Design Process Data Storage and Organize Data Scraping

Falentino Sembiring\textsuperscript{1}, Dian Permata Sari\textsuperscript{2}

\textsuperscript{1} Department Sistem Information, Universitas Nusa Putra
\textsuperscript{2} Departement Education System and Information Technology, Indonesian Education University.

\textsuperscript{1} Falentino.sembiring@nusaputra.ac.id, \textsuperscript{2} Dianpermatasari@upi.edu.

Abstract - In this study Web scraping will explain the process of retrieving urls from similar sites for the erosion process and storing url data on daily, weekly, monthly, and annual databases, so that url data can be valid and invalid urls will be filtered. filtering will be done to make it easier for a number of processes to be moved into the database. The next process will distinguish url based on available content data based on title, tags, keywords like SEO. Each step will be stored in the data warehouse to create the url data center. Hopefully this is the stage to collect data for big data. Problems are limited by designing web crawlers by searching for similar sites and storing processes in the database. From the database it will be directed to the data warehouse. after in the data warehouse, data will be processed in the interface to the user divided by classification.

Keywords : Web Scraping, Storing data, Similar site, Data warehouse.


Keywords : Web Scraping, Penyaringan data, Situs Serupa, Data warehouse.
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. 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.

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.
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

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.
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 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`
- `from bs4 import Beautiful Soup`

For more details, see the following information in figure 5.

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.

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. 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


