IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Implementation of Web Scraper Bot: Web Harvesting

NandanAS[1], Rahul S Niranjan[2], PB Rahul Choudhary[3], Karthik Srinivas[4], Nilesh Kumar Singh[5], Kavya P Hathwar[6]

Student, Department of CSE, BNM Institute of Technology, Bangalore, India [1]

Student, Department of CSE, BNMIT, Bangalore, India [2, 3, 6]

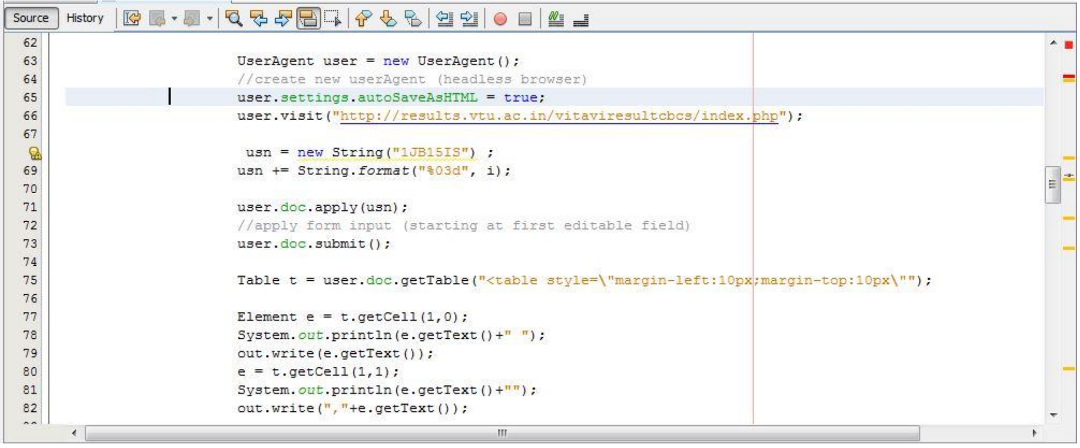
Student, Department of ECE, JSSATE, Bangalore, India[4, 5]

Abstract: Web Harvesting or Web-Scraping also called web data extraction, are various methods of collecting information from across the internet. It is essentially a form of data mining. Programs are written to mine the data and to convert it into a meaningful and useful structure. In this paper we demonstrate a code/program written to harvest web-data from a particular web-site and to display the same in different file formats. The university results of a class of Students arescraped from the web and are being stored and calculated. This code can be reused several number of times and may also be altered to suit the desired/intended application. Creating a customized score sheet of all students in the college or university is a tedious task. In this paper a web scraper bot is employed to do the same within minutes.

Keywords: Web Harvesting or Web-Scraping, essentially a form of data mining, reused several number of times, Creating a customized score sheet, web scraper bot, regular expression (regex).

I. INTRODUCTION

Web mining is employed to extract data from the web [1]. It is the application of data mining techniques to extract knowledge from web data, including web documents, usage logs of website and also web content which can be further conditioned to provide monetary benefits. Figure1 Shows the Taxonomy of web mining. Web content mining is the process of extracting useful information from the contents of web documents.Content data is the collection of facts a web page is designed to contain [2]. It may consist of text, image, audio, video or structured records such as list and tables. Web structure mining is the process of discovering structure information from the web. It is further divided into hyperlinks and document structure. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behaviours at a website. Web usage mining itself can be classified further depending on the kind of usage data considered.



```
UserAgent user = new UserAgent();
//create new userAgent (headless browser)
user.settings.autoSaveAsHTML = true;
user.visit("http://results.vtu.ac.in/vitaviresultcbcs/index.php");

 usn = new String("1JB15IS") ;
usn += String.format("%03d", i);

user.doc.apply(usn);
//apply form input (starting at first editable field)
user.doc.submit();

Table t = user.doc.getTable("<table style=\"margin-left:10px;margin-top:10px\"");

Element e = t.getCell(1,0);
System.out.println(e.getText()+" ");
out.write(e.getText());
e = t.getCell(1,1);
System.out.println(e.getText()+"");
out.write(","+e.getText());
```

Figure 2 shows the URL of a university score display website

II. PROBLEM STATEMENT AND POSSIBLE SOLUTION

A. Problem

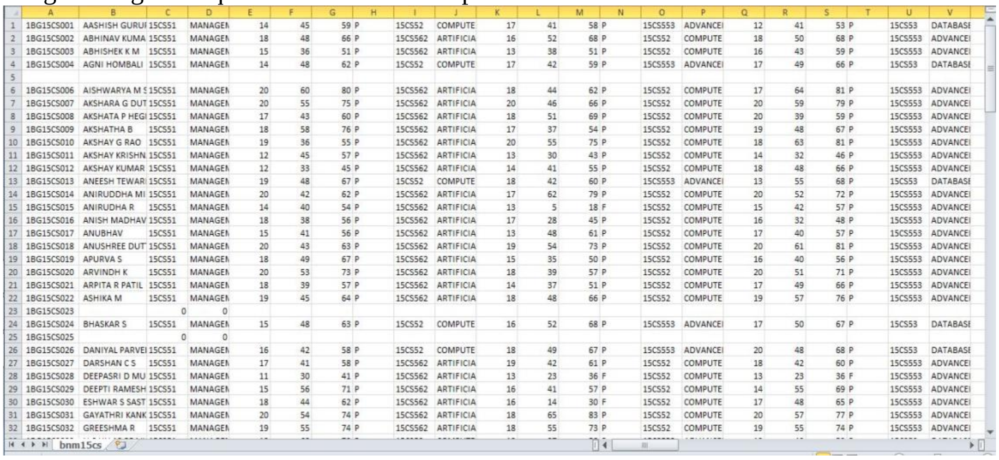
Creating a customized score sheet of all the students in a university is a tedious task. A DEO has to fill more than 50,000 rows on an average. CGPA calculation, credit score realization and categorization is another overhead that cannot be ignored.

DOI 10.17148/IJARCCE.2018.7362

B. Solution

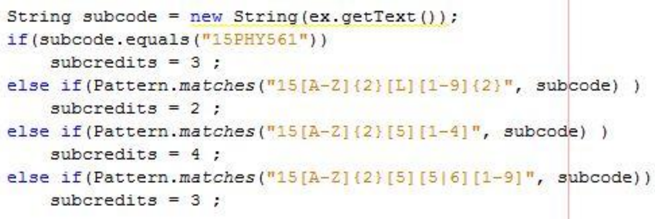
- A Web Scraper Bot

1. An intelligent way of handling things and is very much required in this space. An Algorithm is the need of the day.
2. People can expect a lot of accuracy in the results obtained.
3. The work of gathering the required data will be completed within minutes/seconds.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	1BG15CS001	AASHISH GURUI	15CS51	MANAGEN	14	45	59	P	15CS52	COMPUTE	17	41	58	P	15CS553	ADVANCEI	12	41	53	P	15CS53	DATABASE
2	1BG15CS002	ABHINAV KUMA	15CS51	MANAGEN	18	48	66	P	15CS562	ARTIFICIA	16	52	68	P	15CS52	COMPUTE	18	50	68	P	15CS553	ADVANCEI
3	1BG15CS003	ABHISHEK K M	15CS51	MANAGEN	15	36	51	P	15CS562	ARTIFICIA	13	38	51	P	15CS52	COMPUTE	16	43	59	P	15CS553	ADVANCEI
4	1BG15CS004	AGNI HOMBALI	15CS51	MANAGEN	14	48	62	P	15CS52	COMPUTE	17	42	59	P	15CS553	ADVANCEI	17	49	66	P	15CS53	DATABASE
5																						
6	1BG15CS006	AISHWARYA M S	15CS51	MANAGEN	20	60	80	P	15CS562	ARTIFICIA	18	44	62	P	15CS52	COMPUTE	17	64	81	P	15CS553	ADVANCEI
7	1BG15CS007	AKSHARA G DUT	15CS51	MANAGEN	20	55	75	P	15CS562	ARTIFICIA	20	46	66	P	15CS52	COMPUTE	20	59	79	P	15CS553	ADVANCEI
8	1BG15CS008	AKSHATA P HEGI	15CS51	MANAGEN	17	43	60	P	15CS562	ARTIFICIA	18	51	69	P	15CS52	COMPUTE	20	39	59	P	15CS553	ADVANCEI
9	1BG15CS009	AKSHATHA B	15CS51	MANAGEN	18	58	76	P	15CS562	ARTIFICIA	17	37	54	P	15CS52	COMPUTE	19	48	67	P	15CS553	ADVANCEI
10	1BG15CS010	AKSHAY G RAO	15CS51	MANAGEN	19	36	55	P	15CS562	ARTIFICIA	20	55	75	P	15CS52	COMPUTE	18	63	81	P	15CS553	ADVANCEI
11	1BG15CS011	AKSHAY KRISHN	15CS51	MANAGEN	12	45	57	P	15CS562	ARTIFICIA	13	30	43	P	15CS52	COMPUTE	14	32	46	P	15CS553	ADVANCEI
12	1BG15CS012	AKSHAY KUMAR	15CS51	MANAGEN	12	33	45	P	15CS562	ARTIFICIA	14	41	55	P	15CS52	COMPUTE	18	48	66	P	15CS553	ADVANCEI
13	1BG15CS013	ANEESH TEWARI	15CS51	MANAGEN	19	48	67	P	15CS52	COMPUTE	18	42	60	P	15CS553	ADVANCEI	13	55	68	P	15CS53	DATABASE
14	1BG15CS014	ANIRUDDHA MI	15CS51	MANAGEN	20	42	62	P	15CS562	ARTIFICIA	17	62	79	P	15CS52	COMPUTE	20	52	72	P	15CS553	ADVANCEI
15	1BG15CS015	ANIRUDHA R	15CS51	MANAGEN	14	40	54	P	15CS562	ARTIFICIA	13	5	18	F	15CS52	COMPUTE	15	42	57	P	15CS553	ADVANCEI
16	1BG15CS016	ANISH MADHAV	15CS51	MANAGEN	18	38	56	P	15CS562	ARTIFICIA	17	28	45	P	15CS52	COMPUTE	16	32	48	P	15CS553	ADVANCEI
17	1BG15CS017	ANUBHAV	15CS51	MANAGEN	15	41	56	P	15CS562	ARTIFICIA	13	48	61	P	15CS52	COMPUTE	17	40	57	P	15CS553	ADVANCEI
18	1BG15CS018	ANUSHREE DUT	15CS51	MANAGEN	20	43	63	P	15CS562	ARTIFICIA	19	54	73	P	15CS52	COMPUTE	20	61	81	P	15CS553	ADVANCEI
19	1BG15CS019	APURVA S	15CS51	MANAGEN	18	49	67	P	15CS562	ARTIFICIA	15	35	50	P	15CS52	COMPUTE	16	40	56	P	15CS553	ADVANCEI
20	1BG15CS020	ARVINDH K	15CS51	MANAGEN	20	53	73	P	15CS562	ARTIFICIA	18	39	57	P	15CS52	COMPUTE	20	51	71	P	15CS553	ADVANCEI
21	1BG15CS021	ARPITA R PATIL	15CS51	MANAGEN	18	39	57	P	15CS562	ARTIFICIA	14	37	51	P	15CS52	COMPUTE	17	49	66	P	15CS553	ADVANCEI
22	1BG15CS022	ASHIKA M	15CS51	MANAGEN	19	45	64	P	15CS562	ARTIFICIA	18	48	66	P	15CS52	COMPUTE	19	57	76	P	15CS553	ADVANCEI
23	1BG15CS023		0	0																		
24	1BG15CS024	BHASKAR S	15CS51	MANAGEN	15	48	63	P	15CS52	COMPUTE	16	52	68	P	15CS553	ADVANCEI	17	50	67	P	15CS53	DATABASE
25	1BG15CS025		0	0																		
26	1BG15CS026	DANIYAL PARVEI	15CS51	MANAGEN	16	42	58	P	15CS52	COMPUTE	18	49	67	P	15CS553	ADVANCEI	20	48	68	P	15CS53	DATABASE
27	1BG15CS027	DARSHAN C S	15CS51	MANAGEN	17	41	58	P	15CS562	ARTIFICIA	19	42	61	P	15CS52	COMPUTE	18	42	60	P	15CS553	ADVANCEI
28	1BG15CS028	DEEPASRI D MU	15CS51	MANAGEN	11	30	41	P	15CS562	ARTIFICIA	13	23	36	F	15CS52	COMPUTE	13	23	36	F	15CS553	ADVANCEI
29	1BG15CS029	DEEPTI RAMESH	15CS51	MANAGEN	15	56	71	P	15CS562	ARTIFICIA	16	41	57	P	15CS52	COMPUTE	14	55	69	P	15CS553	ADVANCEI
30	1BG15CS030	ESHWAR S SAST	15CS51	MANAGEN	18	44	62	P	15CS562	ARTIFICIA	16	14	30	F	15CS52	COMPUTE	17	48	65	P	15CS553	ADVANCEI
31	1BG15CS031	GAYATHRI KANK	15CS51	MANAGEN	20	54	74	P	15CS562	ARTIFICIA	18	65	83	P	15CS52	COMPUTE	20	57	77	P	15CS553	ADVANCEI
32	1BG15CS032	GREESHMA R	15CS51	MANAGEN	19	55	74	P	15CS562	ARTIFICIA	18	55	73	P	15CS52	COMPUTE	19	55	74	P	15CS553	ADVANCEI

Figure 3 shows the USNs being scraped from the web and are displayed in CSV format.



```
String subcode = new String(ex.getText());
if(subcode.equals("15PHY561"))
    subcredits = 3 ;
else if(Pattern.matches("15[A-Z]{2}[L][1-9]{2}", subcode) )
    subcredits = 2 ;
else if(Pattern.matches("15[A-Z]{2}[5][1-4]", subcode) )
    subcredits = 4 ;
else if(Pattern.matches("15[A-Z]{2}[5][5|6][1-9]", subcode))
    subcredits = 3 ;
```

Figure 4 shows the if-else statement for subject credits

III. METHODOLOGY

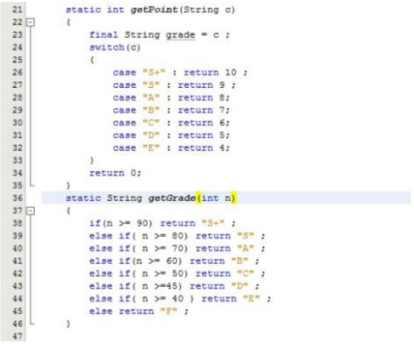
A java code is written to systematically harvest the data from the web. The URL of interest is selected and is pointed to by the admin to the web scrapper bot.Figure 2 shows the URL of a university score display website. The University Seat Number (USN) is incremented from its initial value to a stoppage value. USNs are scraped from the web and are displayed in CSV format as shown in figure 3.

- SUBJECT CREDITS

Each Subject has its own credit Points. Core subjects carry 4 points. Optional subjects carry 3 points. Lab/Practical sessions carry 2 points. Due to the variation in credit points the CGPA calculation is bit tedious when done manually. The web scraper bot does the job for you. Figure 4 shows the code for the above.

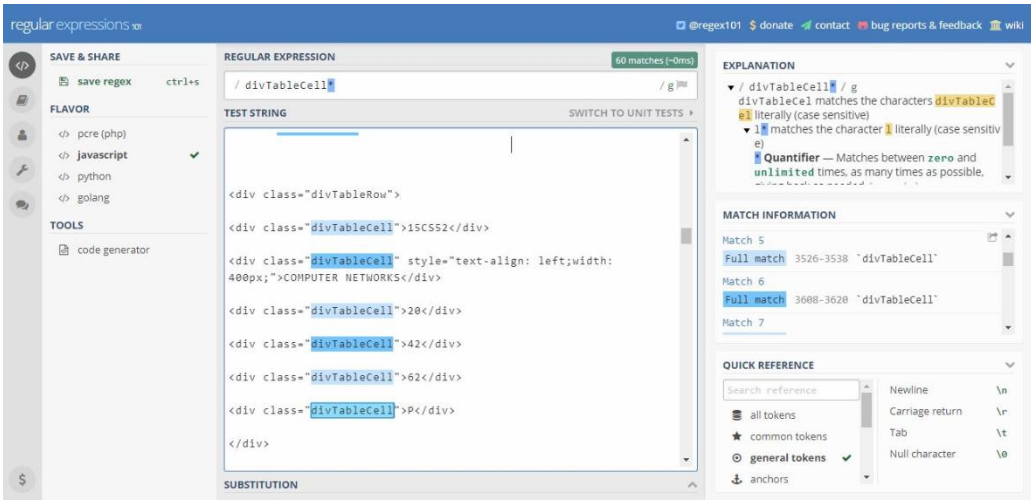
- Case Statements for awarding GRADE

Figure 5 shows the code, where case statement is used to calculate the GRADE of a result of a candidate. Figure 6 shows pattern matching using regex.Everything is parsed using regular expression (regex) [3]. Regex can be used to provide an effective and compact solution to a problem.



```
static int getPoint(String c)
{
    final String grade = c ;
    switch(c)
    {
        case "S+" : return 10 ;
        case "S" : return 9 ;
        case "A" : return 8;
        case "B" : return 7;
        case "C" : return 6;
        case "D" : return 5;
        case "E" : return 4;
    }
    return 0;
}
static String getGrade(int n)
{
    if(n >= 90) return "S+" ;
    else if( n >= 80) return "S" ;
    else if( n >= 70) return "A" ;
    else if(n >= 60) return "B" ;
    else if( n >= 50) return "C" ;
    else if( n >=45) return "D" ;
    else if( n >= 40 ) return "E" ;
    else return "F" ;
}
```

Figure 5 shows the code, where case statement is used to calculate the GRADE of a result of a candidate.



```
/ divTableCell* / g
```

```
<div class="divTableRow">

<div class="divTableCell">15CS52</div>

<div class="divTableCell" style="text-align: left;width:
400px;">COMPUTER NETWORKS</div>

<div class="divTableCell">20</div>

<div class="divTableCell">42</div>

<div class="divTableCell">62</div>

<div class="divTableCell">P</div>

</div>
```

/ divTableCell* / g
divTableCel matches the characters divTableCel literally (case sensitive)
1* matches the character 1 literally (case sensitive)
* Quantifier — Matches between zero and unlimited times, as many times as possible,

```
Match 5
Full match 3526-3538 `divTableCell`
Match 6
Full match 3608-3620 `divTableCell`
Match 7
```

Figure 6 shows pattern matching using regex.

IV. RESULTS

The flow diagram of a ‘VTU Result Scraper Bot’ which calculates the CGPA and awards grades to a candidate belonging to certain stream is as shown in figure 7.The code is executed and the results are displayed in the output window. Figure 8 shows the results with CGPA and total credit points [4].

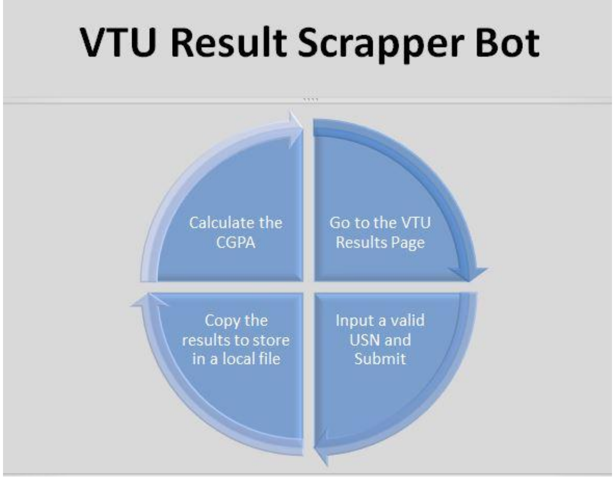
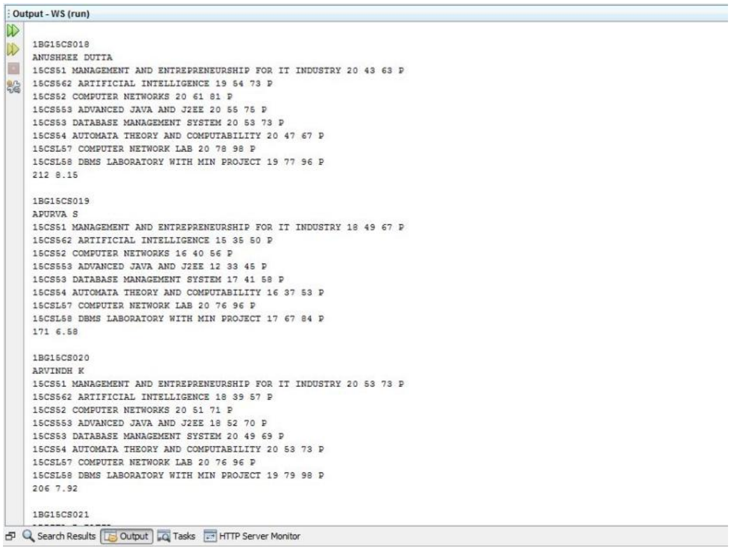

VTU Result Scrapper Bot
Go to the VTU Results Page
Input a valid USN and Submit
Copy the results to store in a local file
Calculate the CGPA

Figure 7 shows ‘VTU Result Scraper Bot’



```
1BG15CS018
ANUSHREE DUTTA
15CS51 MANAGEMENT AND ENTREPRENEURSHIP FOR IT INDUSTRY 20 43 63 P
15CS562 ARTIFICIAL INTELLIGENCE 19 54 73 P
15CS52 COMPUTER NETWORKS 20 61 81 P
15CS553 ADVANCED JAVA AND J2EE 20 55 75 P
15CS53 DATABASE MANAGEMENT SYSTEM 20 53 73 P
15CS54 AUTOMATA THEORY AND COMPUTABILITY 20 47 67 P
15CSL57 COMPUTER NETWORK LAB 20 78 98 P
15CSL58 DBMS LABORATORY WITH MIN PROJECT 19 77 96 P
212 8.15

1BG15CS019
APURVA S
15CS51 MANAGEMENT AND ENTREPRENEURSHIP FOR IT INDUSTRY 18 49 67 P
15CS562 ARTIFICIAL INTELLIGENCE 15 35 50 P
15CS52 COMPUTER NETWORKS 16 40 56 P
15CS553 ADVANCED JAVA AND J2EE 12 33 45 P
15CS53 DATABASE MANAGEMENT SYSTEM 17 41 58 P
15CS54 AUTOMATA THEORY AND COMPUTABILITY 16 37 53 P
15CSL57 COMPUTER NETWORK LAB 20 76 96 P
15CSL58 DBMS LABORATORY WITH MIN PROJECT 17 67 84 P
171 6.58

1BG15CS020
ARVINDH K
15CS51 MANAGEMENT AND ENTREPRENEURSHIP FOR IT INDUSTRY 20 53 73 P
15CS562 ARTIFICIAL INTELLIGENCE 18 39 57 P
15CS52 COMPUTER NETWORKS 20 51 71 P
15CS553 ADVANCED JAVA AND J2EE 18 52 70 P
15CS53 DATABASE MANAGEMENT SYSTEM 20 49 69 P
15CS54 AUTOMATA THEORY AND COMPUTABILITY 20 53 73 P
15CSL57 COMPUTER NETWORK LAB 20 76 96 P
15CSL58 DBMS LABORATORY WITH MIN PROJECT 19 79 98 P
206 7.92

1BG15CS021
```

Figure 8shows the results with CGPA and total credit points.

DOI 10.17148/IJARCCE.2018.7362

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 7, Issue 3, March 2018

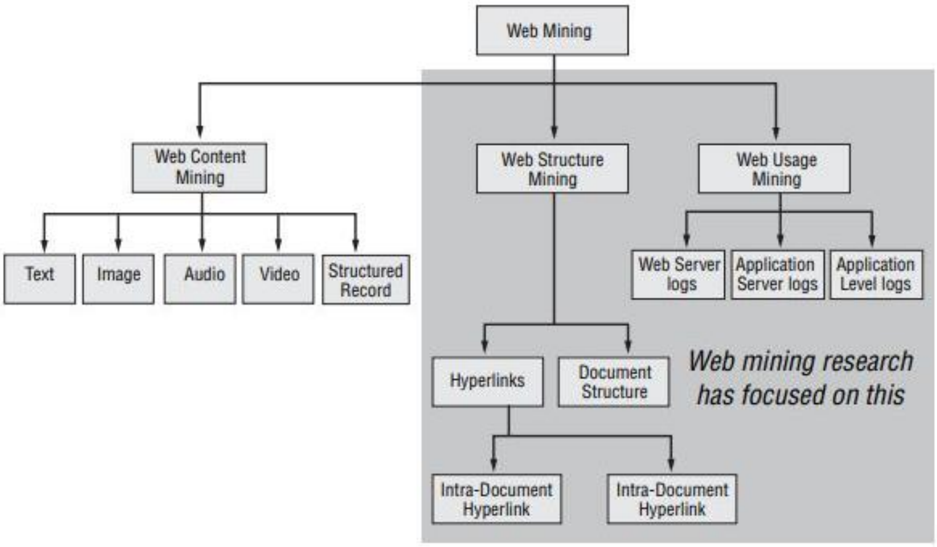


Figure 1 shows the taxonomy of web mining

V. CONCLUSIONS

In this paper we have developed a 'VTU Web Scraper Bot'. An intelligent way of handling things and is very much required in this space. An Algorithm is written and executed. People can expect a lot of accuracy in the results obtained. The work of gathering the required data will be completed within minutes/seconds. The 'VTU Result Scraper Bot' calculates the CGPA and awards grades to a candidate belonging to certain stream automatically. Creating a customized score sheet of all students in the college or university is a tedious task.This code can be reused several number of times and may also be altered to suit the desired/intended application.

REFERENCES

[1] Data Mining, Internet Marketing and Web Mining - ijarcce
[2] https://www.ijarcce.com/upload/2017/march-17/IJARCCE%20117.pdf
[3] Web Mining - Data Analysis and Management Research Group dmr.cs.umn.edu/papers/p2004_4.pdf
[4] RegExr: Learn, Build, & Test RegEx https://regexr.com/
[5] (cbcs) regulations governing the - VTU-www.vtu.ac.in/pdf/regulation/becbcs.pdf

OUR GUIDE

VISHESH S born on 13th June 1992, hails from Bangalore (Karnataka) and has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka in 2015. He also worked as an intern under Dr. Shivananju BN, former Research Scholar, Department of Instrumentation, IISc, Bangalore. His research interests include Embedded Systems, Wireless Communication, BAN and Medical Electronics. He is also the Founder and Managing Director of the corporate company Konigtronics Private Limited. He has guided over a hundred students/interns/professionals in their research work and projects. He is also the co-author of many International Research Papers. He is currently pursuing his MBA in e-Business and PG Diploma in International Business. Presently Konigtronics Private Limited has extended its services in the field of Software Engineering and Webpage Designing. Konigtronics also conducts technical and non-technical workshops on various topics.

DOI 10.17148/IJARCCE.2018.7362