A SEMI- AUTOMATIC DATA SCRAPING METHOD FOR NEWS EVENT DETECTION

Bhavaniya[1], Haripriya[2], Karthika[3,] Jeyalakshmi[4]

UG Students[123],Assistant Professor[4] , Department of Information Technology , SRM Valliammai Engineering College , Kattankulathur , Kanchipuram , Tamil Nadu , India

ABSTRACT

A simple Approach for news extraction using web scrapping smart parser is useful system which can be used to get the top stories from the top news sites. As the internet and World Wide Web continue to expand more users, increased rate of crime occurs. Web Scraping tools is used to extract and analyse the information on these web sites to identify traffickers. The proposed system for this project is a web scraper that is able to access and extract data from web sites using a web application as an interface for user interaction. The extracted data is then stored in a database as the web application allows the user to search through and query the saved findings. when the system has been fully implemented, it restructures the data and displays to the user.

Keyword: Extractor, Parser, HTTP request

---

1. INTRODUCTION

World Wide Web contains the different type of data and storing huge heterogeneous data online, it provides a main issue of extracting information at the limited time. To get required information easily, efficiently and correctly a strong concept is needed that easily mine the required information within fraction of seconds. This vast amount of information is used by ordinary web users through out the world and the automated crawlers that traverse the web for various purposes, such as web mining.Hence in the proposed system the news which displays from multiple newspaper in single application, also this system provides the news which is regional to the user.

1.1 OVERVIEW

Data extraction is the method of retrieving data from unstructured or structured data sources for data processing. The majority of data extraction is getting in the form of unstructured data sources and different data formats. This unstructured data can be in any form such as tables , indexes,text and numeric. After receiving a user query a web database returns the relevant data values in structured format.

2. LITERATURE SURVEY

2.1 A Review on Web Scrapping ands its Application - Vidhi Singrodia, Anirban Mitra and Subrata Paul International Conference on Computer Communication and Informatics , 2019.

This paper focuses on various aspects of web scrapping, beginning with the basic introduction and a brief discussion on various software's and tools for web scraping. It also explains the process of web scrapping with an elaboration on the various types of web scrapping techniques and finally concluded with the pros and cons of web

scrapping and in detail description on various fields where it can be applied. These data includes Open Government Data, Big Data, Business Intelligence, aggregators, comparators, and development of new applications.

2.2 Web Crawling - Based Search Engine using Python - Sanya goel, Mudit Bansal, Atul Kumar Srivastava, Neha, International Conference on Electronics, Communication and Aerospace Technology, 2019.

A data mining powdered search engine is used for education sector. Acquiring information on schools and colleges from the internet is big task also many institutes can be missed as they don't have good SEO.Internet based services for educational institutes with web site crawler has many features to give statistics. The decision making is made easier about the detailed information of education institutes nearby using location-based search, important dates, documents, online forms, contact numbers and the procedures.

2.3 Resource Description Framework Generation for Tropical Disease using Web Scrapping - Amalia. A, Afifa. R. M and Herriyance, IEEE International Conference on Communication, Networks and Satellite, 2018

Introduction Tropical diseases are diseases that commonly occur in tropical areas like Indonesia. These people usually depend on a search engine to search the information about diseases and drug especially for general illness like coughs, colds and fever. our solution is to build a search engine based on semantic web technology. This research aims to generate an RDF serialization that describes the relationship of tropical diseases and treatments terms. Thus the implementation of a web scrapping method to extracting vocabularies from some popular health websites.

2.4 A Specific Process to Generate Datasets Containing Public Transport Accessibility Information – Caceres . P , Sierra-Alonso. A , Vela. B , Cavero . J.M and Cuesta .C.E , International Conference, 2018.

Google Maps is a relavant tool that is employed to calculate routes and find points of interest , while Google Transit Feed Specification is a format used to specify public transport agents to provide a feed complying with the specification. Google maps does not provide detailed information about specific facilities such as accessibility information . It is difficult to find specific and detailed accessibility information that can be downloaded and processed , thus the systematic method is used for extraction of data from the internet to generation of open nd linked datasets.

2.5 Cloud Based Web Scrapping for Big Data Application- Chaulagain.R.S, Pandey.S , Basnet.S.R and Shakya.S ,International Conference on Smart Cloud , 2017.

There is a rapid growth of internet users and data generated by those users on the internet. Several challenges are faced due to scraping large amount of data such as encountering captcha. This paper describes about the cloud-based web scraping architecture that able to handle storage and computing resources with elasticity on demand using Amazon web services. Selenium is one of the tool for web scraping because of web drivers which supports and stimulates a real user working with a browser.

2.6 Web scraping and storing data in a Database , a case study used in car market - Dorde Petrovic , Ilja Stanisevic , International Conference on Telecommunication Forum in 2017.

Searching the web is to extract the useful data and information has become a routine job. The data on the sites can be found in tables , articles , comments , nested in different HTML tags. It is good way to collect information which can be used for further analyzes, thus this process of web scraping of data from different locations on internet and their storage in a database ,for the purpose of collecting and analyzing data of the used cars market.

3. EXISTING SYSTEM

Currently someone needs to reads the news then they must visit the particular websites. In advance to the websites are cause plenty of advertisements. That is the main issues with the current system. Hence a powerful and efficient

technique to extract useful and important information from the huge bulk of data present. Web mining is a data mining technique which discover the hidden data in web log. The users experiences heavy loss of time and knowledge discovery consumes a lot of system resources.

3.1 DISADVANTAGES

- Introduction Storage issue for a large amount of data
- Catching specific data fails in certain condition
- It contains several link before getting the actual page .
- Browsing several web page to get needed information
- Enormous amount of new data can uploaded

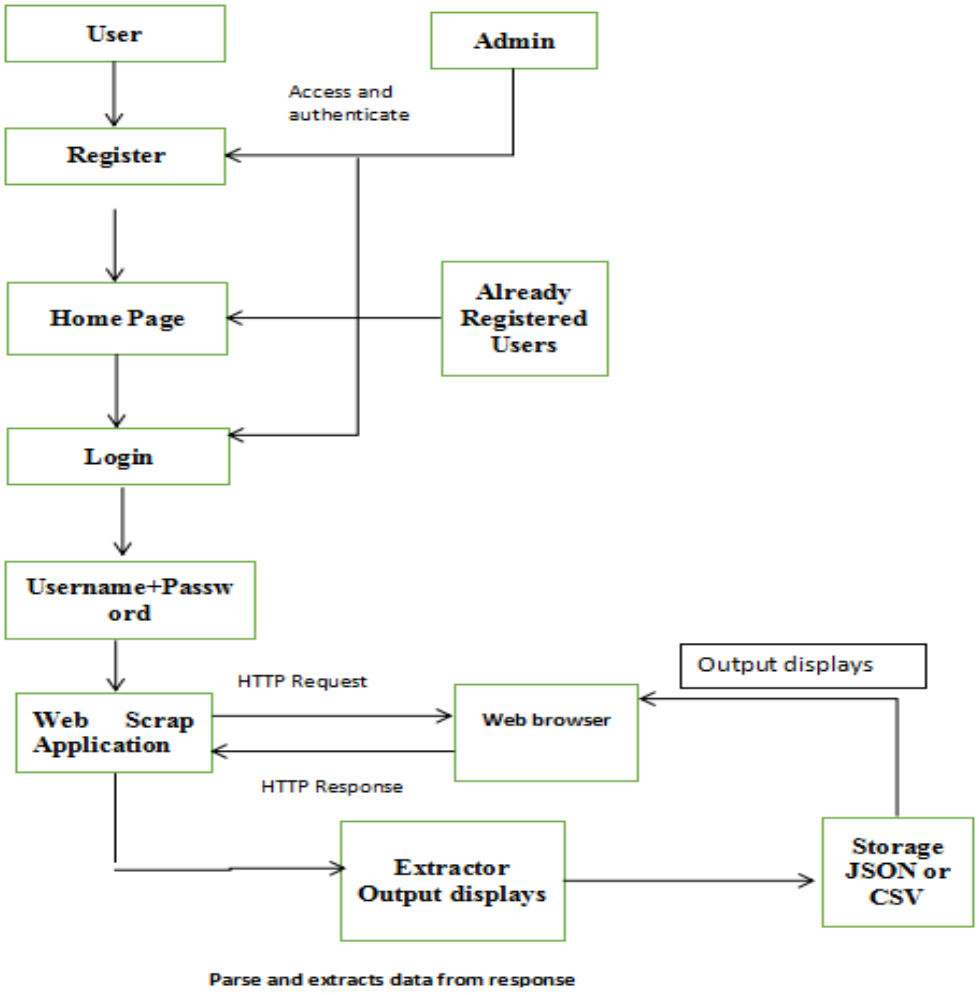
4 . PROPOSED SYSTEM

The smart parser scrapes the top news and latest news from the top news sites and displays it to the user using the web scraping technologies. This method is used for the extraction of information applied in entire webpage and they contain information only from the required content blocks of web page. A user menu will be given and from the menu user can select any of the site and read the news from that site, in addition to this user can also select any one of the news website as default to be displayed in the home page. Web scraping is the major technology used to develop the sytem. The most important feature of this system is that the regional news based on users location is also scraped and displayed in the application.

4.1 ADVANTAGES

- An essential service at low cost.
- Able to extract data from multiple resources
- Searched information is contained in single page
- Easy to implement.
- It consume data in less time with low internet
- Extraction of data is more accurate.

5. BLOCK DIAGRAM:



6. HARDWARE REOUIREMENTS

- 1 TB Harddisk
- 8GB Ram
- Intel core i5processor
- Internet connection
- 1.44MB Floppy drive

7. SOFTWARE REOUIREMENTS

- Anaconda IDE

- Windows 10 OS
- Python language
- SQLite

8 .IMPLEMENTATION

The implementation part of the system starts by caught the URL of the website. After taking the link of each information of the system, scraps the content of those links. System keeps monitoring the change in the contents of the site. System can be implemented using a python library known as Beautiful soup. It can also be used to implement the monitoring of changes.Document Object Modelling defines the style , structure and content of the xml file. Web site parser is a tool that will allow you to gather information through the parameters that have created. System scraps the contents from the site using html tags, css classes and ids. Finally restructures the data and displays it to the user. The summary of the scraped content is provided using natural language processing toolkit.

10. CONCLUSIONS

Web is a rapid growing research area it consists of Web usage Web structure and the Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web content mining aims to extract/mine useful information or knowledge from web page contents In our paper we try to identify the problems path and prevent the unauthorized accessibility.

11. REFERENCES

[1]. Web Crawling - Based Search Engine using Python - Sanya goel, Mudit Bansal, Atul Kumar Srivastava, Neha, International Conference on Electronics, Communication and Aerospace Technology, 2019.

[2]. A Review on Web Scraping and its Application - Vidhi Singrodia , Anirvin Matra , Suprata Paul ,International Conference on Computer Communication and Informatics ,2019

[3]. Recursive Stock Price Prediction with Machine Learning and Web Scraping for Specify Time Period - Ayush Ray , Aman Upadhyay , Bhupesh Gour , Asif Ullah khan, International Confrence on Wireless and Optical Communication networks ,2019.

[4]. Exploiting Filtering Approach with Web Scraping for Online Shopping- Shakra Mehak, Rabia Zafar , Sharaz Aslam , Sohail Masood Bhatti, International Conference on Computing , Mathematics and Engineering Technologies, 2019

[5]. Resource Description Framework Generation for Tropical Disease using Web Scrapping - Amalia. A, Afifa. R. M and Herriyance. H, IEEE International Conference on Communication, Networks and Satellite, 2018.

[6]. A Specific Process to Generate Datasets Containing Public Transport Accessibility Information – Caceres . P , Sierra-Alonso. A , Vela. B , Cavero . J.M and Cuesta .C.E , International Conference, 2018.

[7]. Cloud Based Web Scrapping for Big Data Application- Chaulagain.R.S, Pandey.S , Basnet.S.R and Shakya.S ,International Conference on Smart Cloud ,2017

[8]. Web scraping and storing data in a Database , a case study used in car market - Dorde Petrovic , Ilja Stanisevic , International Conference on Telecommunication Forum, 2017.