INTEGRATED (Information Technology and Vocational Education)
Volume 1 No. 1 | April 2019 : 19-22
ISSN Cetak: 2656-9760
ISSN Online: 2657-0769

Design process data storage and organize data scraping

Falentino Sembiring[1], Dian Permata Sari[2]
[1] Department Sistem Information, Universitas Nusa Putra
[2] Departement Education System and Information Technology, Indonesian Education University.
[1] falentino.sembiring@nusaputra.ac.id, [2] dianpermatasari@upi.edu

Abstract- In this study, Web scraping will describe a process of retrieving urls from similar sites for the erosion process and storing url data in a daily, weekly, monthly and yearly database, so that url data can be valid and invalid urls will be monitored. This filtering will be done to facilitate a number of processes to be moved into the database. The next process is the url based on the available data content will be differentiated based on titles, tags, keywords such as SEO. Each step will be stored in the data warehouse to create a data center url. Hopefully this is the step to collect data for big data. Limited Issues by designing a web crawler by searching for similar sites and storing processes in a database. From the database it will be directed to the data warehouse. Once in the data warehouse, data will be processed in the interface to the user as well as divided by classification. and data that already exists in the data warehouse is used for data mining / data mining by using a number of algorithms to index data and present it in a user interface.

Keywords: Data Warehouse, Similar Site, Storing Data, Web Scraping.

Abstrak -Dalam penelitian ini, Web scraping akan menjelaskan suatu proses pengambilan url dari situs serupa untuk proses erosi dan menyimpan data url pada database harian, mingguan, bulanan, dan tahunan, sehingga data url dapat valid dan url yang tidak valid akan disaring. penyaringan ini akan dilakukan untuk memudahkan sejumlah proses untuk dipindahkan ke dalam database. Proses selanjutnya adalah url berdasarkan data konten yang tersedia akan dibedakan berdasarkan judul, tag, kata kunci seperti SEO. Setiap langkah akan disimpan di gudang data untuk membuat pusat data url. Semoga ini adalah tahap mengumpulkan data untuk big data. Masalah dibatasi dengan mendesain crawlers web dengan mencari situs serupa dan menyimpan proses dalam database. Dari basis data itu akan diarahkan ke data warehouse. Setelah di gudang data, data akan diproses di antarmuka ke pengguna serta dibagi dengan klasifikasi. Data yang sudah ada di data warehouse digunakan untuk data mining/penambangan data dengan menggunakan sejumlah algoritma untuk mengindeks data dan menyajikannya dalam antarmuka pengguna.

***Keywords*: Data Warehouse, Penyaringan Data, Situs Serupa, Web Scraping.**

Makalah dikirim: 30 Maret 2019; Revisi: 7 April 2019; Diterima: 15 April 2019; Publish: 30 April 2019

I. INTRODUCTION

Web scraping is the process of extracting content in the form of data or information from a website. Web scraping is used because the data needed is not available in the RSS or API [1]. In addition to extracting content, data and information, this technique is also used to automate the data retrieval process or called a robot.

Web scraping is often known as screen scraping. Web scraping cannot be included in the field of data mining, because data mining implies an attempt to understand semantic patterns or trends from the large amount of data that has been obtained. Application of web scraping (also called intelligent, automated, or autonomous agents) only focuses on how to obtain data through retrieval and extraction of data with varying data sizes [2].

The results of data scraping will be processed and stored in the database. the data on the database will filter to check content, data and information to be stored in a collection of databases. Some databases will be managed on their storage so that a process is needed that can increase the speed of extracting data from web scraping. So that web pages or websites can be taken with a customized search engine to search for field keywords. This might involve or not use keywords that are limited to a list of controlled vocabulary. Hope the author of this method is commonly used by indexing search engines.

II. BACKGROUND AND RELATED RESEARCH

Machine learning is the data analysis process that automates analytical model building and can be treated as an optimization problem. Manual data manipulation and processing to produce meaningful results is impossible with big data due to the massive quantity. To discover hidden information within a huge dataset without human intervention, computer algorithms are applied iteratively to learn from such data (L'Heureux et al., 2017).

Big data is an “imprecise description of a rich and complicated set of characteristics, practices, techniques, ethical issues, and outcomes all associated with data” (Japec et al., 2015, p. 839).

Web scraping is essentially a form of data mining. Items like weather reports, auction details, market pricing, or any other list of collected data can be sought in Web scraping efforts. The practice of Web scraping has drawn a lot of controversy because the terms of use for some websites do not allow certain kinds of data mining. Despite the legal challenges, Web scraping promises to become a popular way of collecting information as these kinds of aggregated data resources become more capable. The level of the web scraping can be seen in the following Figure 1.

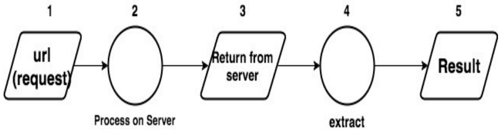


Figure 1. The level of web scraping [1]

A database is a data structure that stores organized information. Most databases contain multiple tables, which may each include several different fields. For example, a company database may include tables for products, employees, and financial records. Each of these tables would have different fields that are relevant to the information stored in the table.

a. Previous Researcher

1. Aprianto
   Aprianto conducted research by implementing Web Scraping and Text Mining for Acquisition and Categorization of Website Information About hydroponics [3]. Prianto gets results with a portal / web page that is able to use topic models in real-time categorizing emerging information all the time on the internet and exploring topic-modeling algorithms based on semantic / ontology similarities resulting in more accurate categorization of topics.
2. Ram Sharan Chaulagain [4] on Cloud Based Web Scraping for Big Data Applications research. Ram compares Web Scraping to Cloud platforms with web scraping saved on local devices.
3. Ahmat Josi [5] makes Application of web scraping techniques on search engine scientific articles. A Search Engine Application and applying web scraping techniques has successfully published scientific journal articles from a number of academic portals both from Indonesia and abroad.

III. DESIGN PROCESS GETTING URL

In designing this web scraping process requires initial data in the form of url data. This data is the beginning of the web scraping process. The following steps will be designed on Figure 2

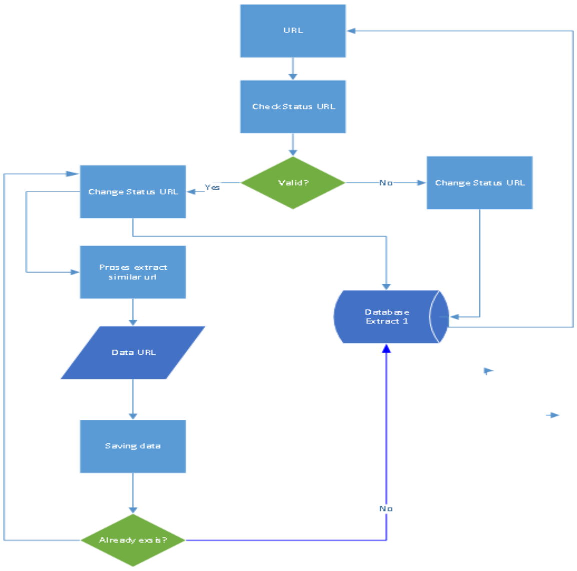


Figure 2. How Scraping Web Works

Makalah dikirim: 30 Maret 2019; Revisi: 7 April 2019; Diterima: 15 April 2019; Publish: 30 April 2019

Existing data in the form of url will be extracted similar url on the website to be contained. The data obtained by web scraping will be checked whether the url address is valid to be accessed, if the url is valid and invalid, the system will change the status of the url into the database and continue the extraction process. The results of the data will be checked for the existence of the database, if it does not exist, it will be saved, but if the data already exists, the next process will continue. Repetition process can be seen in the following Figure 3 repetition process to storage

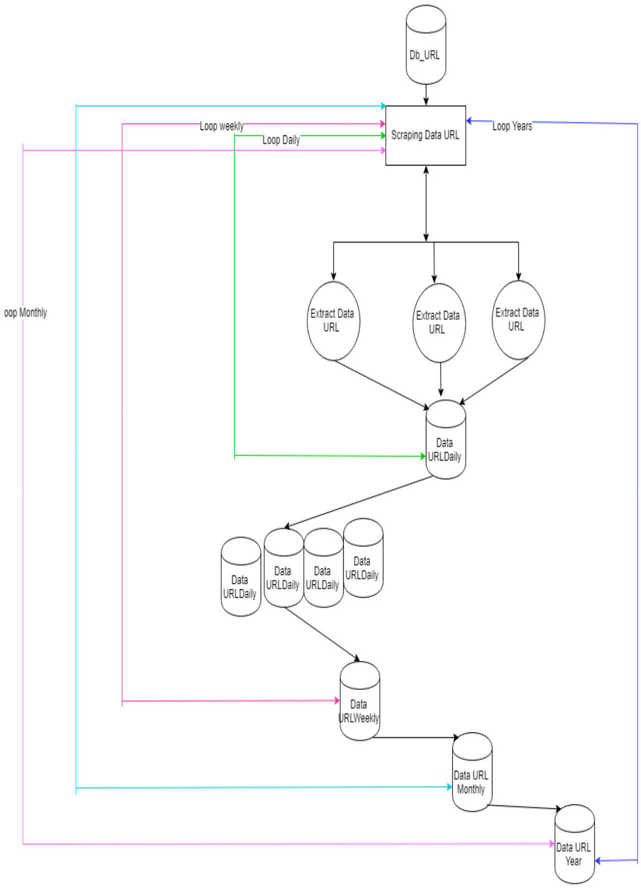


Figure 3. Process of storage and repetition of extraction

Domain data in the form of URL obtained from manual input. The url domain data will be extracted using the scraping URL system with python in 3 stages and the results of each step will be stored in a temporary database. Daily extraction data will be re-filtered in the system to ensure the URL Domain is valid and domain data will be looped at the daily, weekly, monthly and yearly stages. Illustration of data usage for search engines can be seen in the following Figure 4.

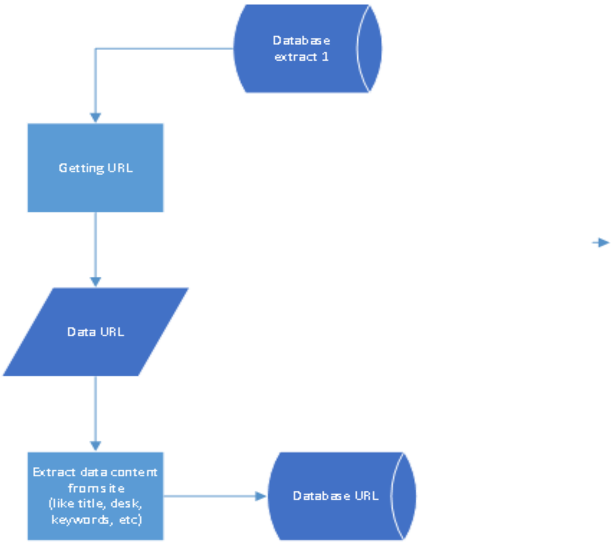

Database extract 1
Getting URL
Data URL
Extract data content from site (like title, desk, keywords, etc)
Database URL

Figure 4. URL data storage

Data that has been stored in the database and filtered is expected to become the database for the data warehouse and will filter based on each category. This data will be presented to the user as a result of data mining and will be displayed in the user interface with the help of a number of algorithms for indexing data and data rank like SEO. For more details, see the following information in Figure 5

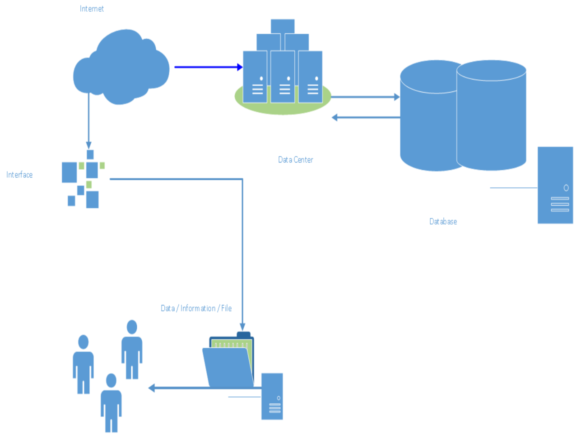


Figure 5. Illustration of data usage for search engines

The hope is that in the research data collected in databases such as data warehouses will be presented to the user by creating an algorithm scheme to display information from the data warehouse.

IV. RESULTS

The results of this study are in the form of a web erosion process design whose data will be stored in the database and will be repeated in the daily, weekly, monthly, and annual ranges. The data generated will be filtered based on the content and keywords like SEO and

Makalah dikirim: 30 Maret 2019; Revisi: 7 April 2019; Diterima: 15 April 2019; Publish: 30 April 2019

the results of filtering will be made classification. The next stage will be stored in the database. The data in the database will be stored again at the next stage into the data warehouse which will be combined with several techniques that are often used with pentaho. data already in the data warehouse is used for data mining / data mining by using a number of algorithms to index data and serve in a user interface

1. Tools

   The tools used in this research are Python programming with version 2.7. Library used :

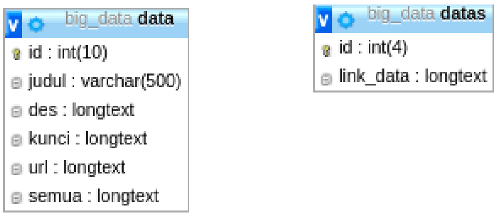
```
    import requests
    import urllib.request
    import time
    from bs4 import Beautiful Soup
import re
inputString = """add1 http://mit.edu.com abc
add2 https://facebook.jp.com.2. abc
add3 www.google.com
add4 https://www.google.com. 123

another test with ipv4 http://192.168.1.1/test.jpg. toto2
website with different port number
www.test.com:8080/test.jpg not port 80
www.website.gov.us/login.html
test with ipv4 (192.168.1.1/test.jpg).
search at google.co.jp/maps.
test with ipv6
2001:0db8:0000:85a3:0000:0000:ac1f:8001/test.jpg."
""
regex=ur"\b((?:https?://)?(?:(?:www\.)?(?:[\da-z\.-]+
)\.(?:[a-z]{2,6})|(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9]
[0-9]?)\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?
)|(?:(?:[0-9a-fA-F]{1,4}:){7,7}[0-9a-fA-F]{1,4}|(?:[0-
9a-fA-F]{1,4}:){1,7}:|(?:[0-9a-fA-F]{1,4}:){1,6}:[0-9
a-fA-F]{1,4}|(?:[0-9a-fA-F]{1,4}:){1,5}(?::[0-9a-fA-F
]{1,4}){1,2}|(?:[0-9a-fA-F]{1,4}:){1,4}(?::[0-9a-fA-F
]{1,4}){1,3}|(?:[0-9a-fA-F]{1,4}:){1,3}(?::[0-9a-fA-F
]{1,4}){1,4}|(?:[0-9a-fA-F]{1,4}:){1,2}(?::[0-9a-fA-F
]{1,4}){1,5}|[0-9a-fA-F]{1,4}:(?:(?::[0-9a-fA-F]{1,4}
){1,6})|:(?:(?::[0-9a-fA-F]{1,4}){1,7}|:)|fe80:(?::[0-9
a-fA-F]{0,4}){0,4}%[0-9a-zA-Z]{1,}|::(?:ffff(?::0{1,4}
){0,1}:){0,1}(?:(?:25[0-5]|(?:2[0-4]|1{0,1}[0-9]){0,1}
[0-9])\.){3,3}(?:25[0-5]|(?:2[0-4]|1{0,1}[0-9]){0,1}[0
-9])|(?:[0-9a-fA-F]{1,4}:){1,4}:(?:(?:25[0-5]|(?:2[0-4
]|1{0,1}[0-9]){0,1}[0-9])\.){3,3}(?:25[0-5]|(?:2[0-4]|
1{0,1}[0-9]){0,1}[0-9]))))(?::[0-9]{1,4}|[1-5][0-9]{4}|
6[0-4][0-9]{3}|65[0-4][0-9]{2}|655[0-2][0-9]|6553[0
-5])?(?:/[\w\.-]*)*/?)\b"

matches = re.findall(regex, inputString)
print(matches)
```

2. Database

   Database system that has been designed using MySQL with version 5.7. Designing a database to store temporary data like the following Figure 6:


big_data data
id : int(10)
judul : varchar(500)
des : longtext
kunci : longtext
url : longtext
semua : longtext

big_data datas
id : int(4)
link_data : longtext

Figure 6. Data storage for temporary

The extraction stage of the data will be stored in the datasets which will be triggered into the python system for extracting the second stage of data which will be stored in the database data and will be filtered again into the final phases to be stored in the data warehouse server

3. Server

   The server used in this study is Apache, which is installed in an open source system, the Linux operating system (Ubuntu).

V. REFERENCES

[1] Adila, Nelawati. "Implementation of web scraping for journal data collection on the SINTA website." *Sinkron: jurnal dan penelitian teknik informatika* 7.4 (2022): 2478-2485.

[2] Boeing, Geoff, and Paul Waddell. "New insights into rental housing markets across the united states: web scraping and analyzing craigslist rental listings." *Journal of Planning Education and Research* 37.4 (2017): 457-476.

[3] Priyanto, Agung, and Muhammad Rifqi Ma'arif. "Implementasi web scrapping dan text mining untuk akuisisi dan kategorisasi informasi dari internet (Studi Kasus: Tutorial Hidroponik)." *Indonesian Journal of Information Systems* 1.1 (2018): 25-33.

[4] Chaulagain, Ram Sharan, et al. "Cloud based web scraping for big data applications." *IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2017.

[5] Josi, Ahmad, and Leon Andretti Abdillah. "Penerapan teknik web scraping pada mesin pencari artikel ilmiah." (2014).