International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I4P110

Volume 68 Issue 4, 54-58, April 2020


*Original Article*

# Survey on Leveraging Django and Redis using Web Scrapping

Abhinav R[1], Abhinav Raman[2], Abilash R[3]

Received Date: 01 March 2020
Revised Date: 17 April 2020
Accepted Date: 20 April 2020

***Abstract*** *- Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. The web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol or through a web browser. Over the years, due to advancements in web development and its technology, various frameworks have come into use, and almost all websites are dynamic with their content being served from CMS. This makes it tough to extract data since there is no common template for extracting data. Hence we use RSS.RSS (originally RDF Site Summary; later, two competing approaches emerged, which used the backronyms Rich Site Summary and Really Simple Syndication, respectively) is a type of web feed that allows users and applications to access updates to websites in a standardized, computer-readable format. This project combines the use of RSS to extract data from websites and serve users in a robust and easy way. The differentiation is that this project uses server-side caching to serve users almost instantaneously without the need to perform data extraction from the requested site all over again. This is done using Redis and Django.*

***Keywords***- *web scrapping , leveraging django, Redis*

## I. INTRODUCTION

Web pages, besides core contents, consists of other elements such as banners, navigational elements, copyright information, external links etc. Extracting relevant information from these data of the semi-structured or unstructured form is useful for many applications, such as document identification, text categorization, clustering, topic tracking etc. There are two types of scaling in the current system for workloads, namely horizontal scaling and vertical scaling. The above two scales prove to be very expensive as the number of users increases. Since web scraping is a process-intensive task, the system may slow down when there are many user requests. To avoid these issues, we use server-side caching for scraping data by using Redis instead of the scaling methods. The priority for the users is given by queue, and queuing is implemented by celery, thus making the system cost-efficient and robust. This technique proves to be very valuable for startup companies.

## II. LITERATURE SURVEY

In paper [1], the creators Belen Vela, Jose, Paloma and Carlos has developed a technological framework for the processing, management and exploitation of open data to promote accessibility to urban public transport. This paper specifically focuses on the data extraction and processing of the existing information on the web concerning public transport and its accessibility for the generation of an open data repository in which to store this information. This method allows the extraction of public transport data and the existing accessibility information from a selected website. The advantage depicted through this paper is that it provides users with a convenient way of accessing public transport for people with special mobility needs. The issue that arises when carrying out the programming task manually is a very time-consuming task owing to the unstructured nature of the sources used to extract data. Therefore the generation of web scrapers for the public transport domain is to be automated in order to reduce the programming effort required.

In paper [2], the creators K Sundaramoorthy and R.Durga developed a platform that dedicates all the latest news from national and international resources and presents them in simplified words. The crawler fetches the content from the RSS feeds of the stored URL. The application developed by them extracts the individual news from the web pages and provides them in a single platform. This software may attempt to automatically recognize the data structure of a page or provide a recording interface that removes the necessity to manually write web-scraping code or some scripting functions that can be used to extract and transform content and database interfaces that can store the scraped data in local databases. The proposed system implies gathering all the latest news using an RSS aggregator and displaying them under a single roof. The advantage of this paper is it gets a new reading experience using its simple user interface and user experience design. It can gather clean news articles from three websites without having the need to manually copy and paste the article. The main issue of this system is it provides only the source URL of the news. In this system, it just displays the name of the digital newspapers available on the web. The user has to go to each and every link to read the news. Hence it is a time-consuming process.

In paper [3], the creators Deborah and Deny Triawan conducted a survey on information retrieval and capabilities on e-commerce websites. The system uses web crawling with scraping techniques that utilize simple HTML DOM as parser tag HTML from the source code of the intended web page. The web extracting program begins by compiling an HTTP request to obtain resources from the targeted websites. This request can be formulated in a URL that contains a GET request containing a POST query. Once the request has been successfully received and processed by the targeted website, the requested resource will be retrieved from the website and then sent back to the web scraping program. Visiting information resources one by one and comparing data from all the information sources visited will add much more time to the process of rediscovering the information. It takes a technique that can gather information from multiple sources into a single entity to facilitate the process of information retrieval. This proposed system uses three e-commerce websites as a source of information. It implements web scraping techniques on search engines by using PHP programming language, and the search results are accumulated using my SQL database. Since the precision rate is 93.9%, there are chances for the system to provide results that are not accurate. Hence the success of the information retrieval is relatively low.

In paper [4], Achmad, Windi Ekaand Muhamat Abdul developed an approach of web scraping based on a sequence of characters that define a search pattern. This approach consists of three steps, i.e. analyzing news website structure, constructing a pattern of regex and implementing the patterns as a set of rules in web scraping. They used two different types of patterns, i.e. content pattern(for extracting original text article of news) and filter pattern(for eliminating non-news elements). This paper is based on usual web scraping methods by extracting general text from HTML pages. However, a little bit different from those works, the proposed approach tries to get five different elements by providing title, publication date, author, clean text article and URL address of news article from HTML page of websites. From the result, we could see that each news website has a unique layout to present its news. Each news website provided one unique HTML element for the article link, title, author, and publication date. These news elements could be extracted easily by providing one corresponding Regex for each. However, this technique could not be implemented on news article content because it contains some non-news elements in every news website. Therefore we have to provide two kinds of Regex, i.e.Regex for filtering and Regex for extraction.

In paper [5], the creators Sandeep Sirsat and Vinay Chavan conducted a survey on pattern matching from news web pages. Although there are many existing methods that formulate the actual content identification problem as a DOM tree node selection problem, each one has some sort of lacunae. The proposed approach is based on the pattern matching technique. This technique uses a simple heuristic for the extraction of core contents from web pages which are mostly semi-structured in nature. This approach also uses devised algorithms that apply regular expressions (Regexes) to identify the correct pattern for extracting the actual text contents from these news documents. The advantage is that it deals with news web pages of any size and extracts core contents with high efficiency and high accuracy. This approach does not exploit the features of the DOM structure. It is template independent and is not dependent on any tag type. It utilises a restricted top-down mapping (RTDM) that is based on the post-order traversal of trees. However, the disadvantage is that it is based on the ambiguous assumption that the new site content could be divided into groups that share common format and layout characteristics. Thus it is not suitable to apply the RTDM approach for the news web pages having heterogeneous structure and page layout.

In paper [6], the creator's Li Zhao and Si-Feng Du designed a paper based on a content management system. A website content management system is a foundational website application platform for website design and information distribution and is an auxiliary tool system for website development. This paper proposes the key method in which the system information distribution module is constructed in the webpage element and template manner to fully reduce the complexity of the system design. This software is defined based on the open architecture and object-oriented method and is developed by using Java based on a pure object-oriented framework. The template can combine the contents with the pages in the design to manage website content and automatically generate the website. Correcting a page event can easily update the whole website. The template can include independent page elements of any webpage. Advantages in the system guarantee excellent cross-platform capability using pure Java. The system can run on Windows and Linux and can migrate to the Unix large scale machine. The disadvantage of the site is that it contains a large number of files and could leave the files prone to errors. Limited flexibility in design is also a major issue.

In paper[7], the creators Shreya Upadhyay and Vishal Pant constructed a web scraper for massive data extraction. There exist a number of automated approaches to data extraction from the web. Most of these approaches are ad-hoc and domain-specific. The paper makes an attempt to highlight the benefits of automatic data extraction tools and their role as a significant component in the development of knowledge-based systems. The need for a robust, automated, easy to use framework for extracting content from the web with minimal human effort across domains appears enticing. The architecture proposed by the authors for a web scraper addresses the gap to harvest data from the web Advantages: The framework offers a feasible and easy approach for parsing and extracting data on a large scale from multiple websites with minimal human interaction. Some of the unique aspects of the design are the simplicity of operations, adaptability to a

wide range of domains, applicability to a wide range of popular file formats and instantaneous provision of data. The issues faced are: Content loaded statically is easy to mine, but content generated dynamic scripting tools like javascript possess their own problem. Such a system is an invaluable tool for organizations operating at the enterprise level or for research institutions by providing unprecedented access to volumes of diverse data.

In paper [8], the creators David Mathew Thomas and Sandeep Mathur analyzed web scraping using python. The web scrapers conniving ethics and procedures are juxtaposed. It explains the working of how the scraper is premeditated. The technique of it is allocated into three fragments: the web scraper draws the desired links from the web, and then the data is extracted to get the data from the source links and finally stowing that data into a CSV file. The python language is used for carrying out the operation. The point of the paper is to remove the information from different sources with the assistance of programming known as the web scrawler scrape utilizing the programming language python adaption 3.6. The database is created, which collects all the unstructured data from various sources and then analyses them by the analytic process of its specifications, assembling, organizing, cleaning, re-analyzing, applying models, algorithms and finally providing the desired results.

| S.NO. | PAPER | TECHNOLOGIES /TECHNIQUES USED | ADVANTAGES | ISSUES | RESULT |
|---|---|---|---|---|---|
| 1.1 | A Semi-Automatic data scraping method for the public transport domain | Uses a semi-automatic generation of web scraper | It provides users with a convenient way for accessing public transport with special mobility needs. | Executing the programing task is very time consuming owing to the unstructured data of the sources used to extract the data. | The proposed paper uses an automatic data scraper to help users access public transport. |
| 1.2 | An aggregation method for news using web scraping method | Uses web crawling method, RSS aggregator. | It helps give the user a new reading system with its new user interface and user experience design. | It only provides the source URL of the system. The user has to go to each link to access the webpage. | The proposed system gathers all the latest news without revisiting those sites manually |
| 1.3 | Increased information retrieval capabilities on e-commerce websites using scraping technique. | It takes a technique that can gather information from multiple sources into a single entity to facilitate the process of information retrieval. | It implements web scraping techniques on search engines by using PHP programming language, and the search results are accumulated using my SQL database | The precision rate is 93.9%. There are chances for the system to provide results that are not accurate. Hence the success of the information retrieval is relatively low. | The proposed system retrieves information from e-commerce websites with a high precision rate. |
| 1.4 | An approach on web scraping on a new website based on regular expression | Web Scraping(extracting from HTML pages) ,using Regex(Regular expression) | The proposed approach tries to get five different elements by providing title, publication date, author, clean text article and URL address of news article from HTML page of websites | This technique could not be implemented on news article content because it contained some non-news elements in every news website | This paper gives an approach by extracting news from websites through regular expressions |

| | | | | | |
|---|---|---|---|---|---|
| 1.5 | Pattern matching for extraction of core contents from news web pages | Restricted top-down matching(RTDM), Regex | It is template independent and is not dependent on any tag type. | Based on the ambiguous assumption that the new site content could be divided into groups that share a common format and layout characteristics | The core contents of news websites are extracted using pattern matching that transforms the contents of news websites automatically to plain text forms. |
| 1.6 | Design and implementation of website content management system | Object-Oriented framework and java programming. | Pure Java can guarantee excellent cross-platform capability | it contains a large number of files and could leave the files prone to errors. Limited flexibility in design is also a major issue. | It enables the user to quickly develop, maintain and manage the high-performance dynamic website |
| 1.7 | Articulating the construction of a web scraper for massive data extraction | Web data extraction | The framework offers a feasible and easy approach for parsing and extracting data on a large scale from multiple websites with minimal human interaction | Content loaded statically is easy to mine, but content generated dynamic scripting tools like javascript possess their own problem | This paper proposes a system that articulates a web scraper with minimal human interaction |
| 1.8 | Data analysis by web scraping using python | Web scraping using Python language | Due to enormous community and library resources for python and the flexibility of coding, it is the most appropriate method for scraping data from the desired website | Because of the autonomous and heterogeneous nature of hidden web content, it has become an ineffective way to search this kind of data | The main outcomes of this paper were a user-friendly search interface, indexing, query processing and effective data extraction technique. |

## IV. CONCLUSION

The system uses a method of web scraping in order to extract data from websites as requested by the user. Most existing systems are inefficient since many users access a website at a time. This makes it difficult to retrieve information, thereby decreasing the rate of success. The proposed system uses server-side caching for scraping data using Redis, which is an inbuilt memory data structure that supports various kinds of data. The priority for the users is given by queue, and the queuing is done by Huey or Celery. This approach makes the system robust, easy to access and cost-efficient.

## REFERENCES

[1] Belen Vela, Jose, Ploma, A Semi-Automatic Data Scraping Method for the Public Transport Domain, IEEE Access, 7(10) (2019) 335-339.

[2] K. Sundaramoorthy, R. Durga, An Aggregation System For News Using Web Scraping Method, IEEE International Conference on Technical Advancements in Computers and Communications(ICTACC), (2017) 1340-1343.

[3] Deborah, Deny, Increased Information Retrieval Capabilities on An E-Commerce Website Using Scraping Techniques, IEEE International Conference on Sustainable Information Engineering And Technology(SIET), (2017) 829-834.

[4] Abdul, Windieka, Muhamat Abdul, An Approach on Web Scraping On News Website Based on Regular Expressions, IEEE 2nd East Indonesia Conference on Computer and Information Technology(Eiconcit), (2018) 906-923.

[5] Sandeep Sirsat, Vinay Chavan, Pattern Matching for Extraction of Core Contents From News Web Pages , IEEE Second International Conference on Web Research, (2016) 51-54 .

[6] LI Zhao, SI-Feng Du, Design And Implementation of Website Content Management System, IEEE International Conference on Information Management And Engineering, (2018) 5-8.

[7] Shreya Upadhyay, Vishal Pant, Articulating The Construction of A Web Scraper for Massive Data Extraction, IEEE Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), (2017) 1-3.

[8] David Mathew Thomas, Sandeep Mathur, Data Analysis by Web Scraping Using Python, IEEE 3rd International Conference on Electronics, Communication and Aerospace Technology(ICECA), (2019) 6179-6186.