DATA CRAWLERS TO COLLECT DATA

[1] Jerripothula Tejasri, [1]K. Sai Abhinav, [1]G. Samyuktha, [1]B. Dileep , [2]Dr. CH.B.N. Lakshmi

[1]Student, Department of CSE, TKR College of Engineering & Technology, Hyderabad, India
[2]Professor,Department of CSE, TKR College of Engineering & Technology, Hyderabad, India

Abstract : A data crawler, often known as a spider or web crawler, is an Internet bot that methodically browses the World Wide Web in order to build search engine indices. Web crawling is used all the time by companies like Google and Facebook to acquire data. By following connections in web pages, web crawling mostly refers to downloading and storing the contents of a vast number of websites. Information on the web is frequently altered or modified without warning or notification. A web crawler scans the internet for fresh or updated content. Users can use various hypertext links to locate their resources. Three steps make up the web crawling process. The spider begins by crawling specific pages on a website. Following that, it continues to index the website's words and content, and finally, it visits all of the site's hyperlinks.Web search engines have additional issues as a result of the large number of web pages available, making the obtained results less relevant to the analysers. Web crawling, on the other hand, has recently focused primarily on acquiring the links to the appropriate documents.Today, multiple techniques and software are used to crawl links from the web that must be further processed for future usage, causing the analyser to become overloaded. This project focuses on crawling the links and extracting the associated information in order to make processing easier for various purposes.

Index Terms **– Auto Scraper, Scraping data for different websites, Storing data in .csv or .xlsx format.**

I. INTRODUCTION

In today's fast-paced world, technology has vastly improved and has made an impact in every aspect of a human life. Only a few lines of code make up a Web crawler. The main idea is to collect a large amount of data from websites such as Amazon, Wikipedia, and whenever there is illegal use of data from different websites, it makes our process easier to find criminals and can help resolve issues, so that the crime rate, which is increasing in today's digital world, can be reduced.

A web crawler is a program that collects data from the internet and stores it in a database for further analysis and organisation. Online crawling is the act of gathering web pages and structuring them in such a way that the search engine can obtain them quickly. The main goal is to do so efficiently and swiftly without causing too much disruption to the distant server's operation. A web crawler starts with a seed, which is a URL or a set of URLs. The crawler goes to the first URL in the list. It searches the web page for links to other web pages and adds them to the existing list of URLs.

The goal of the project is to raise awareness among the statistical community regarding the benefits and drawbacks of using online scraping technology for official statistics.

The project goes over the technological, data security, and legal requirements for web crawlers in great depth.

Web crawling began with a map of the internet, showing how each website was linked to the others. It was also used by search engines to find and index new pages on the internet. Web crawlers were also used to test the vulnerability of a website by testing it and assessing whether or not any issues were discovered. Crawlers collect data that can later be used and processed to classify documents and provide insights into the data acquired. Anyone with a basic understanding of programming may construct a crawler. Making an efficient crawler, on the other hand, is more complicated and time consuming.
The crawler built can also be utilised in Machine Learning for problems like price prediction and classification.

CHARACTERISTICS OF CRAWLERS

Crawlers have unique characteristics. Crawlers that crawl the internet must have the following essential features in order to fulfill their job, as well as the servers that store data and the web as a whole.

1. Robustness: It is defined as the ability to withstand a variety of conditions. Spider traps are loops in the web that are intended to deceive the crawler into recursively crawling a single domain and being stranded there. They create a never-ending web page loop. Such traps necessitate the crawler's resilience. These traps aren't always set up to deceive the crawler; they could instead be the result of poor website design.
2. Politeness: When a crawler can visit a web server, there are rules that must be followed. These etiquette norms must be adhered to. A server's purpose is to serve requests that aren't related to the one for which it was created. If the server is hampered, the crawler may be blocked entirely by the server. As a result, it is preferable to adhere to the server's policies.
3. Distributed: The crawler must be able to work in a distributed environment. It may have several images of itself crawling the internet at the same time in perfect coordination.
4. Scalable: Scalability should be a priority for the crawler. It should be able to scale up and down as needed, with the ability to add new machines and bandwidth.
5. Efficiency and Performance: System resources such as computing power, network bandwidth, and storage should be used sparingly. The crawler's efficiency is determined by these criteria.
6. Quality: The crawler should be able to tell the difference between useful information and useless information. Servers are primarily responsible for serving other requests that contain a large amount of data that may or may not be beneficial. This content should be filtered out by crawlers.

II LITERATURE SURVEY

Web crawlers have been utilised by several researchers to collect web data for their studies. These research articles are useful in analysing current work and identifying gaps that need to be filled in the current work. In the subject of online mining, web crawling can be used to automatically discover and extract information from the internet.

In Automatic Price Collection On The Internet research paper, the use of web crawling technology can help to increase statistical data quality while also reducing data collection workload. Using automated price gathering methods allows statisticians to respond more quickly to the growing number of data sources available on the internet. Any method implementation necessitates meticulous planning in a variety of areas. Legal and data security concerns must be addressed right away.

In An Effective Implementation Of Web Crawling Technology To Retrieve Data From –The World Wide Web (www), Web Crawlers are an important component of web crawlers. The elite web slithering method is an important part of various web administrations. The establishment of such frameworks is far from trivial: These crawlers' data management spanned a large area. It's critical to keep a good balance between irregular access memory and plate access. A web crawler allows search engines and other users to keep their databases up to date on a regular basis. Web crawlers are an important aspect of search engines, and the algorithms and design of these crawlers are guarded as trade secrets.

III ARCHITECTURE

A crawler must have both a solid crawling approach and a highly optimised architecture, as discussed in the previous sections. Online crawling or spidering software is used by web search engines and some other websites to update their web content or indices of other websites' web content. Web crawlers save pages to be processed by a search engine, which indexes the pages so that users may find information more quickly.

Building a high-performance system that can download hundreds of millions of pages over many weeks involves a variety of issues in system design, I/O and network efficiency, as well as robustness and manageability.

Web crawlers are an important aspect of search engines, and the algorithms and design of these crawlers are guarded as trade secrets. When crawler designs are published, there is frequently a significant lack of information that makes it difficult for others to duplicate the work. There are also growing concerns about "search engine spamming," which prevents major search engines from disclosing their algorithmic ranking methods.

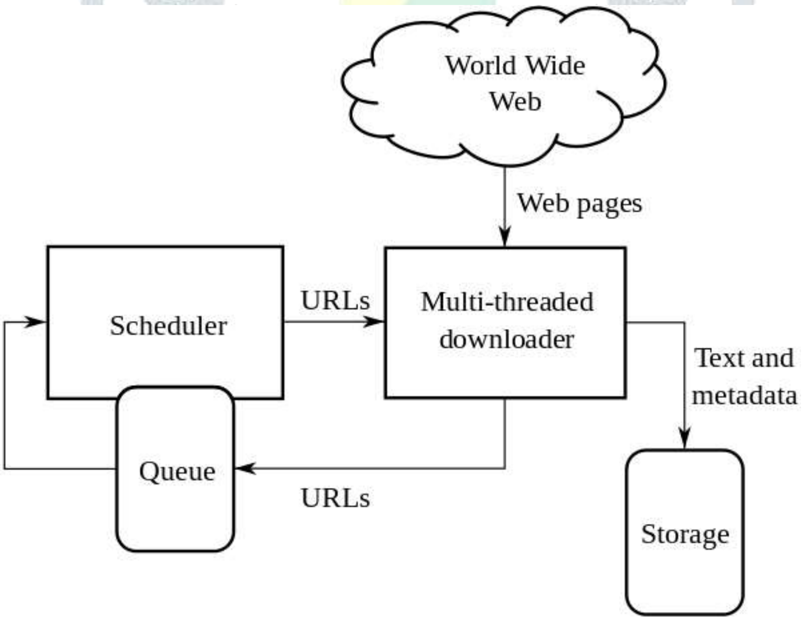

Fig. 3.1 Crawler Architecture

3.1 METHODOLOGY

Typically, bots accomplish simple and structurally repetitive jobs at a rate far higher than is achievable for people alone. The most common usage of bots is in web spiders, in which an automated script fetches, scans and files web server information.

Lists of widely used servers and popular pages are the common beginning points. The spider starts with a popular site and indexes the words on its pages and follows each link on the website. Throughout the most used parts of the Web, the spider mechanism begins to travel rapidly.

The relative importance of each website: Most web-based crawlers do not crawl through the entire internet that is publicly available and are not designed to do so, but instead decide which pages to crawl for, based on the number of other pages linked to that page, the number of page visitors and other factors that indicate the likelihood of important information being contained on this page.

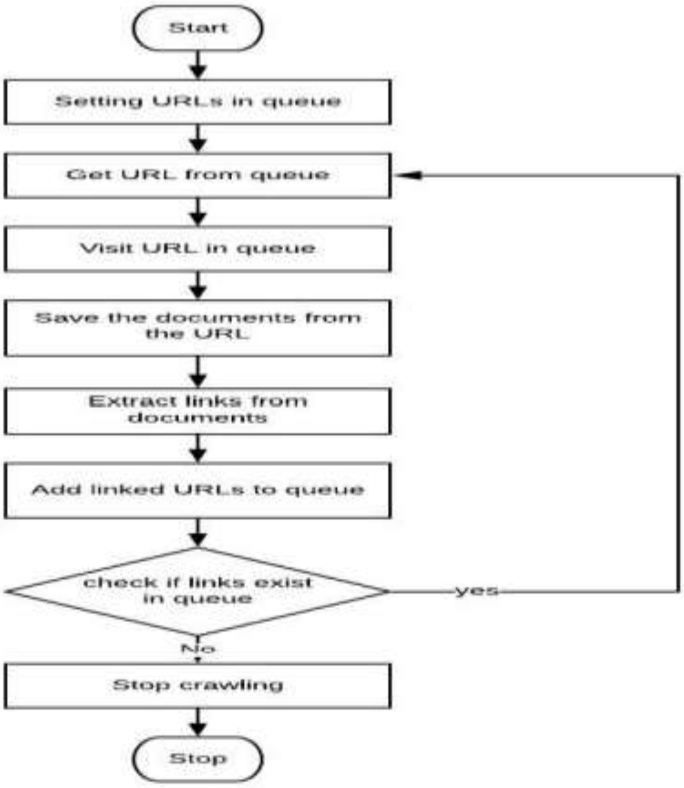

Fig 3.2 Flow diagram of web crawlers

IV ALGORITHM

WEB CRAWLER WORKING PROCEDURE

Data crawler and scraping is implemented using Autoscraper. It is a fast, light -weight and reliable tool for data collection. First import all the required libraries. After importing all the libraries, a class instance of AutoScraper is initialised. Using this class instance a URL and a wanted list is passed as parameters, wanted list contains all the necessary data that is required to be scraped from the target website. A model is built and trained on test case. Then it can be used for querying and fetching relevant information on various things.

```
Import Libraries;
Process_Function (max_pages){ → [Working Function] Initializing Value;
While (page <= max_pages) → [Variable ‗page‘ is less than ‗max_page‘]
Set Url = (Web_Link); → [Webpage URL from where information is to be extracted]
Create Object; → [Create object for get all the information, returned values of the HTTP request]
Object = Url Request;
Convert Object to Plain_Text;
Create Object; → [Create object to pull out the information using‗Autoscraper]
For() { → [ Used traverse through the information (HTML file) that is pulled out]
Declare Variable;
Variable = Concat_Text; Display Variable;
}
Increment value;
```

Crawler_web (Argument) → [Function is called with argument_1‘]

DOWNLOAD AND STORE DATA USING HYPER LINK PROCEDURE

The data that has scraped in the first step can now be downloaded into various file formats, depending upon the user requirement. Few available formats are: .xlsx, .csv. The data can be stored in the form of the above mentioned formats and can be used for Machine Learning problems like Regression and Classification.

```
1:Import Libraries;
2: Declare Variable;
3: Variable = Link Extracted; → [link extracted by the web crawler in saved in the variable]
4: Function_download_csv (csv_url); → [Function_download_csv()‘ with parameter_csv_url‘]
5: Create Object;
6: Object = Extract Information; [Object is created to extract the information from the link]
7: Declare Variable;
8: Variable = Read and Stored Data;
9: Declare Variable2;
10: Variable2 = String_format (Information); → [information is converted to string format and stored in variable]
11: Declare Variable3;
12: Variable3 = String_Spilit; → [String split in variables and saved in_Variables3]
13: Create Object2;
14: Object2 = Create a CSV file; → [Object is created to create a csv file]
15: Open CSV File;19
16: For (){ }→ [Using a_for‘ loop each line variable is written on the csv file]
17: CSV_File = Closed and Stored;
```

V RESULTS AND DISCUSSION

The idea is to have high-quality, authoritative information on a web page that is quoted on many other web pages and receives a lot of visits, and thus it is particularly important for a search engine that is indexed – just as a library can be sure that many copies of a book are retrieved by many people. Web pages review: Web content is continuously updated, deleted or transferred to new sites. Web crawlers are required to review pages on a regular basis to ensure the newest content version is indexed. Requirements for robots.txt: Web crawlers determine which sites to crawl based on robots.txt protocol (also known as the robots exclusion protocol). They check the file robots.txt hosted by the web server of that page before they crawl through a webpage. A robots.txt file is a text file which sets the rules for bots which access the website or application hosted. The rules specify which pages the bots can go and which links can be followed.

The AutoScraper object includes the destination url and a wanted_list and sets certain criteria for the parameters given. Note that the review put in the wanted list is the first review on the whole website It will only highlight tags that perfectly match the items of interest. Once you print results, all the elements identified in the webscraper will be shown. Web crawlers which are build using autoscraper tool is now capable of collecting the required information that user is going to ask, For example user want to scrape the information of price and model names of electronic gadgets A crawler model is built and trained based on the earlier given arguments like URL and wanted_list. Using the trained model, one can now query about all the information relevant to them from the website. Data can be scraped successfully using the autoscraper model, this scraped data can now be stored in supported file formats like: .csv, .xlsx. These formats are proiminently used in the fields like Data Analytics and Machine Learning. Using the above generated datasets. An analyst can now work on price fluctuations and predictions and classification type of problems.

At first, Request – Response cycle checks the availability of the website’s server. This lets us know whether the site is up and running or has downtime. A HTTP response code stating 200[OK] means that the site is working properly and crawling can be performed on it. A test model is built, so that the crawler can understand the structure of the website we are crawling on. This model in trained based on few parameters that are relevant to user. Once the model is trained it generates data which has been scraped from the website.
The generated data is unorganized. In order to separate the data according based on their type, certain rules are generated, which makes it easy to find and use for further queries. These generated rules are grouped together. Aliases are created for readability and understandability purposes. The entire rules and test data are stored in the form a schema, using which , user can perform querying on the website regarding all the required information. This schema is stored and loaded when needed for new query in order to crawl and scrape data from a website. There is no need of training the model for this purpose. Loading the previously stored schema is enough.

Crawler also generates new rules based on user query and groups them together according to their types. It can be easily understood. Based on the aliases provided in the initial phase, data is scraped and served.

This scraped data can now be stored locally, in supported file formats like .csv, .xlsx. These files can now be used in prediction and classification type of problems, as they are prominently used in Data Analytics and Machine Learning fields.

In general, testing is the process of determining how well something performs. In the case of humans, testing determines the level of knowledge or skill attained. Testing is performed at critical checkpoints in the entire process to verify whether objectives are being fulfilled in computer hardware and software development. The process of determining the correctness and quality of the software product and service under test is known as software testing. It appears to have been created to determine whether the product meets the client's specific requirements, wants, and preferences. At the end of the day, testing is the process of executing a framework or application with the objective of identifying bugs, faults, or flaws. Testing's job is to point out flaws and provide Dev (Developers) a hint to assist them repair them correctly according to the criteria.

- Find as many faults (or defects) in a given product as feasible.
- Demonstrate that a certain software product meets the requirements.
- Validate the quality of software with the least amount of money and effort.
- Create high-quality test cases, run effective tests, and send out accurate and useful problem reports.

When user gives a irrelevant URL or invalid URL to the crawler, crawler recognizes it and prompts user with “Please enter a valid URL to test its request response cycle” message. As the provided URL is irrelevant its request – response cycle cannot be determined.During the initial phase of building a trained scraper model, URL of the target website provided has to be working properly in order to train scraper model. When an invalid URL is given by the user, “Please enter a valid URL” message is displayed.
After training the model, user can now be able to query about all the relevant information that is available in the target website. If a query given to the crawler in not in its trained schema, or not available in target website. A message is displayed saying, “Unable to scrape the data. Please Try again with a valid query”.When an empty pandas dataframe is initialized, there is no data to store, i.e., there is zero rows and columns in the dataframe. This dataframe is of no use and this issue may have occurred due to user’s invalid query or target website on downtime. In such cases,crawler prevents the storage of data and user cannot get any prompt requesting them to store the data. A message is also displayed saying “No data is available to scrape and store”.In the final phase, user can now be able to locally store all the data scraped by the crawler from the target website. User is prompted to enter a file format supported by the program. Upon entering any invalid file format, the program restricts user from saving the scraped data locally by displaying “Please enter a valid file format” message.

VI CONCLUSION

Based on the aforementioned data, it can be inferred that as the number of crawled sites grows, the system improves. Following an initial spike, it is discovered that when the overall number of pages grows, the durable pages that must be crawled occur at more frequency. This demonstrates that when applied to a real-time program that processes millions of data, the performance figures will inevitably achieve their maximum efficiency, resulting in the most efficient Web Crawling System. The information integration with simultaneous metatag extraction is an attractive feature of this system. The major goal of the proposed system is to create a database of pages and links from the Internet. It also focuses on recrawling regularly changing web sites in order to maintain the database's contents up to date. The inclusion of a metatag extractor in the crawling process also opens up a world of options dependent on the analyzer's needs. This eliminates the requirement for separate extraction modules to be included in the projects. Future work might focus on decreasing the amount of bandwidth necessary to create this system and making it compatible with higher-level connections.

REFERENCES

[1] An-Effective-Implementation-Of-Web-Crawling-Technology-To-Retrieve Data-From-The-World-Wide-Web-www
http://www.ijstr.org/final-print/jan2020/An-Effective-Implementation-Of-Web Crawling-Technology-To-Retrieve-Data-From-The-World-Wide-Web-www.pdf
[2] An Efficient Approach for Web Indexing of Big Data through Hyperlinks in Web Crawling
https://www.hindawi.com/journals/tswj/2015/739286/
[3] Web Crawler: Extracting the Web Data
https://www.researchgate.net/publication/287397481_Web_Crawler_Extracting_the_Web_Data
[4] Automatic data collection on the Internet (web scraping)
Author: Ingolf Boettcher
https://ec.europa.eu/eurostat/cros/system/files/Boettcher_Automatic%20price%20collection%20on%20the%20Internet.pdf
[5] A Smart Web Crawler for a Concept Based Semantic Search Engine
Author: Vinay Kancherla
San Jose State University
https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://duckduckgo.com/&httpsredir=1&article=1380&context=etd_projects55

[6] Web Crawling-based Search Engine using Python
https://www.researchgate.net/publication/335031903_Web_Crawling based_Search_Engine_using_Python
[7] Autoscraper
https://pypi.org/project/autoscraper/
[8] Pandas
https://pandas.pydata.org/docs/getting_started/index.html