Volume 3 Issue 2: 2015
ISSN (O): 2348-4098
ISSN (P): 2395-4752

A WEB SCRAPER FOR EXTRACTING ALUMNI INFORMATION FROM SOCIAL NETWORKING WEBSITES

[1]MUKUND WAGH, [2]SUPARNIKA MOHATA, [3] SHANTANU KAMDI, [4]RASHMI MOHOD, [5]DIKSHA PACHE, [6]HIMANSHU KAUTKAR

[1] Student, Department of Computer Technology, YCCE, Nagpur, India, mukundwagh@gmail.com
[2]Student, Department of Computer Technology, YCCE, Nagpur, India, suparnika.mohata@gmail.com
[3]Student, Department of Computer Technology, YCCE, Nagpur, India, kamdishantanu@gmail.com
[4]Student, Department of Computer Technology, YCCE, Nagpur, India, rashmimohod@gmail.com
[5]Student, Department of Computer Technology, YCCE, Nagpur, India, dikshapache18@gmail.com
[6]Student, Department of Computer Technology, YCCE, Nagpur, India, kautkarhimanshu37@gmail.com

ABSTRACT

Scrapers are used for extracting information from the repository of web pages which can be stored in well-defined structure to be used for various purposes. The project is about design of an efficient web scrapper on keywords college name, batch, year, etc. for extracting YCCE alumni information by scraping social networking websites Facebook and LinkedIn and design of a database of extracted information. The generated reports using user interface can be used for various purposes.

Index Terms: *Web Scraper, Social Networking Websites, Extraction, Keyword, Parsing, Profiles.*

1. INTRODUCTION

The computer world is giving birth to extensive amount of data every day, to search the required data, concept of scraper wasemerged. In general terms, web scraping isthe way of fetching the data from WebPages using html parsing technique. Either the regular expression in terms of tags is used or the direct tag is being searched. In this paper, we are unveiling the algorithmic and application based representation of the scraper for extracting the alumni information through the social networking websites and creating a database to store the scraped information which can be accessed by user interface for batch wise or name wise retrieval of the records. We are using graph API of Facebook for scraping profiles of alumni from Facebook while HTML parsing is used for scraping LinkedIn profiles

2. RELATED WORK

From the study of related works in data mining techniques,it is clear that scraping of websites can be implemented in variety of ways.

Ramakrishnan R. [1], implemented Advanced Multimedia Answer generation by scraping information through web . It uses novel multimedia question answering (MMQA). This technique can enrich community-contributed textual answers in cQA with appropriate media data. It consists of three components: Answer medium selection, Query generation for multimedia search, Multimedia data selection and presentation. The algorithms used to create the system are: Stemming Algorithm & stop word removal, Naive Bayes, Bigram text classification, POS Histogram.
Vidya V.L. [2] developed various information extraction techniques like SoftMealy, OLERA, IEPAD, RoadRunner, EXALG, NET, FivaTech. It also gives the comparision between the extraction technologies on the basis of supervision type, learning algorithm and the number of pages. Among the above mentioned web data extraction techniques, some techniques extract flat records and some other techniques are trying to extracts nested records.

V. B. Kadam [3], analyzed HTML aware web scraping techniques. The techniques discussed by him includes the RoadRunner, W4F (Wysiwyg Web Wrapper Factory), XWRAP, IEPAD, FiVaTech, DELA (Data Extraction and Label Assignment for Web Databases), DEPTA (Data Extraction based on Partial Tree Alignment), ViPER (Visual perception based Extraction of Records), ViNTs (Visual information and Tag structure basedwrapper generator), CTVS (Data Extraction and Alignment using Combining Tag and Data Value Similarity), Mining Data Records in Web Pages, MSE (Multiple Section Extraction).

Vasani Krunal A [4], introduced a solution on the tree edit distance problem which is related to semantic analysis and improving the performance of tree edit distance problem. It also focuses on higher bound time complexity of this algorithm.

Shridevi Swami, [5], used an approach for the atomic data extraction on Web Scraping framework which uses Tag and Value similarity together for automatically extracting data from query result pages. Web data extraction system, automatically and repeatedly extracts data from dynamic web pages and can deliver the extracted data to a database or some other application.

Govind M. Upadhyay [6], focused on the value of web content mining. The paper gives an insight into its techniques, processes and its applications in the current cut-throat business environment as well in research and extracting contents for educational purposes. It further explains how using web content mining plays an integral role by getting rich set of contents and uses those

Volume 3 Issue 2: 2015
ISSN (O): 2348-4098
ISSN (P): 2395-4752

contents in the decision making in the corporate environment, education and research.

Anthony J. Dreyer [7], analyzed the legal framework surrounding scraping, addressing both the grounds for potential claims against scrapers. Common theories of liability arising from scraping are copyright infringement, trespass to chattels, breach of contract, and violation of the Computer Fraud and Abuse Act (CFAA).This article discusses the leading cases applying these legal theories to website scraping, and concludes that the most effective way to create potential claims against scrapers is through carefully drafted prohibitions in a website's terms of use.

A. Shingate [8], explained development of a Website where users can get an optimized result for the different opinions on different products or events or services on different social networking websites. The project designed mainly deals with checking different opinions so that we can get a quick idea of different users based on their opinions.

P. P. Singh Bedi [9], introduced the technique of designing of the web scraper using prolog server pages. The authors attempt to establish a technique to scrap HTML pages and utilize it as per the requirements of the data and its data type. It also provides the information about how GUI can help to access the information extracted from web pages

Richard Baron Penman [10], concluded about a tool called Site Scraper that aims to address problems occurring while extracting content of the web pages. Use of site scraper allows user to focus on content rather on the structure of the web page.

3. METHODOLOGY

3.1 HTML Parsing

Many websites are consisting of large sets of pages generated dynamically from database. Data of the same perspectives are typically encoded into similar pages by a common script or template. In data mining, a program that detects such kind of templates in a particular information source, extracts its information and translates it into a relational form, is called a wrapper or scraper.

3.2 Graph API

The Graph API is the primary way to get data from Facebook's platform. The Graph API was launched in March 2010 with the intention of replacing the older REST API. There are three methods of using the Graph API**:** requesting data, posting data, and deleting data. Some data can be requested without authentication of the user while most data is dependent on authentication. If the user provides permission to access the profile information, only then the developers can seek the data.

3.3 Proposed Algorithm

Algorithm for scraping Facebook profiles:

1. Login to Facebook with user credentials.
2. Fire the search query and load all the elements of the page by infinite scrolling.
3. Scrap URL elements of each person from HTML document.
4. Store all the scraped elements in excel sheet and replace URL with Graph API.
5. Store Profession of each person from HTML document.
6. Store the JSON array of each searched person from graph.Facebook.com
7. Convert all records stored in excel sheet.
8. Load all the scraped data in excel to access database.
9. Access the database using UI.

Algorithm for scraping LinkedIn profiles:

1. Login to LinkedIn with user credentials.
2. Fire the search query in given text box and load all elements with next page loading click event.
3. Scrap profiles of alumni through HTML class of each element.
4. Put the inner text of the tags into variables.
5. The values of variable can be stored in access file using SQL queries.

Graph API of Facebook gives access to developers open the profiles of Facebook user in HTML document and retrieves the information of the public data of user. HTML parsing allows us to store the text fields present in the inner tags as the information of each user.

3.4 Flowchart

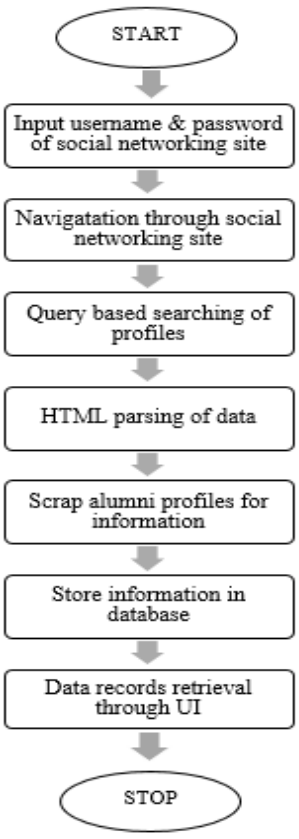


Figure 1: Work flow chart for scraper

The work flow of both modules is similar in certain context. The user inserts login credential which will be filled automatically in the navigated web browser window of the social networking site. The searching of the profile is based on particular key constraint, for example, here

Volume 3 Issue 2: 2015
ISSN (O): 2348-4098
ISSN (P): 2395-4752

the keyword is "people who studied at Yeshwantrao Chavan College of Engineering".

The keyword is transferred to search field of website and start giving the profile of alumni. The profiles are then fetched as records by procedure mentioned in the algorithm given above for respective websites. The information is stored in access database and can be retrieved directly through the user interface designed for it.

4. DERIVED RESULTS

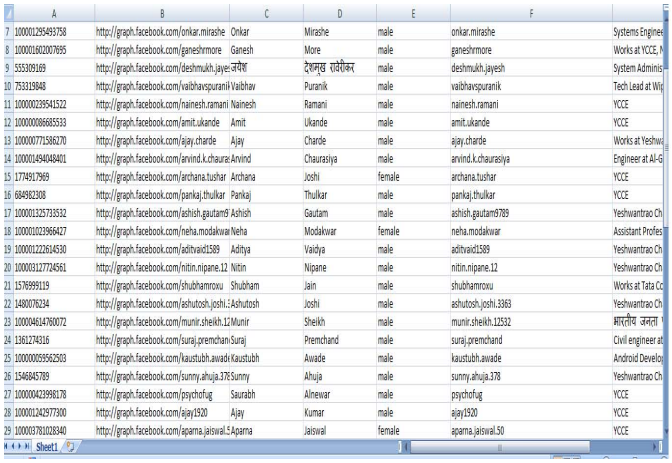
The scraper selection form shown in Figure 2 allows to choose scraper for LinkedIn/Facebook. The click event of the button of LinkedIn and Facebook forwards the user control to the chosen social networking website.

Figure 2: Scraping UI for LinkedIn / Facebook

The keywords for searching alumni records are received by the search box of the Facebook page as shown in Figure. 3. The click event of search button on this webpage is invoked. This makes the required profiles to get displayed on the web page. The webpage shown in Figure 3 shows the list of matched profiles.

Figure 3: Dynamic keyword based searching of alumni profiles from Facebook

The excel sheet shown in Figure 4 contains all the information of the alumni which has been fetched from Facebook. The fetched information includes the fields such as First name, Last Name, Gender, Work, Facebook ID and URL. The information fetched from Facebook is within the bound of privacy. The information which is kept private by the user, would not be fetched during scraping.



	A	B	C	D	E	F	
7	100001295493758	http://graph.facebook.com/onkar.mirashe	Onkar	Mirashe	male	onkar.mirashe	Systems Enginee
8	100001802007695	http://graph.facebook.com/ganeshrmore	Ganesh	More	male	ganeshrmore	Works at YCCE, M
9	555369169	http://graph.facebook.com/deshmukh.jayes	जयेश	देशमुख	male	deshmukh.jayesh	System Adminis
10	753319848	http://graph.facebook.com/vaibhavspuranik	Vaibhav	Puranik	male	vaibhavspuranik	Tech Lead at Wi
11	100000239541522	http://graph.facebook.com/nainesh.ramani	Nainesh	Ramani	male	nainesh.ramani	YCCE
12	100000086685533	http://graph.facebook.com/amit.ukande	Amit	Ukande	male	amit.ukande	YCCE
13	100000771586270	http://graph.facebook.com/ajay.charde	Ajay	Charde	male	ajay.charde	Works at Yeshw
14	100001494048401	http://graph.facebook.com/arvind.k.chaura	Arvind	Chaurasiya	male	arvind.k.chaurasiya	Engineer at Al-Gi
15	1774917969	http://graph.facebook.com/archana.tushar	Archana	Joshi	female	archana.tushar	YCCE
16	684982308	http://graph.facebook.com/pankaj.thulkar	Pankaj	Thulkar	male	pankaj.thulkar	YCCE
17	100001325733532	http://graph.facebook.com/ashish.gautam97	Ashish	Gautam	male	ashish.gautam9789	Yeshwantrao Ch
18	100001023966427	http://graph.facebook.com/neha.modakwar	Neha	Modakwar	female	neha.modakwar	Assistant Profes
19	100001222614530	http://graph.facebook.com/adityaidc1589	Aditya	Vaidya	male	adityaidc1589	Yeshwantrao Ch
20	100003127724561	http://graph.facebook.com/nitin.nipane.12	Nitin	Nipane	male	nitin.nipane.12	Yeshwantrao Ch
21	1576999119	http://graph.facebook.com/shubhamroxu	Shubham	Jain	male	shubhamroxu	Works at Tata Co
22	1480076234	http://graph.facebook.com/ashutosh.joshi.3	Ashutosh	Joshi	male	ashutosh.joshi.3363	Yeshwantrao Ch
23	100004614760072	http://graph.facebook.com/munir.sheikh.12	Munir	Sheikh	male	munir.sheikh.12532	भारतीय जनता
24	1361274316	http://graph.facebook.com/suraj.premchan	Suraj	Premchand	male	suraj.premchand	Civil engineer at
25	100000059562503	http://graph.facebook.com/kaustubh.awadk	Kaustubh	Awade	male	kaustubh.awade	Android Develo
26	1546845789	http://graph.facebook.com/sunny.ahuja.378	Sunny	Ahuja	male	sunny.ahuja.378	Yeshwantrao Ch
27	100000423998178	http://graph.facebook.com/psychofug	Saurabh	Alnewar	male	psychofug	YCCE
28	100001242977300	http://graph.facebook.com/ajay1920	Ajay	Kumar	male	ajay1920	YCCE
29	100003781028340	http://graph.facebook.com/aparna.jaiswal.5	Aparna	Jaiswal	female	aparna.jaiswal.50	YCCE

Figure 4: Storing of Alumni complete information fetched from Facebook

After successful login in LinkedIn, we need to enter the keyword for the searching and scraping profiles of alumni. The keywords are reflected in search box of LinkedIn webpage after we click on the 'Search' button on web browser. The matched results are displayed in search result container.

Figure. 5 shows the search container elements of the LinkedIn search result. The "Scrap" button click makes the required HTML tag inner text to get stored into the access file. This access file checks for the redundant data and store only those records which are unique. The LinkedIn scraper scraps the "Full Name", "Location", "Past Profile" and the "Current Profile" of the alumni.

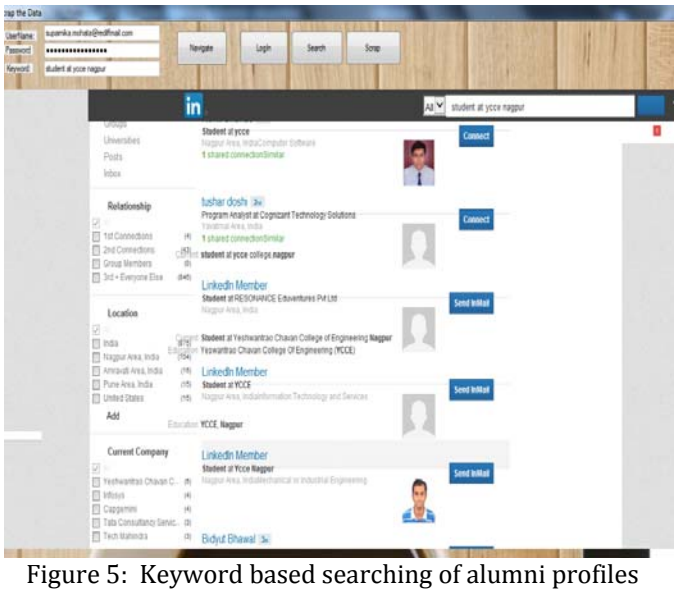
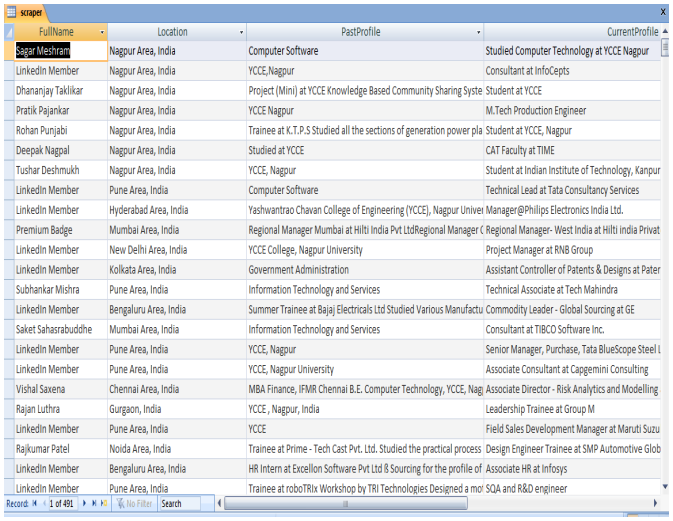


Figure 5: Keyword based searching of alumni profiles from LinkedIn

Figure 6 shows the information of about 25 alumni which is fetched from LinkedIn Profiles. The access file records shown above involve "Full Name", "Location", "Past Profile" and "Current Profile" of alumni.

Volume 3 Issue 2: 2015
ISSN (O): 2348-4098
ISSN (P): 2395-4752



scraper

FullName	Location	PastProfile	CurrentProfile
Sagar Meshram	Nagpur Area, India	Computer Software	Studied Computer Technology at YCCE Nagpur
LinkedIn Member	Nagpur Area, India	YCCE,Nagpur	Consultant at InfoCepts
Dhananjay Taklikar	Nagpur Area, India	Project (Mini) at YCCE Knowledge Based Community Sharing Syste	Student at YCCE
Pratik Pajankar	Nagpur Area, India	YCCE Nagpur	M.Tech Production Engineer
Rohan Punjabi	Nagpur Area, India	Trainee at K.T.P.S Studied all the sections of generation power pla	Student at YCCE, Nagpur
Deepak Nagpal	Nagpur Area, India	Studied at YCCE	CAT Faculty at TIME
Tushar Deshmukh	Nagpur Area, India	YCCE, Nagpur	Student at Indian Institute of Technology, Kanpur
LinkedIn Member	Pune Area, India	Computer Software	Technical Lead at Tata Consultancy Services
LinkedIn Member	Hyderabad Area, India	Yashwantrao Chavan College of Engineering (YCCE), Nagpur Univer	Manager@Philips Electronics India Ltd.
Premium Badge	Mumbai Area, India	Regional Manager Mumbai at Hilti India Pvt LtdRegional Manager (	Regional Manager- West India at Hilti india Privat
LinkedIn Member	New Delhi Area, India	YCCE College, Nagpur University	Project Manager at RNB Group
LinkedIn Member	Kolkata Area, India	Government Administration	Assistant Controller of Patents & Designs at Pater
Subhankar Mishra	Pune Area, India	Information Technology and Services	Technical Associate at Tech Mahindra
LinkedIn Member	Bengaluru Area, India	Summer Trainee at Bajaj Electricals Ltd Studied Various Manufactu	Commodity Leader - Global Sourcing at GE
Saket Sahasrabuddhe	Mumbai Area, India	Information Technology and Services	Consultant at TIBCO Software Inc.
LinkedIn Member	Pune Area, India	YCCE, Nagpur	Senior Manager, Purchase, Tata BlueScope Steel L
LinkedIn Member	Pune Area, India	YCCE, Nagpur University	Associate Consultant at Capgemini Consulting
Vishal Saxena	Chennai Area, India	MBA Finance, IFMR Chennai B.E. Computer Technology, YCCE, Nag	Associate Director - Risk Analytics and Modelling
Rajan Luthra	Gurgaon, India	YCCE , Nagpur, India	Leadership Trainee at Group M
LinkedIn Member	Pune Area, India	YCCE	Field Sales Development Manager at Maruti Suzu
Rajkumar Patel	Noida Area, India	Trainee at Prime - Tech Cast Pvt. Ltd. Studied the practical process	Design Engineer Trainee at SMP Automotive Glob
LinkedIn Member	Bengaluru Area, India	HR Intern at Excellon Software Pvt Ltd 8 Sourcing for the profile of	Associate HR at Infosys
LinkedIn Member	Pune Area, India	Trainee at roboTRIx Workshop by TRI Technologies Designed a mo	SQA and R&D engineer

Record: 1 of 491 | No Filter | Search

Figure 6: Alumni Information fetched from LinkedIn

The merged database of LinkedIn and Facebook are as given below in figure 7. Some fields of Facebook retrieved information and LinkedIn retrieved data are not same. So we have replaced it by dummy keyword i.e. 0.

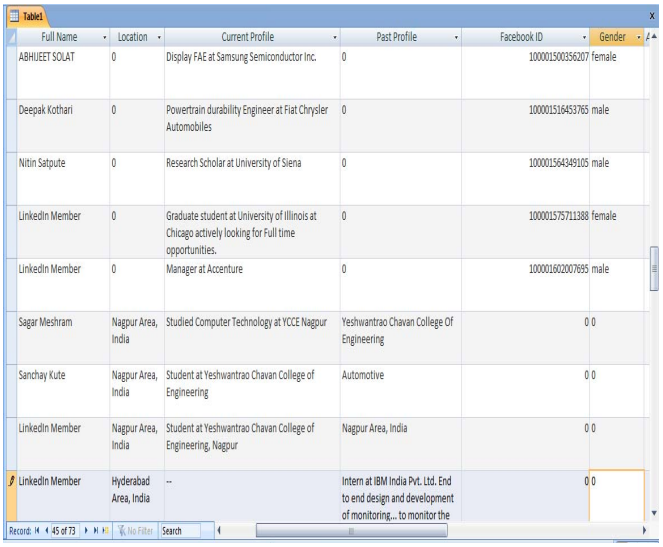


Table1

Full Name	Location	Current Profile	Past Profile	Facebook ID	Gender
ABHIJEET SOLAT	0	Display FAE at Samsung Semiconductor Inc.	0	100001500356207	female
Deepak Kothari	0	Powertrain durability Engineer at Fiat Chrysler Automobiles	0	100001516453765	male
Nitin Satpute	0	Research Scholar at University of Siena	0	100001564349105	male
LinkedIn Member	0	Graduate student at University of Illinois at Chicago actively looking for Full time opportunities.	0	100001575711388	female
LinkedIn Member	0	Manager at Accenture	0	100001602007695	male
Sagar Meshram	Nagpur Area, India	Studied Computer Technology at YCCE Nagpur	Yeshwantrao Chavan College Of Engineering	0	0
Sanchay Kute	Nagpur Area, India	Student at Yeshwantrao Chavan College of Engineering	Automotive	0	0
LinkedIn Member	Nagpur Area, India	Student at Yeshwantrao Chavan College of Engineering, Nagpur	Nagpur Area, India	0	0
LinkedIn Member	Hyderabad Area, India	--	Intern at IBM India Pvt. Ltd. End to end design and development of monitoring... to monitor the	0	0

Record: 45 of 73 | No Filter | Search

Figure 7: Sample record from the merged database of LinkedIn and Facebook data

5. CONCLUSION

This web scraper is effectively able to extract the profiles of alumni and fetch relevant information from social networking site such as Facebook and LinkedIn. The algorithm specifies working of scraper for LinkedIn and Facebook, both HTML parsing and Graph API are used for retrieval of information. The retrieved information through scraper is useful for various purposes. The scraper is not violating any authentication issue and the scraped information is used for creation of alumni database.

REFERENCES

[1]. Ramakrishnan.R, Jayalakshmi.A, Priyadharshani.S, "Advanced Multimedia Answer Generation by Scraping Information through Web", *International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 12, December 2014.*

[2]. Vidya.V.L., "A Survey of Web Data Extraction Techniques",*International Journal of Advance Research inComputer Science and Management Studies, Volume 2, Issue 9, September 2014.*

[3]. Vinayak B. Kadam, Ganesh K. Pakle, "A Survey on HTML Structure Aware and Tree Based Web Data Scraping Technique", *International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014.*

[4]. Vasani Krunal A, "Content evocation using web scraping and semantic illustration" , *IOSR Journal of Computer Engineering, Volume 16, Issue 3, May-Jun. 2014.*

[5]. Shridevi Swami, Pujashree Vidap, "Web Scraping Framework based on Combining Tag and Value Similarity",*International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013.*

[6]. Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013.*

[7]. Anthony J. Dreyer and Jamie Stockton, "Internet Data Scraping, a Primer for Counseling Clients", *New York Law Journal Special Section, July 15, 2013.*

[8]. Abhinav Shingate, Nayan Tayade, Rahul More, ParagZaware, "Opinion Mining: Opinion Extractor from Social Networking Sites [Single Page Result].", *International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4, April 2012.*

[9]. Parminder Pal Singh Bedi, Sumit Kumar, "Web scraping and implementation using prolog server pages in semantic web", *International Journal of Research in Engineering & Applied Sciences, Volume 2, Issue 2,February 2012*.

[10]. Richard Baron Penman, Timothy Baldwin, David Martinez, "Web Scraping Made Simple with Site Scraper", *International Journal of Research in Engineering & Applied Sciences, Volume 4, Issue 3, May 2000.*