmasetya2)
Figure 1. Web Scraping using Depth first search [6]

Figure 2. Flowchart of Web Scraping for Education News using Depth-First Search
Figure 2 shows the flowchart that describes the data searching technique based on keywords, looking fit of the data using a depth-first search, store data, and analyze results. This study will begin with the process of loading the URL in online news intended for the keyword "education". After that, the proceed with depth-first search starts by taking the start date, and the expiration date of the news, the URL news, and a category, and will be repeated until it is found news that matches the search. Search result URL will continue to scratch and crawl data in accordance with keywords. Web scraping and web crawling have different tasks even though they are used on web pages. In the web scraping, users can take the raw data in a web browser, perform text mining, building a data set is then stored and analyzed to obtain data in real-time.

Web crawlers can move between websites automatically for web scraping, and was assigned to make a copy of the relevant, seek information, and collect data on the web site. The process of extracting information then performed, and after scraping and crawling process data, news data will be exported to an Excel file format (.csv), and stored in a NoSQL database.

3. RESULTS AND DISCUSSION

This study implemented in Python Programming web scraping. An adding a search method by Depth First Search. The search method will be used as an alternative to obtaining relevant results in accordance with the three objectives of online news. Web scraping process is done by a number of attempts to produce valid data.

Python programming is used as a means of annihilation with a written script method Depth First Search in the search process data. The initial step user enters a keyword and date for the news; these measures will continue until the data on that date is found, then perform the sorting/extraction of data to choose the news that is relevant to the desired search. Results of this web scraping is the document/data compiled and stored using Excel format (.CSV).

After the data extraction process is complete, the process of comparing the data acquisition with three online news followed, namely Indonesia CNN, Detikcom, and Liputan6. Retrieving data with web scraping in this study had a span of one year, beginning in January 2019 - December 2019. The data were collected in accordance with relevant keywords, i.e. about education news. The results of the three-web scraping of online news will be stored and analyzed all the results obtained, to ensure that the news by keyword searchable.

Figure 3. Result of Web Scraping in CNN Indonesia news

Figure 3 results of the scaling Web in Indonesia CNN web site, data is retrieved varied, relevant education news, but there is some news unrelated to education news. CNN Indonesia consists of five categories, namely, lifestyle, tv, national, international, technology, and economics. On average CNN Indonesia will send 239 news every month, with the keyword education news, all the news can be included in all categories.
Figure 4. Result of Web scraping in Detikcom news

Figure 4 results from web scraping Detikcom. Detikcom online news has a lot of users, because the categories suggested by these sites are very diverse, ranging from the category of news, finance, the most popular searches, technology, sports, automotive, until the health category. It makes the user very comfortable using the site with ease of data of all kinds of things. The average Detikcom will send more than 500 news every month for a specific category. As more data is obtained from this site, we need to filter to the information being used is valid. Because not all data transmitted includes accurate data.

Figure 5. Result of Web Scraping in Liputan 6 news

Figure 5 is the result of scraping Liputan6 website. Liputan6 has a slightly different category with two online news that have been discussed, namely events, liputan6 citizens, economy, territory, bibs news, hot, bank, and celebrities. Based on two other online news, web sites Coverage 6 show fewer data scrape. Within one year from the average monthly post data as much as 28 education news records, this is very much different from the results memos between Indonesia CNN and Detikcom.

Based on Figure 3, Figure 4, and Figure 5, the number of results scraping of three online news gets mixed results. Online news scraping Detikcom has the most substantial amount of data to be searched by keyword education news, but data obtained by the analysis of the needs of scraping back right post. Analysis...
of the data scraping results obtained, the next step is the comparison of data scraping the relevant and valid data less relevant even unrelated to the keywords that are searched, is evidenced in Table 1. In Table 1 are shown all data scraping of three online news, from January 2019 until December 2019.

Table 1. Comparison web scraping with relevant data

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of Scraping Data Results</th>
<th>Number of Relevant Data Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CNN Indonesia Detik com Liputan 6</td>
<td>CNN Indonesia Detik com Liputan 6</td>
</tr>
<tr>
<td>January</td>
<td>196 702 21</td>
<td>21 696 21</td>
</tr>
<tr>
<td>February</td>
<td>202 792 26</td>
<td>32 779 26</td>
</tr>
<tr>
<td>March</td>
<td>258 1107 34</td>
<td>51 1067 34</td>
</tr>
<tr>
<td>April</td>
<td>148 700 11</td>
<td>15 679 11</td>
</tr>
<tr>
<td>May</td>
<td>136 633 39</td>
<td>45 621 39</td>
</tr>
<tr>
<td>June</td>
<td>182 660 17</td>
<td>69 643 17</td>
</tr>
<tr>
<td>July</td>
<td>242 938 24</td>
<td>71 914 24</td>
</tr>
<tr>
<td>August</td>
<td>279 895 23</td>
<td>58 881 23</td>
</tr>
<tr>
<td>September</td>
<td>250 853 29</td>
<td>64 831 29</td>
</tr>
<tr>
<td>October</td>
<td>353 1213 40</td>
<td>121 1180 40</td>
</tr>
<tr>
<td>November</td>
<td>321 1192 43</td>
<td>99 1154 43</td>
</tr>
<tr>
<td>December</td>
<td>309 937 32</td>
<td>127 925 32</td>
</tr>
<tr>
<td>Total</td>
<td>2.876 10.622 339</td>
<td>773 10.370 339</td>
</tr>
</tbody>
</table>

Table 1 presents a recap of the results of a web scraping produced by CNN Indonesia, Detikcom, and Liputan6. Based on tests conducted, the results can be seen most news posting them on the website Detikcom, in the first year, there were 10,622 observation posts news about education news, with the amount of data relevant to as many as 10,370. Figure 6 displays the comparison between the data scraping with data relevant to the site Detikcom data discrepancies are not much different, which means that the results of the keyword search feature pertinent to the news category.

![Graph of data Scraping result CNN Indonesia](image1)

![Graph of data Scraping result Detikcom](image2)

![Graph of data Scraping result Liputan6](image3)

Figure 6. Graphs of data scraping results with three online news
Figure 6 graphs the results of the data erasure with three online news, the results of the second-largest amount of data on CNN Indonesia, the scraping of the entire record 2,876 news reporting, with relevant data as much as 773. From the result, it can be seen that every month the news about the news post more than 300 education news. Based on Figure 6, there is more than half the data be erased using the keyword search feature on the CNN website Indonesia is not optimal, many irrelevant posts with results taken. Results of data extraction Liputan6 within one year, there was news about education news that a record 339, extraction of all data relevant to the keywords searched. Although the post every month a little bit, but Liputan6 search feature on the site, displaying data corresponding to a keyword input from the user. Based on the three online news that has been studied, the results revealed a significant difference between the amount of data scraping and also the amount of relevant data.

4. CONCLUSION

Based on the implementation and testing of web scraping done with Python Programming with Depth First Search method, it can be concluded that the feature by entering a specific keyword search will affect the result of scraping. Depth First Search will significantly affect the process of searching in web scraping because this method will perform keyword and selecting post similarity data then find the destination URL. From scraping web for three online news, it can be concluded that the results obtained are very diverse, Detikcom produces result data erasure most, Liputan6 sites has a number of deletion of data is the smallest. Still, the data collected as a whole is very relevant than Detikcom and CNN Indonesia. But, CNN Indonesia has a lot of irrelevant data.

5. REFERENCES

Exploiting Web Scraping for Education News Analysis Using Depth-First Search Algorithm
(Endah Ratna Arumı, Pristi Sukmasetya)