Web Scraping News Portals for the Ease of News Reading

Dilip Suthar[1], Ajit Kumar Trivedi[2], Rishabh Kumar Pandey[3]

[1]*Student, Dept. of Computer Engineering, SLRTCE, India*
[2]*Student, Dept. of Computer Engineering, SLRTCE, India*
[3]*Student, Dept. of Computer Engineering, SLRTCE, India*

Abstract - *There are various news publishers online. They publish their content on multiple platforms. Imagine opening multiple websites every day to get some quality news. It makes information gathering much more difficult and time consuming. In today's fast changing world, keeping yourself updated with quality information is very important. Now, is there a way we can make it easier? A news aggregator can help in making this easier. In a news aggregator, a user chooses the websites he wants to read. Then the news aggregator collects the news articles from selected websites. And, the user is just a click away to get information from various websites. This task otherwise takes too much time from our daily routine. Our goal was to create an online news aggregator platform where the user can select the news portals they want to follow from the given news portal options and then the aggregator should collect articles from those news websites and display it on the user's news feed. Users will have the option of choosing the news category.*

Key Words: Web scraping, Web crawling, News Aggregation, Information Gathering, Beautifulsoup, requests, python

1.INTRODUCTION

A news aggregator is a web application that aggregates news articles from various news portals in one location for easy access. To make the news data relevant and meaningful, it must be updated frequently. The news should be categorized to make it easy to navigate and personalize the user experience. It should not contain any duplicate data, and news should be from primary sources to be useful and the system should have less hardware requirement.

The news aggregator provides the news updates from different portals in a systematized way to the users. News aggregation is a good way to remain updated with the latest news. It is time-saving and convenient to use. In times like these when information is more important than anything, they are the best way to compile data and store it into a form which is easy to process. Businesses that use news aggregation will be ahead of their competitors and ready to handle any future challenges.

1.1 Web Scraping

Web scraping is a process of obtaining data from websites. Unstructured data is converted into structured data, then that data is stored into a database or in any other form. This data can be used for various applications. There are many ways of scraping data from websites, some of them are using online services, APIs, creating your own program. Many websites provide API to access their data in structured form. There are also some websites which don't allow users to access their data. For situations like these Web Scraping is used to scrape data from websites.

Web scrapers can be developed using programming languages using third party libraries. These libraries provide the HTTP connection with features like SSL certificates and authentication After establishing the connection parsing is done using third party libraries.

2. LITERATURE SURVEY

[1] In this paper the author has talked about different aspects of web scraping and has also reviewed the tools available for web scraping. The author puts the case for an easy to use web scraper. The author has provided an

insight into things related to constructing web scrapers. The author also creates a web scraping framework for an e-learning Platform to harvest resources.

[2] In this journal the author gives an introduction to web scraping and also discusses the different softwares and tools for web scraping. The author had also explained the process of various types of web scraping techniques with their pros and cons with detailed information on where they can be applied.

3. PROPOSED SYSTEM

Our proposed system works in two phases. Getting the preference of the user. Serving the personalised news on the news feed. User first logs in into the system, and choses the news portals and category from the given options, then personalised news is served on the news feed of the user.

News aggregators scrape the data from different news portals in different categories and store them into a database. The objective of web scraping is to scrape the data from identified websites and convert it into a form which can be stored into traditional databases.

Figure 1 and 2 provide the architecture and overview of news aggregators. News scrapers and websites are integrated seamlessly to work together. The working of the system can be understood using the user case. If a user wants to read the personalized news, he needs to choose the news portals and news category from the given options. The news aggregator aggregates the news from the selected news portals and stores them into a database according to the category of the news. Then news is displayed on the news feed of the user.

A news aggregator works in three phases, It scrapes the web for the news articles. Then it stores the image, link, and title of the article in the database. The stored objects in the database are served to the news feed. The client gets information on his news feed.

Fig -1: Use Case Diagram

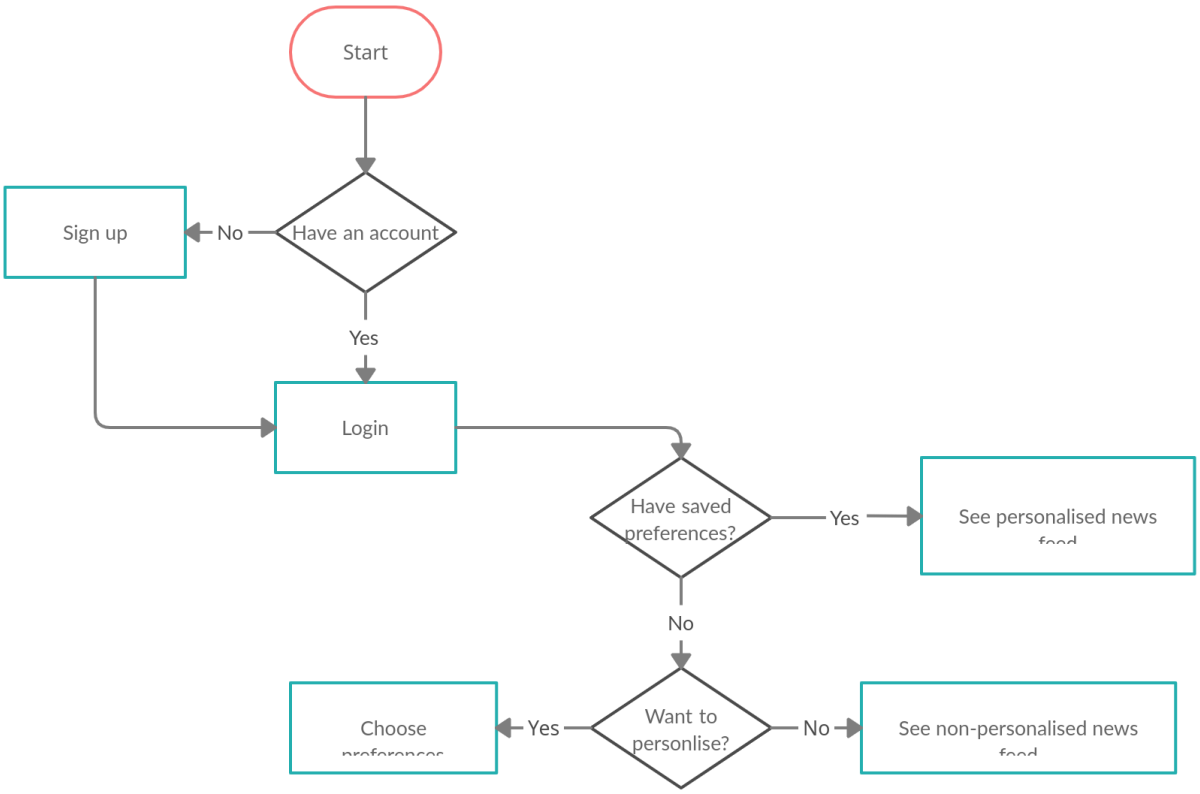

Chart -1: Flowchart of the news Agrregator

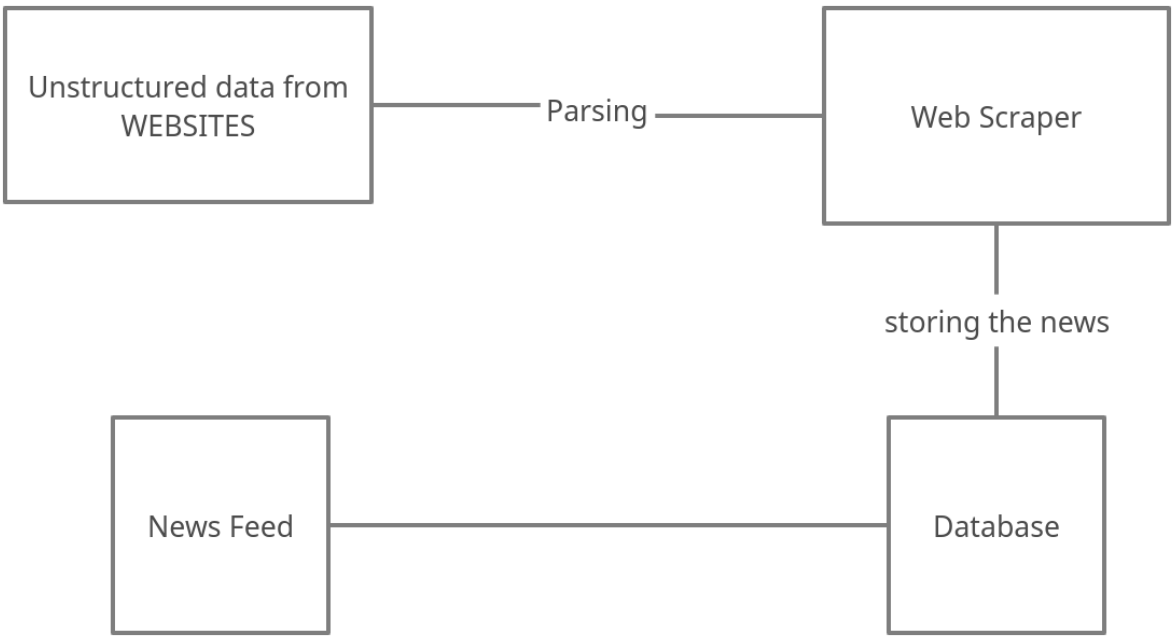

Fig -2: Working of News aggregator

4. METHODOLOGY AND IMPLEMENTATION

The following methodology was used in designing the news aggregator.

- Firstly, the URL of the webpage is required to be accessed to fetch the content. URLs were stored in a python dictionary.

- For getting the HTML structure of the website, HTTP connection has to be established to the web server.
- Web pages of news articles have a basic html structure, which can be accessed and then required data can be fetched using python libraries.
- Then the scraped article was stored in the database.

The following system was created using python, django and Postgresql.

- To get started, python with the following libraries needs to be installed - requests, BeautifulSoup, django, psycopg2 and Dateutil. We also need to install Postgresql.
- Next step is to create a web app to render the news. Web apps have been developed using django in the backend and HTML, CSS and JS in the frontend.
- Then the next step is to create the web scraper for the news portals. Since different websites have different structure, we cannot use the same scraper for all news portals, due to this different web scraper was created for the different websites.
- After establishing the HTTP connection using the requests module of python, BeautifulSoup is used to parse the html code of the webpage, then required data is parsed and stored into the database.
- Then webscraper and Web app were integrated to make the system.
- Then the articles are filtered according to the user's preference and rendered to the news feed.

The programming language used for the project was python. It is an open source general purpose programming language with thousands of open source libraries for different range of works, which makes it easy to do a variety of work from web development to web scraping. Using these libraries made it easy to implement the system and reduced the size of code which made the code readable to other individuals. And because of these features python was chosen to implement this project.

The web app was built on django web framework, which is an open source python web framework for rapid development of projects. A web framework is a tool consisting of components required for web app development. Django's inbuilt modules help in making the development process hassle free so one can focus on writing code without worrying about other side stuff.

5. CONCLUSION

People waste so much of their time getting quality news. News aggregators can solve that problem by providing news of their interest at one place. Although we knew before our research that news aggregators are popular and many see them as essential knowledge resources, we lacked insights on the role they play in software development and how they could be improved.

6. REFERENCES

[1] Upadhyay, Shreya; Pant, Vishal; Bhasin, Shivansh; Pattanshetti, -- [IEEE 2017 Second International Conference on Electrical, Computer and Communication Technologies
[2] Singrodia, Vidhi; Mitra, Anirban; Paul, Subrata -- [IEEE 2019 International Conference on Computer Communication and Informatics (ICCCI) - Coimbatore, Tamil Nadu, In