Web Crawling-based Search Engine using Python

SANYA GOEL
Electronics and Telecommunications
ASET, Amity University, Uttar Pradesh,
Noida, India
goelsanya38@gmail.com

MUDIT BANSAL
Electronics and Telecommunications
ASET, Amity University, Uttar Pradesh,
Noida, India
muditbansal95@gmail.com

ATUL KUMAR SRIVASTAVA
Dept. of Electronics and Telecommunications
ASET, Amity University, Uttar Pradesh,
Noida, India
aksrivastava1@amity.edu

NEHA ARORA
Electronics and Telecommunications
ASET, Amity University, Uttar Pradesh,
Noida, India
narora2@amity.edu

Abstract*—A data mining powered search engine for the help of education sector. Getting information on schools and colleges from the internet are a big task also many institutes can be missed as they don't have good SEO. Internet based services for admission in school and colleges with site crawler and many features to give statistics. The decision making is made easier with easy access to detailed information of educational institutes nearby using location-based search, important dates, documents, online forms, contact numbers and the procedure. There are websites like www.shiksha.com and www.pathshala.com tells about the college admissions and certified courses, they are crawling over 14,000 institutions, 40,000 plus courses and have a registered database of more than 3.5 million students. So, taking inspiration from them we are trying to create our own website for school admission in Delhi NCR.*

Keywords—Web crawling, Web scrapping, Topical crawling and Web content mining

I. Introduction

The website is based on educational background. Every parent/student will have personalized experience on website as getting admissions in schools and colleges is a big task. Internet based services for admission in school and colleges with site crawler and many features to give statistics.

The decision making is empowered with easy access to detailed information of the schools near by according to their GPRS filter, important dates & documents, online forms, contact numbers and the procedure.

There are websites like www.shiksha.com and www.pathshala.com tells about the college admissions and certified courses. They are crawling about 15,000 institutions plus courses and have a registered database of more than 3.51 million students. So, taking inspiration from them we are trying to create our own website for school admission in Delhi NCR.

Main objective is to that we should be able to crawl the useful information about the school and provide it to parents.

Aim is to make this website one-stop-solution for selection and admission of students. This could be real time world thing. We can expand this website by providing the school details from all over the country.

One major goal is to make this website India's top educational website by providing what all information parents need. And after some time, we could have active ask and answer community where experts can help the parents for selection and other requirements. And over the years service providers can also gain profit.

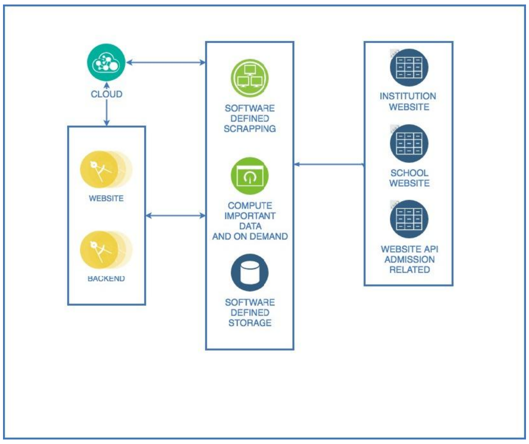
CLOUD
SOFTWARE DEFINED SCRAPPING
INSTITUTION WEBSITE
WEBSITE
COMPUTE IMPORTANT DATA AND ON DEMAND
SCHOOL WEBSITE
BACKEND
WEBSITE API ADMISSION RELATED
SOFTWARE DEFINED STORAGE

Fig 1.

II. METHODOLGY

- **WEB SCRAPPING**

Web scrapping is a programming system of extricating data from sites. This method for the most part centers around the change of unstructured information (HTML design) on the web into organized information (database or spreadsheet).

The automated indexing of in data from the Internet is about there for ancient internet days. Although web scrapping is definitely not another term, in years past the

978-1-7281-0167-5/19/$31.00 ©2019 IEEE

training has been all the more ordinarily known as screen scratching, information mining, web collecting, or comparable varieties. General accord today appears to support web scrapping.

There are a few different ways to scrap data from the web. Utilization of APIs being presumably the most ideal approach to extricate information from a site. All vast sites like Twitter, Facebook, Google, Twitter, Stack Overflow give APIs to get to their information in a progressively organized way. On the off chance that you can get what you require through an API, it is quite often favored methodology over web rejecting. This is supposing that you are gaining admittance to organized information from the supplier, for what reason would you need to make a motor to extricate a similar data.

Web scratching is the act of social occasion information through any methods other than a program interfacing with an API (or, clearly, through a human utilizing a web program). This is most normally done by automated bots (spiders) which reads a web server, demands information (generally as the HTML and different documents that involve web pages), and after that parses that information to separate required data.

- **WEB CRAWLING**

Web crawler is a web bot that is utilized for web ordering in World Wide Web. All kinds of web search tools use web crawler to give proficient results. Actually, it gathers all or some explicit hyperlinks and HTML content from different websites and see them in an appropriate manner. When there are tremendous number of connections to slither, even the biggest crawler fails. For this reason, web indexes mid 2000 were terrible at giving significant results, but now this procedure has enhanced much, and legitimate outcomes are given in a moment

The web crawler here is made in python3. Python is high level programming language that includes object-oriented, basic, practical programming and a huge standard library. For the web crawler two standard libraries is utilized-solicitations and BeautfulSoup4 demands gives a simple method to associate with internet and BeautifulSoup4 is utilized for some specific string activities.

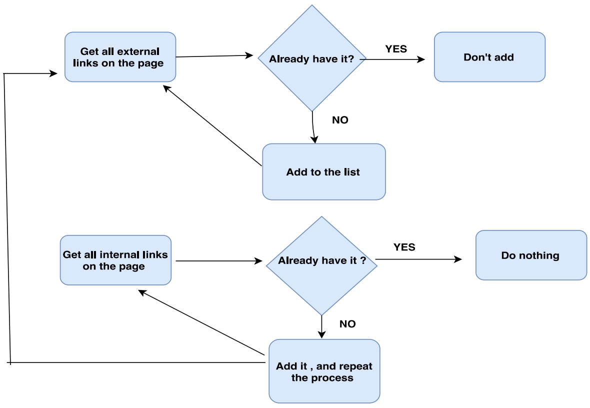
Get all external links on the page
Already have it?
YES
Don't add
NO
Add to the list
Get all internal links on the page
Already have it ?
YES
Do nothing
NO
Add it , and repeat the process

Fig 2

Modules used for web scrapping and web crawling are:

Beautiful Soup4

Beautiful Soup is a Python library for getting information out of HTML, XML, and other markup dialects. Let's assume you've discovered a few website pages that show information applicable to your examination, for example, date or address data, yet that don't give any method for downloading the information specifically. Beautiful Soup encourages you pull specific substance from a site page, expel the HTML markup, and spare the data. It is an apparatus for web scratching that causes you tidy up and parse the archives you have pulled down from the web.

Beautifulsoup4 is used for extracting names and URLS from an HTML page. At starting of the script import the library. After that create a soup object which need to pass from beautifulsoup. It doesn't fetch the webpage. So, for this we use urllib2 in combination with this.

Urllib.request

Urllib is a Python module that can be utilized for opening URLs. It characterizes capacities and classes to help in URL activities.

With Python you can likewise get to and recover information from the web like XML, HTML, JSON, and so on. You can likewise utilize Python to work with this information straightforwardly. In this instructional exercise we will perceive how we can recover information from the web. For instance, here we utilized a guru99 video URL, and we will get to this video URL utilizing Python and also print HTML document of this URL

Parsing the data

Given page information, we need to remove intriguing data. You could utilize the BeautifulSoup module to parse the returned HTML information. Python is one of the dialects that is broadly used to scrap information from website pages. This is a simple method to accumulate data.

III. WORKING

First, the site mapping code crawls websites and the data is made available for the scrappers to selectively work on the provided links. Then the data found from the scrapper and crawler are kept in managed database.

The website is PHP based and the server end processes are handled by python programming, making the whole operation light for processing reducing the cost and time drastically

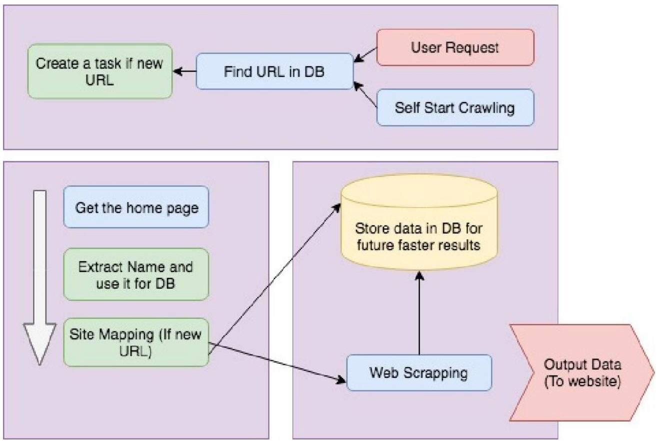

Fig 3

IV. Conclusion

The search engines can be powerful tools when created with versatile core systems, Search portals powered with smart web crawlers, given the computing power can scrap the whole internet to find the required data, which can be helpful for analysts and researchers.

In this case the search engine will always display more options for institutes than others as it’s going to scrap the internet and doesn’t need direct human data feeding.

Some basic maintenance and the web portal can be working 24/7 providing information to the users.

V. References

[1] S. Amudha, “Web Crawler for Mining Web Data” in International Research Journal of Engineering and Technology Volume: 04 Issue: 02, Feb -2017

[2] Ayar Pranav, Sandip Chauhan, “Efficient Focused Web Crawling Approach for Search Engine”, International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 5, May 2015

[3] Mini Singh Ahuja, Dr Jatinder Singh Bal, Varnica, “Web Crawler: Extracting the Web Data”, International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014

[4] Nemeslaki Andras, Pocsarovszky Karoly, “Web Crawler Research Methodology”, 22nd European Regional ITS Conference, Budapest, 18-21 September 2011

[5] Christopher Olston, Marc Najork, “Web Crawling”, Foundations and Trends in Information Retrieval Vol. 4, No. 3 (2010) 175–246

[6] Mike Thelwall, “Methodologies for Crawler Based Web Surveys”

[7] Vandana Shrivastava, “A Methodical Study of Web Crawler”, VandanaShrivastava Journal of Engineering Research and Application, Vol. 8, Issue 11 (Part -I) Nov 2018

[8] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik, “Study of Web Crawler and its Different Types”, IOSR Journal of Computer Engineering, Volume 16, Issue 1, Ver. VI (Feb. 2014)