NEWSONE- AN AGGREGATION SYSTEM FOR NEWS USING WEB SCRAPING METHOD

K.Sundaramoorthy
Professor & Head
Department of Information Technology
Agni College of Technology
ithod@act.edu.in

R.Durga
Assistant Professor
Department of Information Technology
Agni College of Technology
emailtodurgasaravanan@gmail.com

S.Nagadarshini
Student
Department of Information Technology
Agni College of Technology
nagadarshini.s@gmail.com

***Abstract*--NewsOne is a dedicated platform that aggregates all the latest news updates from multiple national and international resources and summarizes them to present in a short and crisp words.This online platform provides a service oriented interaction among the users from across web. The main motto of this application is to access the news fast. It will bring news directly without wasting any time for searching and news loading.NewsOne uses web scraping/crawling method to extract the content from various news websites. The basic concept deals with news feeds, the admin or sub admin will add all the news URL into a database. The crawler fetches the content from the RSS feeds of the stored URL. A bot is employed that will dynamically extract the contents at certain intervals.**

This paper describes a model to perform categorization which extracts useful information for classifying a document into category by referring to URL. Users will get a flexible experience on this application.it allows the readers to read the news based on interest. This can be possible by enabling them to choose the categories of news such as Just-In, Technology, Health, Science, Sports, Business and Economics and Entertainment. It provides the reader to read the news for free of cost and in the fastest way possible. Users can get complete up-to-the-minute daily news coverage and headlines from over 100+ fully licensed and trusted news sources nationally and worldwide. Lastly some recommendation and thoughts are laid out for the future development of the work.

Keywords – Web crawling, News aggregator, Scraping, XML parsing

I. INTRODUCTION

The world of web is extremely huge in terms of web pages with large amount of informative contents available in different formats like text, graphical, audio-video, etc. which leads to inconsistency in retrieval of data due to its irrelevance for which the user looking for. As the result, the probability to find the exact information is very nominal. So, when we talk about news, any reader can obtain information which may delivered from newspapers, magazines, television news channel or widely accessed internet provides news from portals, blogs and other social media.

This application acts as a gateway to multiple news sites, organizes news by topic. News One extracts the individual news from the web pages and provides them in a single platform. It improves the quality of results because the contents are short and crisp. It also reduces the searching and reading time. The news feeds, which will be added by the admin / sub admin. About 126 licensed and trusted news sources are aggregated. Users can get complete up-to-the-minute daily news coverage and headlines.

Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured. NewsOne adapts a technique called as Web scraping/Web crawling which is part of Data Science. Web scraping is a computer software technique of extracting information from websites. This technique mostly focuses on the transformation of unstructured data (HTML/XML format) on the web into structured data (database).It is a powerful tool for working with data on the web. With a web scraper, you can mine data about a set of products, get a large corpus of text or quantitative data.There are many software tools available that can be used to customize web-scraping solutions. 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. Using a web scraper NewsOne extracts the news from different websites, stores those information in a database and displays the quantitative information to the users.

II. SYSTEM DESIGN

The NewsOne system architecture is as shown in Figure 1.

978-1-5090-4797-0/17 $31.00 © 2017 IEEE
DOI 10.1109/ICTACC.2017.43

EXISTING SYSTEM:

In the existing system, not all the informations are available under a single roof. Therefore this becomes a problem of time consumption. The user has to verify multiple websites in order to derive at a conclusion. Therefore not all the websites will have the same information, they will have a unified way of presenting their own information. 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 websites which provides news in a consolidated manner. Such as indiapress.org, drudge-report.org, snopes.com. These websites also uses RSS feed and acts as a news aggregator.

Disadvantages

- Provides only the source url of the news. In this system, it just displays the name of the digital newspapers available in the web. The user has to go to each and every link to read the news. Hence it is time consuming one.

- Does not state any description about the fact or news. It provides only the title of the news. The user doesn't get the information what he is looking for. Thus, it leads to chaos.

- Not Categorized. This System provides the latest news in a short and crisp manner. But the disadvantage is that it does not categorize the news. Thus, the reader couldn't able to read the news on the favourite subject of interest.

Proposed System:

In order to get the daily news in an individual news website is a time consuming activity. User cannot visit each news web site separately to get the latest and trending news on different topics of interest. Thus, one can use news aggregators are to bring all the current breaking news from multiple channels into a single dashboard. NewsOne is an innovative news aggregator that offers unique features making it a type of RSS feed reader and in a unique way to discover amazing new content around the web. It offers a simple and refined user interface where reader can easily access all your favourite content from around the web. From around 100+ sources, reader can find the exact news you are looking for within minutes. User can read all the news they want on their favourite topics in this system.

This multi-layered format of our NewsOne website application encourages users to pursue their own path and select information that matches their interests. It is used for creating a unified website application for people those who are very curious to keep on updating their knowledge with latest news. This proposed system also helps the people who are preparing for competitive examinations like IAS, IBPS, and other government exams and it is also helpful for those who are participating in quiz contests or interviews where the questions will be asked from the current affairs. It also shows the weather updates on the current location.

This website doesn't produce much original content, but instead it will provide 'curate' content created by others which is done by using a combination of human editorial judgment and computer algorithms. The results are presented with a short description with images from the original article. In order to read the full news article, users can click through to the website of the original content creator. The websites such as Yahoo News and MSN mainly show content from contractual partners. NewsOne is a 'Pure' aggregator, which is similar to Google News. It generally does not make any payments to the original authors of the news content. This website is made by 'crawling' the web and then using some algorithms and editorial judgments to organize the content.

A user may forget about smaller niche websites, such as local news sites. Instead, a user may focus on popular sites such as The Hindu and the Times of India, together with a few personal favourite news websites. This affects the user of not gaining wider knowledge. Thus, The NewsOne Aggregator might allow users to become informed about the quality of a wider range of outlets, leading them to outlets that match their interests more closely. This in turn, increases a user's overall news consumption as well as their consumption of news aggregators.

Advantages of the proposed System

- Latest news from different sources can be made available to people

- It provides news and content in its broad categories (Just-In, Technology, Business and Economics, Sports, Science, Entertainment, Health) from across the world.

- It updates you with breaking and trending news across the day

- Get a new reading experience with its innovative and simple UI/UX.

3. Working principle of NewsOne

The NewsOne is based on Client-Server architecture. The client and server interact through the internet connection. The front end is a web browser where the client sends a site page URL. The request hits the application server. The application server maintains a hypertext link pool where 100+ URL is present. The application server fetches the content from those URLs using a web scraping or web crawling technique. The content in those links gets updated dynamically. These contents are stored in a database server. The server fetches the content from the data base and provides response to the client. This application is hosted in a Cloud.

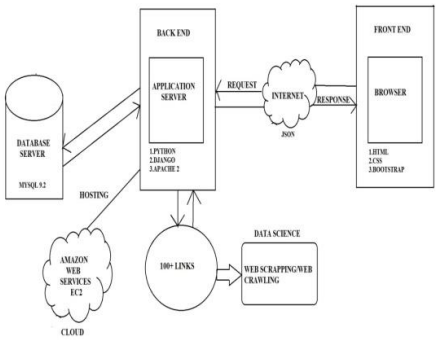

Figure 1. Architecture of NewsOne Application

3.1. Web Module

3.1.1 User Interface

The web module will give interface to the client. By utilizing this interface the client can send its demand for web daily newspaper. User gets a new reading experience with its innovative and simple UI/UX. NewsOne Web Application has following categories.

- Just In
- Business and Economics
- Tech (Technology)
- Entertainment
- Science
- Health
- Weather
- LifeScoop

A very simple and easy to use User Interface. Just by clicking the desired category the reader can get the updated news of that particular domain.

3.1.2 Application Logic

The tier architecture provides an architectural style that allows building s flexible and reusable web application. The most used style is the three tier architecture. In this type of architecture the presentation logic, business logic and data handling are realized as separate tiers.

The RSS fetcher is responsible for retrieving RSS feeds from selected sites at specified time intervals. It is written in Python and uses Beautiful Soup library to parse RSS feeds, and MySQL DB to store each parsed RSS document into a MySQL database.

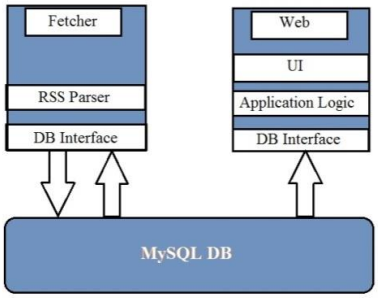

Figure 2. Represents the application logic of NewsOne

The fetcher is multi-threaded and is scheduled by Cron to run every 12 hours.The web application is responsible for interacting with the user and presenting them with the previously fetched RSS documents according to their interest.

3.2. Database Schema

The Schema contains three classes. Each class contains ID as its primary key.

The three classes are

- Category
- Link
- Item

The class Category contains name of various categories of news along with its ID. The class Link contains a pool of 126 URLs along with its ID and Category as its reference. The class Item contains title, description, image, date and author. The Link used as references in the class Item.

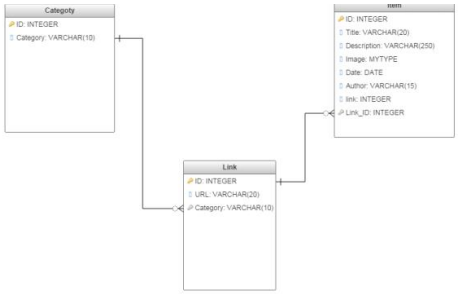

Figure 3. A schema representation of NewsOne Web aggregator

3.3. Web Scraping/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.

Our goal here is to get links (url) from various news websites and the associated data. Once those links are scraped, they are categorized accordingly.

The website links crawled using BS4 or scrapy are then categorized and stored in database according to the category. This crawling is triggered automatically twice a day like normal morning/ evening newspaper. The trigger does not have any connection with Front End Browser, the process takes place only with Back End App Server connection.

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.

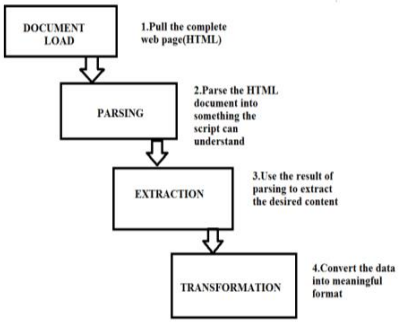

Figure 4. Shows the Stages of Web Crawling

1. Document Load

It uses theRequest/response handlers which are managers that make http requests to a group of urls, and fetch the response objects as html contents and pass this data to the next module. Python uses opening process libraries for performing request/response url.

2. Data parsing/data cleansing

This is the module where the data which is fetched is processed and cleaned. Here the unstructured data is transformed into structured during this processing.

3. Data Extraction

In this module the idea behind is web scraping which is to retrieve data that is already exists on a website and convert it into a format that is suitable for analysis. Web-Pages are rendered by the browser. BS4 is essentially a set of wrapper functions that make it simple to select common HTML/XML elements.

4. Transformation/Serialization

After you get the cleaned data from the parsing and cleaning module, the data serialization module is used to serialize the data according to the data models. This is the final module that will output data in a standard format such as JSON that can be stored in databases.

4. IMPLEMENTATION TECHNIQUE

Parsing or syntactic analysis is the process of analysing a string of symbols, either in natural language or in computer languages, according to the rules of a formal grammar.

4.1 THE XML DOM TREE GENERATION

XML parsing is basically taking in XML code and extracting relevant information like the title of the news, link of the website, creator name, published date, description of the news with an image. XML parsing is the process of taking raw XML code, reading it, and generating a DOM tree object structure from it.

The general idea behind web scraping is to retrieve data that already exists on a website and convert it into a format that is usable for further analysis. BS4 is a set of wrapper functions that is used to select XML and HTML elements. It is a class that is used to parse the XML files directly.BS4 or BeautifulSoup is a DOM based tool in which the parser makes a single sequential pass through the file to parse the XML file. The parser does not save any of the tags or the contents inside the tags. So it leads to very fast parsing because the XML file contents is not changed by the parser and the parser makes only one pass through the file.In contrast, bs4 class constructs a DOM (Document Object Model) object. It means that the entire contents of the XML file are stored in memory. DOM is a convention used in HTML, XHTML, and XML for representing and interacting with objects. The elements in an XML document may have attributes. Even though it is a slower form of parsing, it allows making changes to the contents of the XML file. The bs4 uses mainly two kinds of objects to perform XML parsing. The Objects are BS4 and tag in order to do XML parsing using bs4. The BeautifulSoup is an object that holds the entire contents of the XML file in a tree-like structure. The tag is an object that stores a HTML or XML tag. The tag object contains number of attributes and methods that manipulates the XML file easily.

4.2. ALGORITHM FOR SCRAPING THE CONTENT FROM XML USING BS4

1. Import the necessary libraries for scraping such as bs4 to parse the data returned from the website

2. Fetch the links of the url using urllib2 library and save it in a variable called page

3. For each link do

a. Parse the XML in the page variable and store it in a BeautifulSoup format.

b. For each data in the item tag do

i. Scrap the title, published date, creator/author, link, description and image.

4. Save the scraped content into the database

5. Set the apscheduler to schedule jobs to run periodically at fixed times, dates, or intervals.

5. RESULTS

The meaningful content will get crawled automatically by a bot. Finally the reader will get the updated news from various news sites in a single website. The Figure 5. Shows the sample UI of NewsOne where the news are categorized into 9 categories and also you can able know the weather condition of your location.

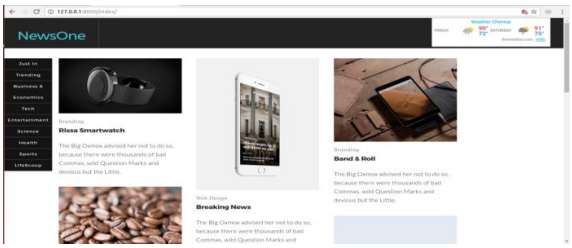

Figure 5. The Sample UI of NewsOne

6. CONCLUSION AND FUTUREWORK

The bigger challenge faced by a reader is that, they don't get the enough clarity on news related of their particular field of interest. So the need arises to fetch information using search engine which leads to more inconvenience in exact findings from the different sources over web and the contents available on web pages speaks a lot on massive topics.

In general, which may results complexity. In most of the cases where the reader's choice might be expecting the news which one wants to read. As a solution the digital newspapers introduced which provides valuable information to the readers. Each news websites contains various different perspective of news and it will be lengthy which is updated frequently. In today's life we do not have enough time to read each and every content of the newspaper from various sources. Hence, the reader prefers some important and summarized information. NewsOne provides a convenient way to keep up with changes on news sites, without re-visiting those sites manually to look for latest news.

In future, one can develop the mobile app of this system so that the mobile user can also easily access this application. A research work is going to enhance the NewsOne 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 news 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.