IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 4, Ver. I (July – Aug. 2015), PP 88-93
www.iosrjournals.org

Relational Web Wrapper: A Web Data Extraction Approach

Neeraj Raheja[1] , Dr.V.K.Katiyar[2]
Associate Professor1, Professor & Head2
1, 2 Department of Computer Science and Engineering, M.M.University, Mullana (Ambala), Haryana, India

Abstract : *The information over the internet is growing at rapid rate, so web data extraction systems are required to extract the required information. One such technique is web wrapper, which is a supervised learning approach in which a template (program) is developed by the programmer to extract some specific data. This research paper provides a web wrapper known as Relational Web Wrapper which extracts related information of the webpage. Finally the performance evaluation of this web wrapper on the basis of time to extract the data and accuracy provided are shown in results. The results shows this web wrapper provides efficient results.*

Keywords*- Web data extraction, web wrapper ,relational data.*

I INTRODUCTION

The world-wide web (WWW) has become one of the most widely used information resources and hence growing at a very fast rate i.e. millions of websites are available today. Because of this rapid growth and large information available web data extraction systems are necessary to use. Web wrapper is one of such technique which is a supervised learning approach where a programmer develops a program (template) and applies it to a number of webpage. Web wrapper may solve a particular problem i.e. may extract a particular pattern of data. Web wrapper translates the relevant data embedded in web pages into a relational (or other regular) structure and store into some specific format like xml, excel sheets etc.. Wrappers may be developed manually, through wrapper induction or automatic approach. The manual generation of wrappers is difficult, error-prone and specific, hence semi-automatic and automatic wrapper construction systems are preferable [6, 7, 11, 12]. Applications of web wrappers include comparison shopping where information from multiple online shops is extracted to compare prices, and online monitoring of news over multiple websites or bulletin boards, say for items relevant to a particular topic.

In this paper, we describe a new system named Relational web wrapper which is wrapper induction approach i.e. a program is developed. This wrapper extracts the related data of a webpage which may further be used for extending the information of the webpage or user.

II. LITERATURE REVIEW

Kushmerick et al. [13] provides two distinct categories of web wrappers i.e. finite-state and relational learning tools. The extraction rules in finite-state tools are formally equivalent to regular grammars or automata, e.g. WIEN [9], Soft Mealy [10] and STALKER [1], while the extraction rules in relational learning tools are essentially in the form of Prolog-like logic programs, e.g. SRV [8], Crystal [5], Webfoot [14], Rapier and Pinocchio [15]. Muslea et al. [1] provides a web wrapper known as STALKER which takes a webpage as input and then searches some tokens which have to be skipped and then add some new tokens which refines or replaces the current set. R. Baumgartner et al. [2, 3] provides a web wrapper known as Lixto, which offers an interactive interface that hides most technical details from a webpage. Lixto does not support a verification set, and does not provide a ranking of candidate tuple sets to decrease user effort. Senellart et al. [4] provides a web wrapper to extract information from the hidden web sources. It searches the HTML codes of the related WebPages over the internet.

III. SYSTEM MODEL

Relational web wrapper takes a webpage as input and extracts the related data of webpage like Title, Summary and Keywords. The extracted data like keyword (related data of the webpage) are used for further searching so that more results may be provided. The extracted data is stored in excel sheets. It provides the user an easy approach to search the related data of the webpage.

Working to relational web wrapper

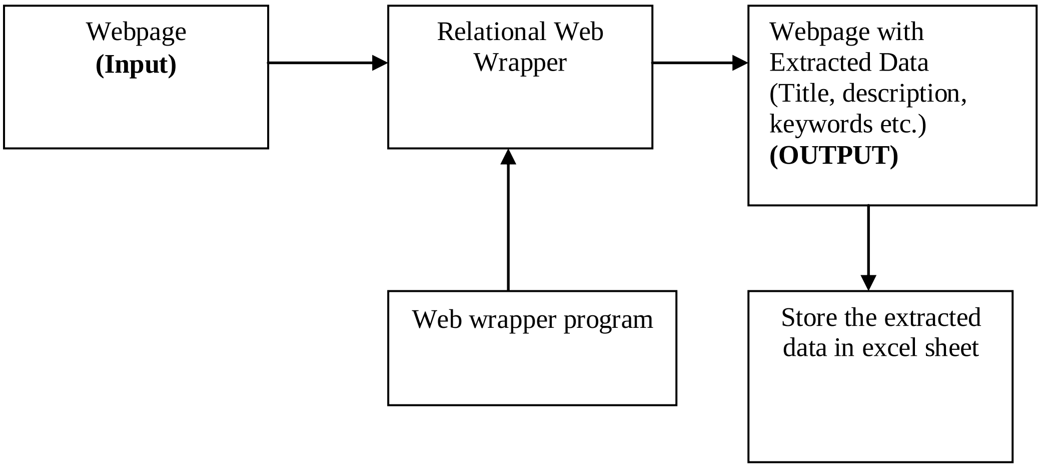

Webpage
(Input)
Relational Web Wrapper
Webpage with Extracted Data (Title, description, keywords etc.) (OUTPUT)
Web wrapper program
Store the extracted data in excel sheet

Fig 1: Relational web wrapper working

II. ALGORITHM FOR RELATIONAL WEB WRAPPER

Suppose start is the starting time to start the process of extracting metadata from the webpage and end is ending time when the metadata had been extracted. M consists of the extracted data and meta_extracted_time is the time to extract the meta data.

Input: A webpage W.
Output: Metadata M from the webpage W.

Begin
Phase 1: Extracting the data from the webpage and calculation of time to extract the data

```
 start = current time ( Time at which the data extraction starts)
a = W.title
 b = W.location.href
 c = W.getElementsByTagName('meta')
M=a
      For (x = 0 To  y = c.length; x < y; x++)
      if (c[x].name.toLowerCase() == "description")
      description = c[x];
endif
endfor
M=M+description
for (var x = 0, y = c.length; x < y; x++) {
      if (c[x].name.toLowerCase () == "keywords") {
      description = c[x];
endif
endfor
M=M+description
 end = current time ( Time at which data extraction ends)
meta_extracted_ time = end - start;
```

Phase 2: Storing data to excel sheets

Suppose x is namespace for excel workbook.

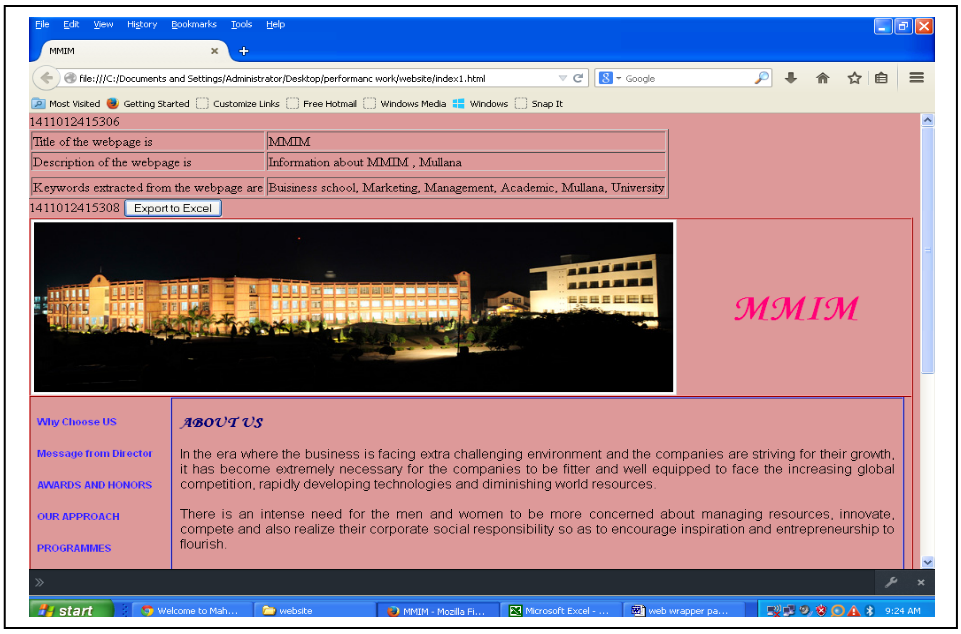
```
   xmlns: x="urn: schemas-microsoft-com: office:excel"
 uri = 'data: application/vnd.ms-excel;base64'
```

Create the excel worksheet using x as following:

```
<xml>
<x:ExcelWorkbook>
<x:ExcelWorksheets>
<x:ExcelWorksheet>
<x: Name> {worksheet}</x:Name>
<x:WorksheetOptions><x:DisplayGridlines/></x:WorksheetOptions>
</x:ExcelWorksheet>
</x:ExcelWorksheets>
</x:ExcelWorkbook>
</xml>

 function B(s)
Begin
return window.btoa (unescape (encodeURIComponent(s)))
End

 function F(s, c)
Begin
return s.replace (/ {(\w+)}/g, function (m, p) {return c[p]}) }
  End

function P(table, name)
Begin
if (!table.nodeType)

 table = W.getElementById(table)
endif

   M = {worksheet: name || 'Worksheet', table: table.innerHTML}
    window.location.href = uri + B (F (template, M))}
End
```

IV. Experimental Results

The relational web wrapper is developed on JavaScript platform. 3 Websites were developed in HTML and JavaScript for experimentation. The results from one of the websites are shown in fig 2, fig 3 and fig 4. Similar results are also available for other websites.



1411012415306

Title of the webpage is	MMIM
Description of the webpage is	Information about MMIM , Mullana
Keywords extracted from the webpage are	Buisiness school, Marketing, Management, Academic, Mullana, University

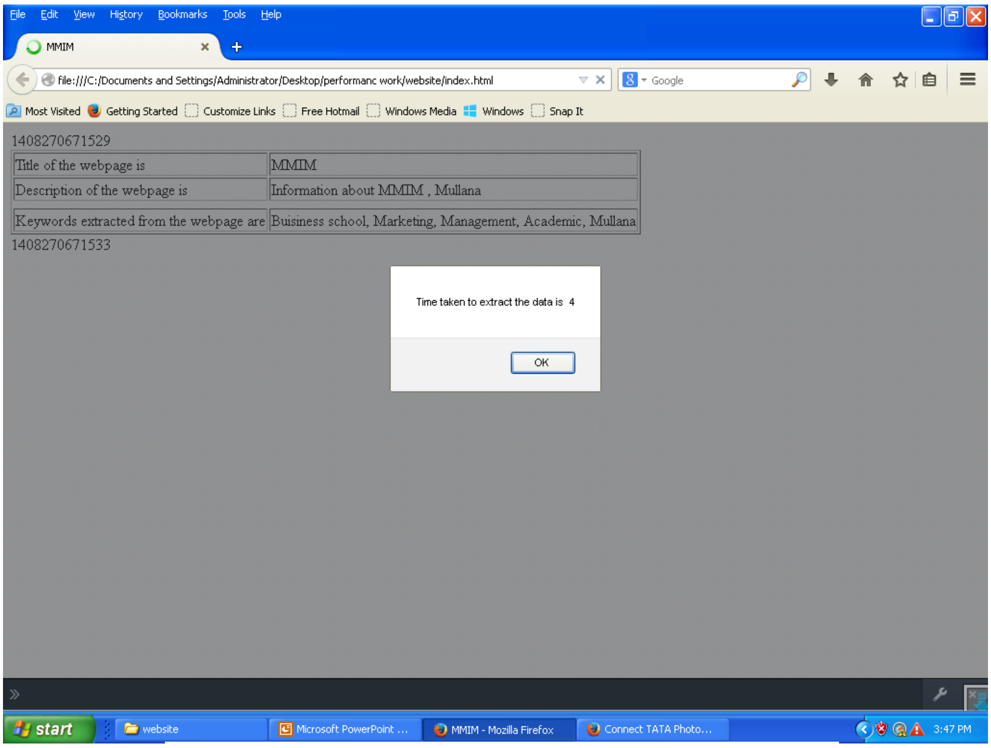
1411012415308 Export to Excel

ABOUT US

In the era where the business is facing extra challenging environment and the companies are striving for their growth, it has become extremely necessary for the companies to be fitter and well equipped to face the increasing global competition, rapidly developing technologies and diminishing world resources.

There is an intense need for the men and women to be more concerned about managing resources, innovate, compete and also realize their corporate social responsibility so as to encourage inspiration and entrepreneurship to flourish.

Fig 2: Webpage with extracted data



1408270671529

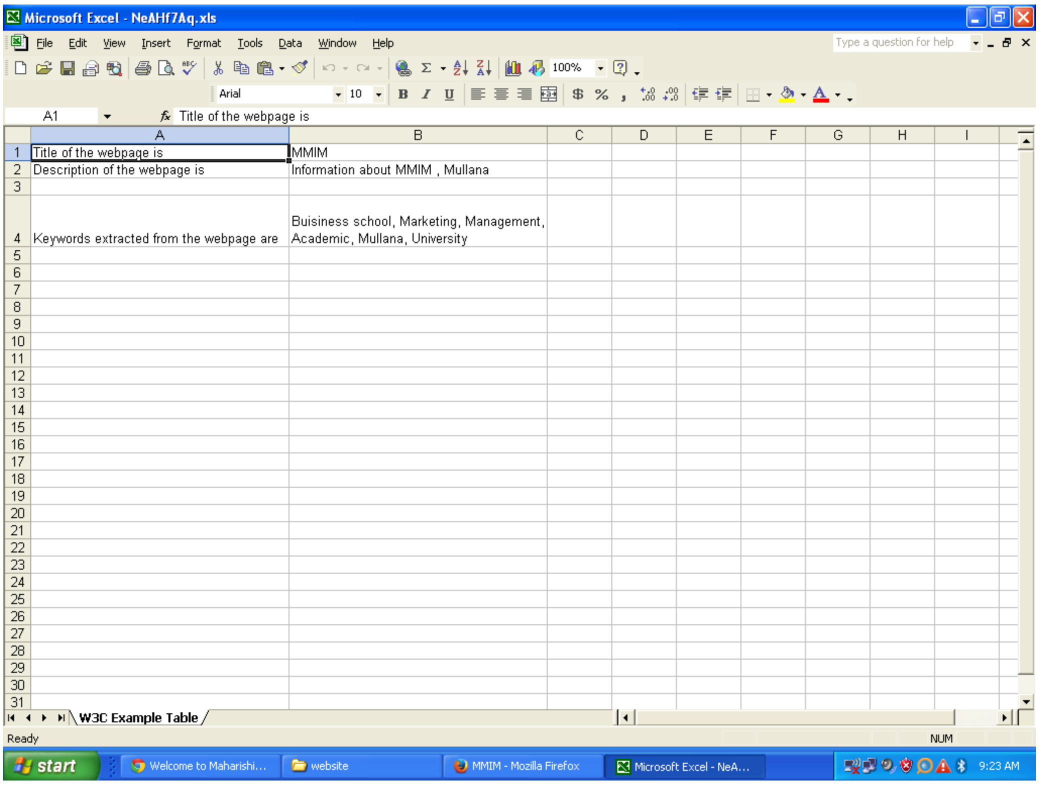
Title of the webpage is	MMIM
Description of the webpage is	Information about MMIM , Mullana
Keywords extracted from the webpage are	Buisiness school, Marketing, Management, Academic, Mullana

1408270671533

Time taken to extract the data is 4

OK

Fig 3: Extracted data and evaluation time



	A	B
1	Title of the webpage is	MMIM
2	Description of the webpage is	Information about MMIM , Mullana
3		
4	Keywords extracted from the webpage are	Buisiness school, Marketing, Management, Academic, Mullana, University

Fig 4: Storing the extracted data to excel sheet

PERFORMANCE EVALUATION OF RELATIONAL WEB WRAPPER

For this purpose 3 websites named MMIM, MMEC and MMCE each consisting of 10 WebPages was developed.

WEBSITE: MMIM

Table 1: Evaluation Time to extract the related data from website MMIM

Name of Webpage	Average Time to Extract the Data (Number of runs)				Average Time (ms)
	2	5	10	20	
Index.html	4.0	4.2	4.8	4.6	4.4
Life.html	6.5	5.4	5.2	5.2	5.575
Prog.html	6.0	5.2	5.0	5.0	5.3
Student.html	5.0	4.4	4.3	4.4	4.525
Approach.html	4.5	4.2	4.2	4.3	4.3
Message.html	4.0	4.4	4.5	4.4	4.325
Facilities.html	6.0	5.0	4.8	4.8	5.15
Knowledge.html	6.0	5.6	5.6	5.6	5.7

WEBSITE: MMEC

Table 2: Evaluation Time to extract the related data from website MMEC

Name of Webpage	Average Time to Extract the Data (Number of runs)				Average Time (ms)
	2	5	10	20	
Index.html	5.0	5.2	5.4	5.6	5.3
Life.html	7.5	7.4	7.2	7.4	7.375
Prog.html	7.2	7.2	7.2	7.3	7.225
Student.html	6.0	6.2	6.3	6.5	6.25
Approach.html	5.5	5.4	5.2	5.4	5.375
Message.html	5.2	5.4	5.6	5.6	5.45
Facilities.html	6.0	6.2	6.2	6.1	6.125
Knowledge.html	6.4	6.6	6.6	6.6	6.55

WEBSITE: MMCE

Table 3: Evaluation Time to extract the related data from website MMCE

Name of Webpage	Average Time to Extract the Data (Number of runs)				Average Time (ms)
	2	5	10	20	
Index.html	6.2	6.2	6.3	6.3	6.25
Life.html	5.5	5.4	5.4	5.4	5.425
Prog.html	6.2	6.2	6.2	6.3	6.225
Student.html	6.1	6.2	6.4	6.2	6.225
Approach.html	6.5	6.4	6.3	6.3	6.375
Message.html	6.2	6.2	6.6	6.6	6.4
Facilities.html	7.2	6.8	6.6	6.6	6.8
Knowledge.html	6.8	6.6	6.7	6.7	6.7

Accuracy: Accuracy is measured on the basis of following formula
Accuracy=Data extracted by the relational web wrapper/ Data extracted on manual basis (Actual existence)

Table 4: Accuracy measurement to extract the related data from website MMEC

Name of Webpage	%age of data extracted by relational web wrapper			%age of data extracted on manual basis		
	Title	**Description**	**Keywords**	**Title**	**Description**	**Keywords**
Index.html	100%	100%	100%	100%	100%	100
Life.html	100%	100%	95%	100%	100%	100
Prog.html	100%	100%	95%	100%	100%	100
Student.html	100%	100%	96%	100%	100%	100
Approach.html	100%	100%	95.4%	100%	100%	100
Message.html	100%	100%	95.4%	100%	100%	100
Facilities.html	100%	100%	96%	100%	100%	100
Knowledge.html	100%	100%	95%	100%	100%	100

The relational web wrapper provides 100% accuracy in extraction of title and description of document and more than 0.95 accuracy in extracting the keywords of most of the webpages.

V. CONCLUSION AND FUTURE WORK

Relational web wrapper provides an example of wrapper induction system which may extract related data or meta-data of the webpage efficiently and with minimal user effort. The web wrapper extracts data efficiently in terms of time taken to extract the data and accuracy provided. As per the definition of web wrapper the relational web wrapper can be applied only to extract meta-data from the webpage (single work).The work may further be extended to more dynamic content and saving the results to xml format. Accuracy of the web wrapper may be extended by applying some more accurate techniques on the same web wrapper. In future work the relational web wrapper developed may be compared with already existing techniques of web data extraction.

References

[1] Muslea I, Minton S and Knoblock CA, "Hierarchical wrapper induction for semi structured information sources", Journal of autonomous agents and multi-agent systems, vol. 4, pp. 93-114, 2001.

[2] R. Baumgartner, S. Flesca and G. Gottlob. "Declarative information extraction, Web crawling, and recursive wrapping with Lixto". In Proc. of Int. Conf. on Logic Programming and Nonmonotonic Reasoning, Vienna, Austria, 2001.

[3] R. Baumgartner, S. Flesca and G. Gottlob. "Visual web information extraction with Lixto". in the VLDB Journal, pp. 119–128, 2001.

[4] Senellart, P, Mittal A, Muschick D, Gilleron, R and Tommasi M, "Automatic wrapper induction from hidden-web sources with domain knowledge", WIDM, pp. 9-16, 2008.

[5] Soderland S, Fisher D, Aseltine J and Lehnert W, "CRYSTAL: Inducing a conceptual dictionary". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[6] Utku Irmak, Torsten Suel, "Interactive Wrapper Generation with Minimal User Effort", International World Wide Web Conference Committee (IW3C2), ACM, Edinburgh, Scotland, 2006.

[7] C. Chang and S. Lui. Iepad, "Information extraction based on pattern discovery" in the proceedings of the International World Wide Web Conference, 2001.

[8] Freitag D., "Information extraction from HTML: Application of a general learning approach" in the proceedings of the Fifteenth international Conference on Artificial Intelligence (AAAI) 1998.

[9] Kushmerick N., Weld D. and Doorenbos R., "Wrapper induction for information extraction" in the proceedings of the Fifteenth International Conference on Artificial Intelligence (IJCAI), pp. 729-735, 1997.

[10] Hsu C.N. and Dung M., "Generating finite-state transducers for semi-structured data extraction from the web". Journal of Information Systems vol. 23, No.8, pp. 521-538, 1998.

[11] W. Cohen, M. Hurst, and L. Jensen. "A flexible learning system for wrapping tables and lists in html documents". In the proceedings of the International World Wide Web Conference, 2002.

[12] Reinhard Pichler and Robert Baumgartner , "Web Data Extraction - Overview and Comparison of selected State-of-the-Art Tools", Wien Seminar at mit Bachelorarbeit, May 2011.

[13] Kushmerick. N., "Adaptive Information Extraction: Core technologies for Information agents". in Intelligent Information Agents R&D in Europe: An Agent Link perspective (Klusch, Bergamaschi, Edwards & Petta, eds.). Lecture Notes in Computer Science 2586, Springer, 2003.

[14] Soderland, S., "Learning to extract text-based information from the World Wide Web", in the proceedings of the third International Conference on Knowledge Discovery and Data Mining (KDD), pp. 251-254, 1997.

[15] Ciravegna F., "Learning to tag for information extraction from text", in the proceedings of the ECAI-2000 Workshop on Machine Learning for Information Extraction, Berlin, August 2000.