International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

EXAM NOTIFICATION GENERATED BY AUTOMATIC SCRAPING OF E-MAIL IDS AND INFORMATION FROM WEBSITES

Shreya Gupta[1], Shrishti Panwar[2], Shivani Singh[3], Shraddha Gautam[4], Deepika Gupta[5]

[1-4]Student, Department of Computer Science Engineering MIET, Meerut, UP, India
[5]Assistant Professor, Department of Computer Science Engineering MIET, Meerut, UP, India

---

Abstract - *Our objective is to develop an online application "Exam notification generated by automatic scraping of E-mail ids and information from websites" which is used to send engineering entrance exam notifications to schools by extracting e-mail ids of the school. This application is developed to automate the task of e-mail ids and engineering entrance exam information extraction.*

This online application will use the JSoup library in java in order to scrap the e-mail id of the schools of a particular area [2]. All the relevant information about the engineering entrance exams that are being conducted after 12th class in India will be extracted by understanding the HTML code and then the notification will be send to the various schools of a particular area using java mail API [1]. Before this, the task to extract the e-mails id of the schools is very tiring. Also we have to search on the websites of different colleges for the information about the entrance exam which is complicated as well as time consuming.

KEYWORDS: Jsoup, Java Mail API, HTML, E-mail ids, Engineering entrance exam

1. INTRODUCTION

In today's world, the technology is becoming advance. Every person wants to get their work to be completed in less time and with high efficiency. The person usually uses the internet for the purpose of communication and gaining knowledge or to find out the information. E-mail is best way of communication when it's about formal communication. It is easy when one wants to send the e-mail only to a few person whom e-mail ids are known. But when it comes to send the e-mail to all the schools of a particular area whose e-mail ids are not known then the process becomes very time consuming and complicated.

The students studying in 12th class searches for the information about various engineering entrance exam. For this purpose, one has to go to websites of particular colleges for the infromation which is very tiring. There may be chances that the student may miss some college. Even due to many reasons, there may be a change in the date of application, exam etc.

The first that comes to mind is to use the technology for making the extraction process efficienct. For this purpose, we have designed an online application by using which we can send the notification about engineering entrance exams of different college to all the schools so the students don't have to search for it and school can convey the information to them. This application uses real time filter which extract e-mail ids of different schools and the relevant information about various engineering entrance exams just by providing the URL of the website [2].This application understands the value of time and identify the broken links on a web page that are no longer working because of the various reasons. The notification of the information like college name, exam name, date of application, exam date etc is send to different schools at just a single click.

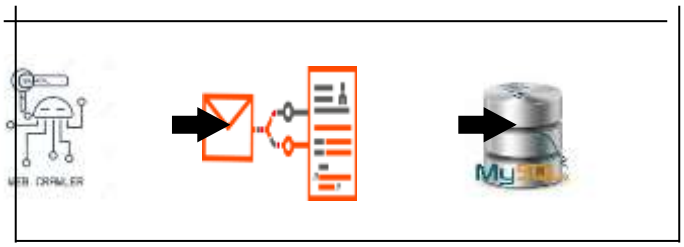

Fig -1: E-mail ids and information parsing

The extractor that are currently available takes the text of only a webpage and if the necessary information is not available on that web page, they return nothing. But our application is capable of extracting all the relevant email ids and information available that one wants to extract. The application we designed automates the task of extracting emails and information. The basic technologies on which we are going to work are java i/o stream, JSoup, MySql, JDBC and java mail API.

2. PROPOSED WORK PLAN

2.1 FLOW CHART OF HOW THIS ONLINE APPLICATION IS DESIGNED

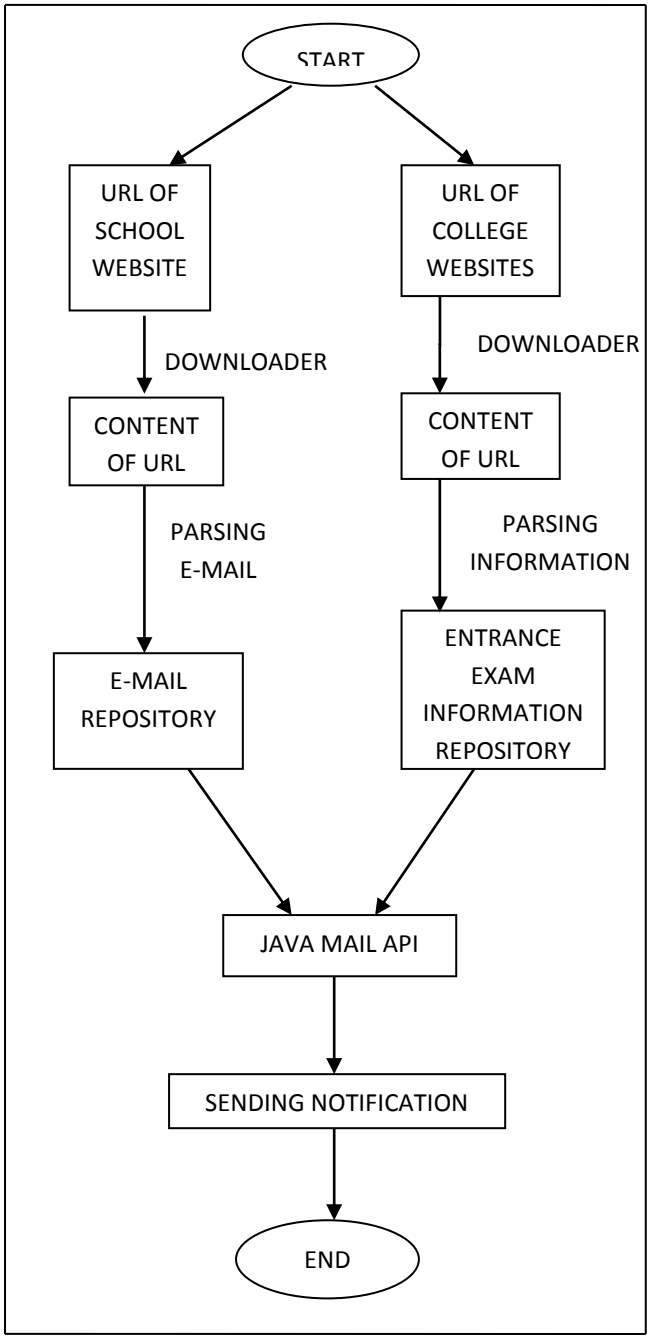

Fig -2: Flow Chart of Robotic Process

2.2 DESCRIPTION OF VARIOUS MODULES

This online application is divided into the following modules –

a. CONTROLLER MODULE

This module focuses on the Graphical User Interface (GUI) designed for the online crawler and is chargeable for dominant the operations of the crawler. The interface allow the user to enter the beginning address, enter the maximum number of URL's to crawl, read the URL's that are being fetched. It basically controls the Fetcher and Parser.

b. FETCHER MODULE

This module starts by fetching the page according to the beginning address or start URL specified by the user. The fetcher module can help to retrieve all the links in a specific page and continues doing that till the maximum number of URL's is reached.

c. PARSER MODULE

This module parses the URL's fetched by the Fetcher module and saves the contents of the fetched pages in the database [3].

d. E-MAIL SENDING MODULE

This module allow user to compose the mail. The user is required to give his/her e-mail id and password, then starts by connecting the e-mail repository and the entrance exam information repository to the java mail API and sending the notifications through mails [1].

2.3 ALGORITHM USED

Algorithm to Fetch Official Websites of the Schools:

Step 1: Read the URL of List of Schools given by the user.
Step 2: Read the content of the URL with the help of URL content fetcher algorithm.
Step 3: Define a regular expression to identify website hyperlinks.
Step 4: Apply logic for pattern matching and store the hyperlinks in a set.
Step 5: Read hyperlinks from the set from Step 4.
Step 6: Read the content of the URL with the help of URL content fetcher algorithm.
Step 7: Define a regular expression to identify official website hyperlinks.
Step 8: Apply logic for pattern matching and store the hyperlinks in another set.

Description. First our algorithm read the content of the user given URL of Schools list and then it finds the hyperlinks from the website content. Then it read the content of fetched hyperlinks one-by-one and finds the official website links from the fetched hyperlinks.

Algorithm to Find the E-mail addresses of the Schools:

Step 1: Read URL of Schools official website.
Step 2: Read the content of the URL with the help of URL content fetcher algorithm.
Step 3: Define a regular expression to identify contact us hyperlinks.
Step 4: Apply logic for pattern matching and store the contact us hyperlinks in a set.
Step 5: Read hyperlinks from the set from Step 4.
Step 6: Read the content of the URL with the help of URL content fetcher algorithm.
Step 7: Define a regular expression to identify e-mail id of the school.
Step 8: Apply logic for pattern matching and store the e-mails in another set.

Description. First our algorithm read the content of the fetched official website link and then it finds the hyperlinks of the departments from the website content. Then it read the content of fetched hyperlinks of the departments one-by-one and finds the e-mail addresses of the faculties from the fetched department hyperlinks.

Algorithm to Find the Information of Engineering entrance exams:

Step 1: Read URL of College official website.
Step 2: Read the content of the URL with the help of URL content fetcher algorithm.
Step 3: Define a regular expression to identify entrance Exam hyperlinks.
Step 4: Apply logic for pattern matching and store the entrance exa,m hyperlinks in a set.
Step 5: Read hyperlinks from the set from Step 4.
Step 6: Read the content of the URL with the help of URL content fetcher algorithm.
Step 7: Define a regular expression to identify the information about the entrance exam.
Step 8: Apply logic for pattern matching and store the information in another set.

Description. First our algorithm read the content of the fetched official website link and then it finds the hyperlinks of the entrance exam from the website content. Then it read the content of fetched hyperlinks of the entrance exam one-by-one and finds the information about the entrance exam from the fetched hyperlinks.

Algorithm for URL Content Fetcher:

Step 1: Create object of URL and URLConnection class.
Step 2: Create object of BufferedReader class.
Step 3: Read the content of the webpage and store it in a string buffer.
Step 4: Return the webpage content.

Description. First our algorithm establishes the connection to the website with the help of URL and URLConnection classes, reads the content of the website with the help of BufferedReader class and stores the content in a string buffer.

3. RESULT ANALYSIS

We are using precision to estimate the efficiency of the proposed system. It is the ratio of necessary information fetched to all of the information fetched. After applying the proposed steps on the given URL address and comparing the results with the desired algorithm, we can say that the given method gives more precision results.

PRECISION RATE = Correct e-mail ids fetched / Total e-mail ids fetched

Table -1: Result Analysis

S. NO.	URL	Total e-mail id Fetched	Correctly Identified e-mail id	Incorrect Result Fetched	Precision Rate
1	http://klischool.com	35	27	8	0.77
2	http://dpsmeerut.in	19	15	4	0.78
3	http://vardhamanacademy.com	51	45	6	0.88
4	http://meerutpublicschool.edu.in	39	33	6	0.84
5	http://mountliteraschool.org	38	35	3	0.92

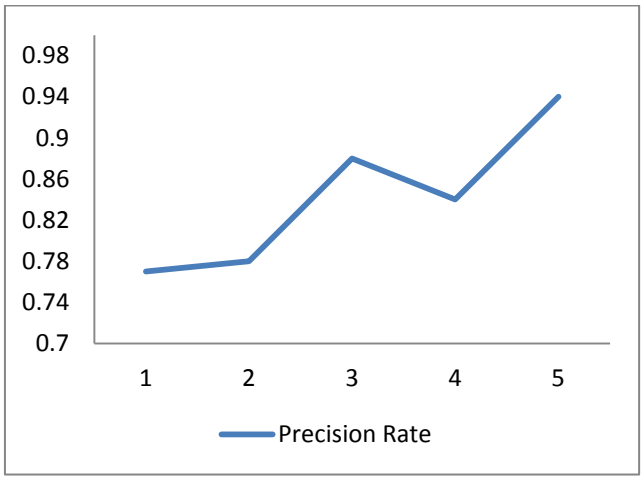

Chart -1: Line Chart for Precision Rate

4. CONCLUSION

This application of automatic scraping is capable of extracting emails and entrance exam information available that one wants to extract from the given URL. When it comes to send e-mails in bulk it is very difficult to send e-mails one by one. Also when the student searches for the entrance exam information, there are chances that they will some colleges. To make this task easy, our application fetches the email ids of the school and the information about various engineering entrance exam from the website URL. Then send e-mails of the fetched information to the fetched e-mail addresses. The application we designed automates the task of extracting emails and information. This application helps the students in getting all the necessary information on time without being even searching for it on web.

REFERENCES

[1] LanTechsoft, https://www.lantechsoft.com/web-email-extractor.html-
[2] S.C.M. de S Sirisuriya, 2015, A Comparative Study on Web Scraping .Proceedings of 8th International Research Conference, KDU.
[3] K. I. Satoto, R. R. Isnanto, R. Kridalukmana, K. T. Martono, "Optimizing MySQL database system on information systems research publications and community service", *2016 3rd International Conference on Information Technology Computer and Electrical Engineering (ICITACEE)*, pp. 1-5, 2016
[4] Jason Buberel, Java regex email, Web Page, https://stackoverflow.com/questions/8204680/java-regex-email
[5] Nikos Maravitsas, Extract HTML Links, Web Page, https://examples.javacodegeeks.com/core-java/util/regex/matcher/extract-html-links-with-java-regular-expression-example