International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Socially Smart an Aggregation System for Social Media using Web Scraping

Namburu Srikanth[1], Vennapusa Tejaswini[2], Chethala Harish[3], D. Praveen Kumar[4]

[1,2,3]Student, Computer Science and Engineering, Hindustan Institute of Technology and Science, Tamilnadu, India
[4]Assistant Professor, Department of Computer Science and Engineering, Hindustan Institute of Technology and Science, Tamilnadu, India

---

Abstract - *Abstract - Socially Smart is a dedicated platform that aggregates all the latest social media posts from multiple social media web-based platforms such as famous reddit blog, Onion News and GitHub and summarizes them to display in a short and crisp words. This online web-based platform provides a service type interaction to users across the web. The main motto of this application is to access the required and useful social media content at same place which leads to save the time being spent on social media. It will fetch the top and trending posts without less priority posts. This Socially Smart platform uses Web Crawling or Web Scraping and API to fetch the top posts form various social media sites. The web crawling method fetches the posts form the mentioned URL, similarly the API Based method also gets the top posts from the websites API URL which is given already. The users will get excellent experience on this platform as it fetches data form the most used social media websites and the data is also the most commented and trending posts only and presents them with simple and pleasing User Interface.*

Keywords – Social Media aggregator, Web Scraping / Web Scraping and API.

1. INTRODUCTION

With the creation of new social media platforms, there is a rapid growth of Social Media users. There web applications meant for different purposes. Following these websites is very time consuming and mostly unimportant information is consumed in these sites. In this busy world it is very difficult for an individual to find time for every social website may it be regional news, coding blogs or reddit. This online platform helps to reduce time spent on these websites by collecting the data from those through web crawling, API and presenting the top and most trending data online.

This web-based application acts as a common point to social media sites. Socially Smart fetches the individual posts form the social media websites and provides them in a single flexible platform. It reduces the time being spent on social media and the efficiency of the useful info is also increases since the posts fetched the most commented and trending posts on respective websites. The struggle to visit every website reduces. The users can get full up-to-minute daily trending posts.

Data Science which is basically a data-driven subject about how enormous data should be handled and finding easier ways to extract knowledge or patterns from the data in various forms, either structured or unstructured. Socially Smart uses a technique called as Web scraping/Web crawling which is part of Data Science. Web Crawling/ Web Scraping is a procedure to acquire the information from the web-based applications which don't have API functionality. This method focuses on converting the formless data (HTML / XML format) on the web to structured data that can be stored in database such as SQL-Lite in meaningful format.

2. RELATED WORK

The Socially Smart system architecture is as shown in Figure 1.

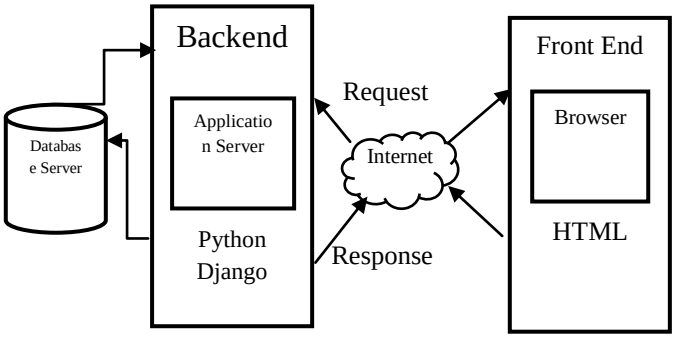

Fig 1. Socially Smart system Architecture

2.1 Existing System:

In the existing system, not all the information is available under a single roof. Therefore, this becomes a problem of time consumption. Not all the websites will have the same information. So, the social media user will need to reach multiple websites in order to see the posts. Therefore, to solve these issues, the proposed system which we are making provides information available for the user under a single roof. There are some existing platforms such as Newsone which provides news in a consolidated manner which acts as a news aggregator.

2.1.1 Existing System: Disadvantages

- Provides the source URL for only the news. In this system it just displays the name of only digital newspapers available in the web.

- Displays all the news that scraped without any narrowing of news to top searched or most followed post.
- Not Categorized, This System provides the latest news in a short manner but is not Categorize from which website it scrapes data from.

2.2 Proposed System:

We have so many social media applications opening every app and using them is time consuming. User cannot visit each and every social media website separately to use it for hours so I propose Socially Smart: An aggregation system for social media using Web scraping for fetching the data form a couple of websites and displays the posts to the user at a single place. First the data is fetched using two different methods one is Web Scraping / Web Crawling and the other one is fetching the data using websites api. Once the data is fetched from the websites it needs to be stored in a database for that purpose, we are using MySQL database. For scraping of data, we use the library called Beautiful Soup.

Once the data is scraped and stored in SQL database we use Django framework for creating a web-based platform and displaying the data from the database. We choose the Django because it is relatively easy to create websites in a short time and without compromising the security of the website. On top of Django the html templates are relatively simple and does the work of displaying title to the users. If the users want to see a particular post, he can click on the title it will take him directly to the post in respective site. The posts have given heading from which the posts are came from and number of comments also for the reddit posts. Along with this user have a functionality to take notes if he finds any post interesting and want to research about it or want to save it, he can simply click on new note and can write title and the link of the post for future study. It will be there permanently until the user decides to delete it. Yes, it has the functionality to delete or update the texts.

2.2.1 Advantages of the proposed system:

- Latest posts from different social media websites can be made available to people.
- It updates you with social media posts time to time when the trending post changes in the websites.
- Get a flexible and simple reading experience with its simple UI/UX.
- Have the notes taking functionality for future reference.

3. Working principle of Socially Smart

3.1 Application Logic:

First it visits the social media websites specified and scrapes the data once done it will store the data in

the SQL lite database deleting the previous data present in it. For this the user needs to give the event triggering click in home page. Once the data is stored in the database the logic will analyse the data and keeps only the high commented and trending posts by deleting the remaining. Now the posts in the database are finally displayed to user in simple UI/UX.

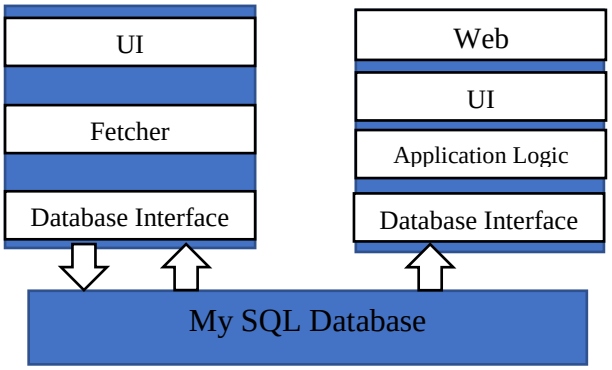

Fig 2. Application Logic

3.2 DATABASE SCHEMA

The schema contains three classes reddit, Onion news and one for the Notes functionality.

The three classes are -

- Headlines - Onion News
- Story - Reddit
- Notes - For Notes App.

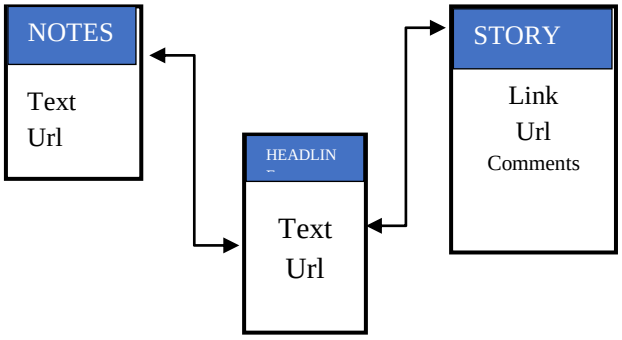

Fig 3. Database Schema

3.3 Web Scraping / Web Crawling:

Web scraping, often called web crawling or web spidering or programmatically going over a collection of web pages and extracting data, is a powerful tool for working with data on the web. It is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table format. Web scraping services is the technique of automating this process, so that instead of manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time.

3.3.1 Web Crawling steps

Here are some basic steps performed by most web crawler:

1. Enter a URL and use a HTTP request to access the URL
2. Now fetch all the contents in the URL and parse the data
3. Store the data in any desired database.
4. Enqueue all the URLs in a page.
5. Use the URLs in queue and repeat from process 1.

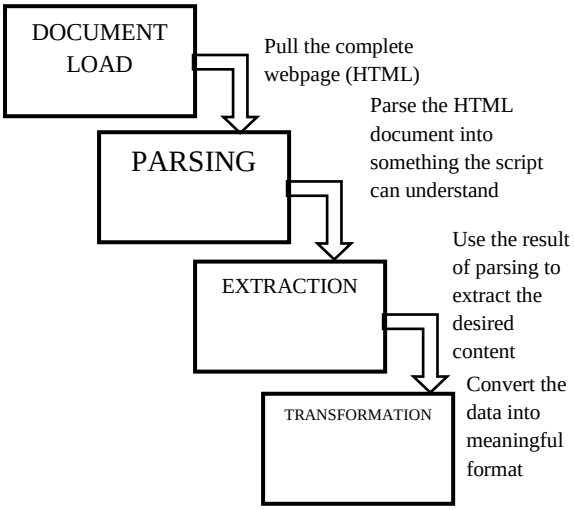

Figure 4. Shows the Stages of Web Crawling

4. RESULTS

The most trending and most commented content will get crawled by a click. Finally, the reader will be able to get the posts from different social media. The user can also take notes and store URL if he wants this functionality is provided. The Figure 4 shows the sample UI of Socially Smart where the posts are categorized by their websites along with comments for reddit posts and the functionality of notes taking.

Fig 5. The sample UI of Socially Smart

5. CONCLUSION:

Socially Smart is a meaningful platform which reduces the user time being spent on social media by fetching all the required data to a single place reducing the time spent on opening every website. Instead of reading all the posts the user gets to read only top posts only so more efficiency is achieved. User can also take notes which altogether makes it a great Social media aggregation platform.

6. FUTUREWORK:

In future one can develop mobile application of this system so that mobile user can easily access this application. Adding new websites such as GitHub, Facebook, twitter and some more news websites will make it very good and flexible.

A research work is going to enhance the Socially Smart which can be personalized and customized according to the reader. For an example, if a reader has interest in sports the sports news will be shown as a priority. And also, additional tab called Education can be added so that the students will get the latest posts about education and current affairs which will help them to crack the competitive examinations

7. REFERENCES:

1. Sandeep Sirsat and Vinay Chavan , " Pattern matching for extraction of core contents from news web pages" IEEE Transaction on Web Research (ICWR),23 June 2016.
2. Kolari P and Joszhi.A, "Application of Web Scraping and Google API service", IEEE Transactions on Knowledge and Data Engineering, Vol.6, No.4, 2014.
3. Deepak Kumar Mahto and Lisha Singh, "A Dive into Web Scraper World", IEEE Transaction on Computing for Sustainable Global Development, Vol.2, No.1, March 2014.
4. Suraj B. Karale and G.A. Patil, "Extracting brief note from Internet Newspaper", IEEE Transaction on Computing for Sustainable Global Development, Vol.45, No.1 March 2016.
5. TehPohey Lee, Abdul Azim Abdul Ghani and Chang Yu Huang,"Survey on application tools of Really Simple Syndication (RSS): A case study at Klang Valley", Vol.3, 2008.
6. S.K. Malik, S.M. Ravi, "Information Extraction Using Web Usage Mining Web Scrapping and Semantic Annotation", IEEE International Conference on Computational Intelligence and Communication Networks (CICN), 2011.
7. Fister, S. Fong, Yan Zhuang, "Data Reconstruction of Abandoned Websites", IEEE 2nd International Symposiumon Computational and Business Intelligence (ISCBI), 2014.

8. Quang Thai Le, D Pishva, "Application of Web Scraping and Google API service to optimize convenience stores distribution", 17th IEEE International Conference on Advanced Communication Technology (ICACT), 2015.

9. Zixuan Zhuang Mehdi Bahrami, Mukesh Singhal, "A cloud-based web crawler architecture", 18th International Conference on Intelligence in Next Generation Networks 978-1-4799-1866-9, pp. 216-223, 2015.