International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

Implementation of Web Data Extraction with Mashup Techniques for HTML Documents

Omkar Somji[1], Sufyan Shaikh[2], Parag Shelke[3], Mayur Tathe[4], B.Mahalakshmi[5]

[1]B.E.(Computer),
Pimpri Chinchwad College of Engg. Nigdi, Pune-44

[2]B.E.(Computer),
Pimpri Chinchwad College of Engg. Nigdi, Pune-44

[3]B.E.(Computer),
Pimpri Chinchwad College of Engg. Nigdi, Pune-44

[4]B.E.(Computer),
Pimpri Chinchwad College of Engg. Nigdi, Pune-44

[5]Asst. Professor,
Pimpri Chinchwad College of Engg. Nigdi, Pune-44

***ABSTRACT*- In the fast growing world the technology has improved a lot, so is the internet improving day by day, but as it grows the content of the web pages also increases, so is the irrelevant data which sometimes is not required by the user. It is possible for the users to extract the required data but as we know that most of the websites are of unstructured HTML format, it becomes more time consuming and costly for the users .HTML format only focuses on the presentation rather than querying the database due to unstructured content of their source pages , thus there is requirement of tool for gathering only the particular data rather than the irrelevant data and hence we propose a fully visual and interactive user interface based system tool(robot) using the programming language java in order to extract the unstructured data from the HTML web pages by creating the MASHUP through DOM tree modeling which uses the processes like data retrieval, data source modeling, data filtering, data integration, data visualization, such that structured data can be delivered to the user as per his/her requirements and without having any knowledge of the programming languages.**

Keywords: **Web Data Extraction: Making Mashup; Mashup Stages; HTML; DOM Tree**

I. INTRODUCTION

The comprehensive, extremely diversified and abundant information pool is nothing but the World Wide Web i.e. internet today, but the data available can be unstructured or semi-structured and there is need for extracting useful information in it. Internet is too much vast and structures of web pages are complex and it is a tough job to seek the essential data which poses a great challenge of how to extract useful information from the web, mostly which is in the form of semi-structured. Moreover, extracting useful information from the internet is necessary, which is required for the best decision-making.

In this paper, we propose a new method for generating a Mashup by using DOM tree modeling system that is automatically generated by using our system tool i.e. ROBOT. The Mashup created is used as a web based application which combines data or a particular function from two or more sources pages (web pages) in order to create new services in the new web pages by using the data warehouse created or by sending the SMS or E-MAIL of the essential information to the user. DOM tree is a useful platform to represent or to interact various HTML web pages document object in a form of tree structure or commonly called node-tree. With the use of DOM tree structure approach then the web browser could interpret the HTML tags structure from a web page very easily. This mean that all the nodes exist inside the HTML tags structure within a web page can be easily identified or modified its data structures.

The purpose of this method is to ease the Internet users to get the required data efficiently and effectively for the unstructured web pages. In order to implement this technique, a new Mashup creator tool called ROBOT has been proposed and developed. The ROBOT tool consists of algorithms having some

set of rules which is capable of creating the Mashup through a process of essential data extraction from the HTML source page. The process of creating the Mashup is initiated by mapping the whole structure of a webpage thorough HTML tags that has been successfully extracted and grouped from the single source page. Afterward, those HTML tags are -+grouped into Root, Parent, and Child Nodes by the ROBOT system. These groups of nodes will be transformed to a DOM tree structure model that automatically builds by ROBOT tool. This DOM tree structure will be used as a main reference for computation process in every stages of Mashup creation, such as: Data Retrieval, Data (Source) Modeling, Data Cleaning/ Filtering, Data Integration and Data Visualization. The structured data obtained is then transferred to the user as per the requirements through techniques like SMS and E-MAIL.

The main goal of this paper is that computer literate users should be able to use ROBOT tool without having to write complicated queries or to program or understand programming concepts, in order to implement web data extraction and creation of the Mashup. The other goal is that the ROBOT tool should be applicable to various web sources especially for data exchange and indirectly solve each issue during the Mashup staging process.

II. LITERATURE REVIEW

A. Web Data Extraction

Web data extraction [3] is a software system that automatically and repeatedly extracts data from web-pages with changing contents and delivers the extracted data to the user. The task of web data extraction is divided into five different functions. (1) Interaction of user with the web-pages to get desired information. (2)Extractor generation and execution where extractor is a program that identifies desired data on target pages extract the data and transform into structured format. (3) Scheduling which allows repeated application of previously generated extractors to their respective target pages. (4) Data transformation involves filtering, transforming and integrating data extracted from one or more sources and structuring the result into a desired output format. (5) Delivering: Resulting structured data to external applications such as email servers or SMS servers.

B. Mashup

Mashup is a web application that combines data or functionality from two or more external sources to create a new service. Making a Mashup need to deal with five basic issues [6, 7] they are:

- Data Retrieval involves extracting data from web pages into a structured data source. In addition to figuring out the rules to extract particular data from HTML pages [1, 3] the structure of data on a page or the location of data which can span multiple web pages can make the process more complicated.
- Data Source Modeling is the process of assigning the attribute name for each data column so a relationship between a new data source and existing data sources can be deduced.
- Data Cleaning/ Filtering is required to fix misspellings and transform extracted data into an appropriate format.
- Data Integration specifies how to combine two or more data sources together using a database join operation.
- Data Visualization takes the final data generated by the user and displays it.

C. RoboMaker

RoboMaker is a robosuite application which functions to create and debug various types of robots. Users can create any type of robot according to the task user want to do. For instance robot which is assign to search data from various different web pages in the internet. Search robot which is assign to search part by part in web-page in HTML format and those parts will be presented in another format such as portal or new web-page. RoboMaker creates an IDE for different types of robots. RoboMaker provides complete features in its GUI from the interactive visual programming, capability of full debugging to easy access online assistance on a sensitive context issue (Heier, 2008).

D. DOM Tree

The Document Object Model (DOM)[2] is a cross-platform and language-independent convention for representing and interacting with objects in HTML and XML documents. The HTML or XML DOM views a HTML or XML document as a tree-structure. The tree structure is called a node-tree. All nodes can be accessed through the tree. Their contents can be modified or deleted, and new elements can be created [4]. The node tree depicted in figure 2 below shows the set of nodes, and the connections between them. The tree starts at the root node and branches out to the text nodes at the lowest level of the tree:

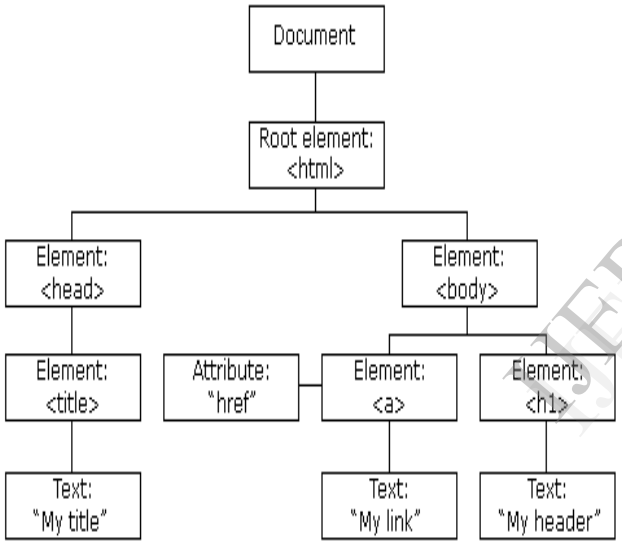

Figure 1 . DOM Tree

III.DESIGN AND ARCHITECTURE

A. *Robot System*

The Robot is designed to implement the web data extraction [1] that will make user to extract more relevant data from HTML web pages without knowledge of programming.

Below diagram show the architecture of system

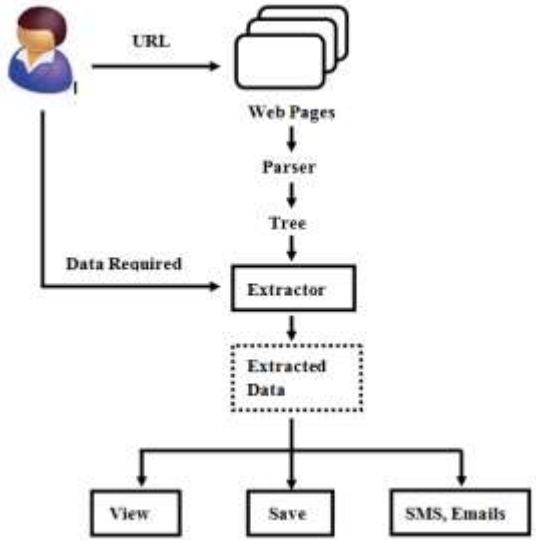

Figure 2. System Architecture

The user must specify URL from the web page that he wants to extract. The URL is given to parser for parsing purpose. Parser generates a parse tree. From the generated tree user can decide which specific data to extract. Extractor uses Mashup process to extract information. Mashup process which includes five steps:

- Data Retrieval involves extraction of data from HTML web pages.
- Data Modeling involves modeling of data.
- Data Filtering/Cleaning involves fixing of misspellings and remove unwanted data.
- Data Integration involves combining of data from two or more sources.
- Data Visualization involves how the data is made visible to the user i.e. the user can either view the data, save in database or SMS, emails.

B.DOM Tree design for Robot

The Document Object Model tree has been used for basic structure approach of web data extraction process. The DOM Tree is constructed based on organization of HTML structures (tags, elements, attributes).Using a DOM Tree is an effective way to identify a list or extracting data from the web page. The DOM Tree can be constructed from the tables

given below.The Robot will extract the relevant data required.For example path for Roll No will be (Body,tr,td,Roll No).

Sr No	Roll no	Marks	Percentage
1	255	40	65
2	261	45	60
3	263	75	55
4	210	80	50

Sr No	Roll no	Marks
1	255	40
2	261	45
3	263	75
4	210	80

Figure 3 .Data Source Tables

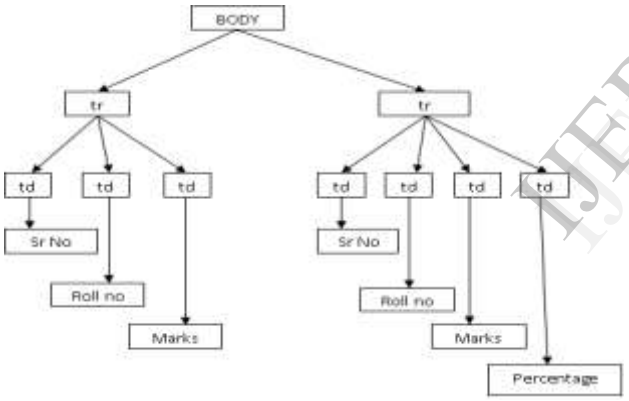

Figure 4 .DOM Tree implemented in Robot System

IV.IMPLEMENTATION OF ROBOT

A. Data Retrieval

Figure 5 .Data Retrieval interface

In this stage, user will extract the data from the HTML source page from the given URL, by clicking on extract button user will get extracted data.

B. Data Source Modeling

In this stage the extracted data will be processed for restructuring by the robot and performing parsing on the HTML tag for each element of data and modeling of that data for user purpose.

C. Data Filtering

In Data Filtering user can remove the irrelevant data and extract the information effectively. For example if there are advertisements on a website the user can filter it and once the data is filtered it will not display advertisements again.

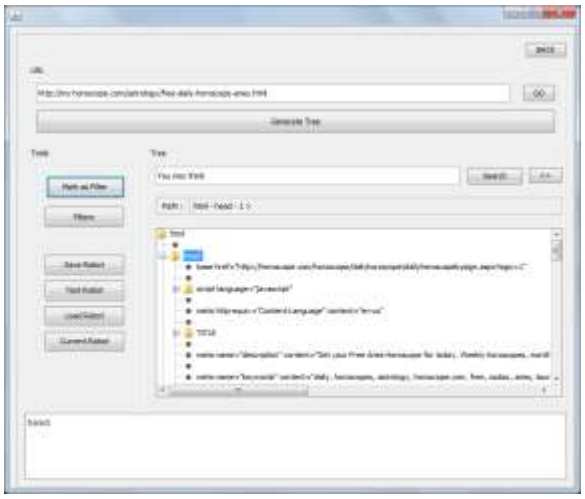

Figure 6 .Before Filtering

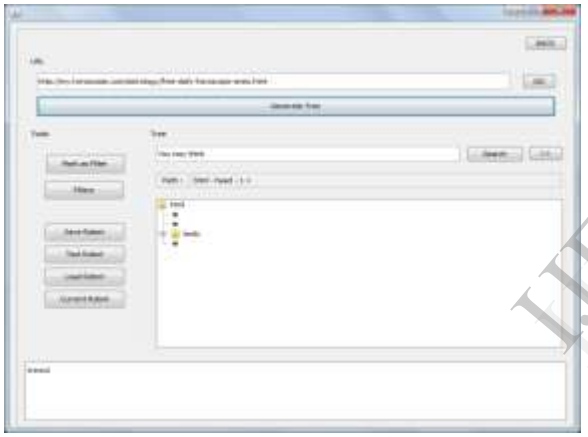

Figure 7. After Filtering

D. Data Integration:-

In Data Integration, the extracted data from different sources can be combined into single data repository. For example there are separate webpage for displaying horoscope of Aries and Leo, if user wants to view both at a time then by applying mashup on both WebPages the user can view the information of both Aries and Leo together.

E. Data Visualization:-

This stage is actually "the end stage" of making mashup. It is simply the representation of data in visual form. Data Visualization takes data generated by the user and displays it in various forms. For example the user wants to know the horoscope of Aries then his horoscope will be sent as SMS or Email to user.

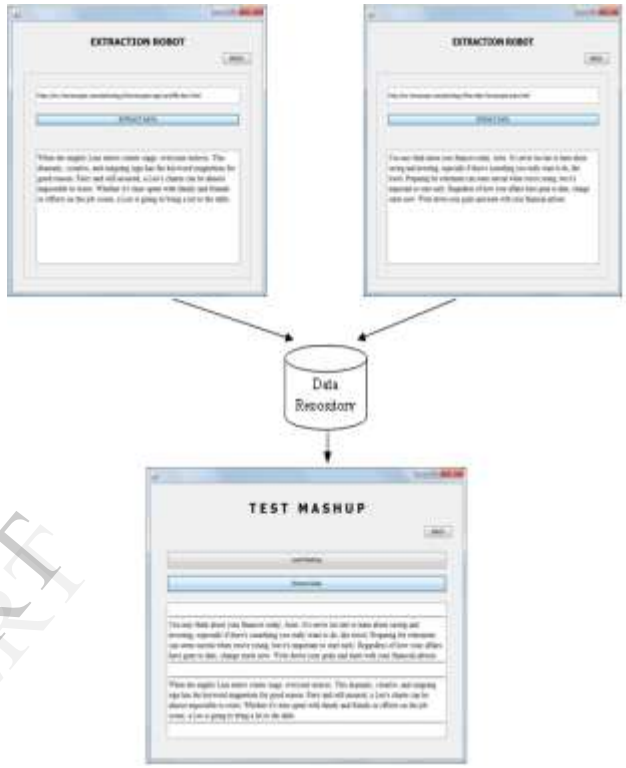

Figure 8 .Data Integration

V. PERFORMANCE AND EVALUATION RESULTS

Evaluation results for implementing visual web extraction and making a mashup between RoboMaker and Robot. The performance measuring criteria are:

1) Number of steps required for building a mashup
2) Precision of extracted data

Figures shows the result evaluation, X axis maps Mashup stages and Y axis shows the number of steps.

	Data Retrieval	Data (Source) Modeling	Data Cleaning/ Filtering	Data Integration	Data Visualization
RoboMaker	5	1	0 (N/A)	0 (N/A)	1
Robot	2	2	1	6	1

Table 1. Result Evaluation of RoboMaker and Robot

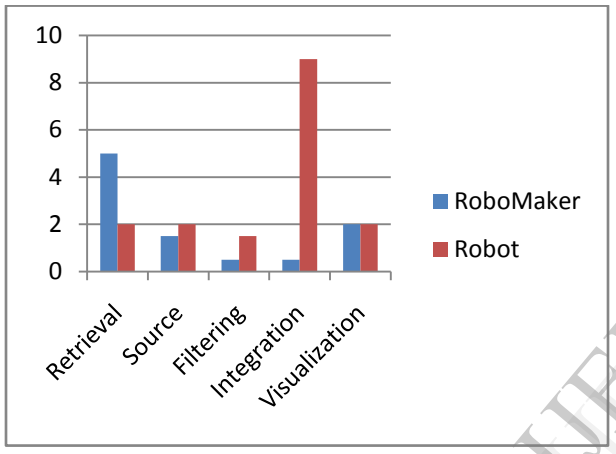

Table 2. Result Evaluation in Clustered Cylinder Chart

VI. CONCLUSION AND FUTURE WORKS

We conclude that this paper positively contributes the new approach and technique to implement web data extraction by creating MASHUP. The interactive user interface will ease the users to perform the extraction data from the source pages without having the expertise in the computer programming. In terms of future works, the capability of the ROBOT system can be improved with the help of semantic web approach [5] also the data extraction from non-HTML file formats can be made possible.

REFERENCES

[1] Rudy AG. Gultom,Riri Firtri Sari, Bagio Budiardjo ”Implementing Web Data Extraction And Making Mashup with Xtractorz”,2010

[2] Jer Lang Hong, Fariza Fauzi“Tree Wrap-data Extraction Using Tree Matching Algorithm”Majlesi Journal of Electrical Engineering Vol. 4, No. 2, June 2010

[3] Robert Baumgartner, Wolfgang Gatterbauer, Georg Gottlob “Web Data Extraction System”, 2010.

[4] W3Schools.com, “The HTML DOM node Tree”, http://www.w3schools.com/htmldom/dom_nodetree.asp, last accessed 12 January 2013

[5] K. Lerman, A. Plangrasopchok, and C. A. Knoblock, “Semantic Labeling of Online Information Sources”, In Pavel Shaiko (Eds.) IJSWIS, Special Issue on Ontology Matching, 2007.

[6]C.A. Knoblock, K. Lerman, S. Minton, and I. Muslea, “Accurately and reliably extracting data from the web: A machine learning approach”, Intelligent Exploration of the Web, Springer-Verlag, Berkeley, CA,2003.

[7] D. Huynh, S. Mazzocchi, and D. Karger, “Piggy Bank: Experience the Semantic Web Inside Your Web Browser”, In Proc. of ISWC, 2005.

[8] Wikipedia, “Mashup (web application hybrid)”, http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid), last accessed 12 October 2009.