Performance Comparison of Web Data Extraction Techniques

Neeraj Raheja
Associate Professor
Department of Computer Science and Engineering,
M.M.University, Mullana (Ambala), Haryana, India

Vijay Kumar Katiyar, PhD
Professor
Department of Computer Science and Engineering,
M.M.University, Mullana (Ambala), Haryana, India

ABSTRACT
Websites in today world consist of a large amount of data as per the requirements of the users. So web data extraction systems helps user in extracting the required data from these types of websites. The basic techniques used for web data extraction are manual and web wrapper. Web wrapper further consists of wrapper induction and automatic approaches. A lot of methods are available which uses wrapper induction and automatic methods. This research work provides performance comparison of manual, web wrapper induction and automatic approaches on the basis of methods chosen as manual (By manual efforts), nX1 (web wrapper induction), DEPTA and MDR (Automatic). The results are compared on the basis of various parameters like precision, recall, F-measure and data extraction time.

Keywords
Web data extraction, manual, web wrapper, nX1, Depta, MDR.

1. INTRODUCTION
The data is growing over the websites at a very fast rate. Hence web data extraction systems are required to extract the required data from this type of large scale website. Web data extraction helps in search engine optimization [14].

Basically there are two types of web data extraction techniques known as manual and web wrapper generation [10]. In manual technique the data is extracted on the basis of requirement by manual efforts. In web wrapper there are two methods known as web wrapper induction and automatic [11] [12].

In web wrapper induction a program known as wrapper is developed by the programmer depending upon a particular pattern or template of the website. In this approach the developer first label a set of sample WebPages and find out the pattern or structure, which most of WebPages are using. Then according to that pattern the web wrapper program is developed. Hence this approach is limited to a number of websites [7]. This approach requires a lot of manual efforts in developing the program. Example of such a system is WIEN, STALKER etc.

In case of automatic technique the supervised learning approach is used and can be used for multiple types of web data extraction. Hence automatic approach can be applied over a number of websites [8] [9]. Examples of such type of systems are ROADRUNNER, DEPTA [3], MDR[4], VIPS etc.

2. LITERATURE REVIEW
A lot of methods are proposed by different researchers depending upon the techniques available for web data extraction i.e. manual, web wrapper and automatic.

Chang et.al [1] proposed a string matching method for extracting the content of the WebPages, it extract the data by observing the distance in the tag tree or DOM tree.

Neeraj Raheja and V.K.Katiyar [2] proposed a method named nX1, which extracts the data from the webpage of a website which uses the template of the form nX1 table. It converts the semi-structured form of data to structured form and then It extracts the data using XSL method.

Liu et.al [3] proposed a technique called DEPTA which uses tree alignment instead of tag strings, which exploits nested tree structures to perform more accurate data extraction.

Liu et.al [4] proposed an automatic approach called MDR which is based on string matching method. It builds the tag tree of the webpage and extracts the content based upon tags used for matching.

Liu et.al [5] proposed an automatic web data extraction method called NET which was used to extract web data from both flat and nested types of data records. Given a page as input NET first builds a tag tree based on visual information of the webpage then it performs a post-order traversal of the build tree and matches all subtrees of the tree using a tree edit distance method and visual cues.

Hiremath et.al [13] proposed an automatic web data extraction approach based on visual clue of the webpage. It works on the basis two steps i.e. Identification and Extraction of the data regions based on visual clues information and Identification of data records and extraction of data items from a data region. This technique is not dependent upon the tags used in the content and can also extract the data from contiguous as well as non-contiguous content.

3. TECHNIQUES USED FOR WEB DATA EXTRACTION
This research work uses both techniques used for web data extraction:

i. **Manual:** This technique uses manual effort for extracting the web data from WebPages.

ii. Web wrapper :

i) Web wrapper induction: The technique chosen for this method is nX1 table method [2].

ii) Automatic: The techniques chosen for this method are DEPTA [3] and MDR [4].

Working of nX1 method (Web Wrapper Induction method)

This method extracts the data of the webpage based upon template. It looks for the WebPages of a particular website and checks their template. If the template found of the format nX1 table then it converts the webpage from HTML (unstructured) format to XML format (structured). Then it uses XSL display method of XML to extract the data of the webpage.

Algorithm for nX1 approach:

Input: A webpage W using the template nX1.
Output: Extracted data from webpage W.
Step 1: Check whether W uses nX1 table type template
If (yes) then goto step 2 else goto step 5.
Step 2: Convert W into XML format (X)
Step 3: Apply XSL query on X
Step 4: Store the results of Step 3.
Step5: End.

Working of MDR (Automatic Methods)

MDR is an automatic method of web data extraction which works on the basis of tags used in the webpage for building the content. It mines the content by matching with the tags specified. It works on the basis of following two concepts:
a) Data Region: Content on a webpage is presented in a contiguous area of the webpage and uses same type of tags to build the content.
b) Tag tree creation: It is build according to the structure of the webpage.

Algorithm for MDR

Step 1: Build the tag tree of the webpage using structure of the webpage.
Step 2: Find generalized nodes in the tag tree with the properties:
i) The nodes with the same parent.
ii) The nodes which are adjacent.
Step 3: Apply string matching method to check whether generalized nodes contains data records.
Step 4: Extract content from each data records from step 3.

Working of DEPTA (Automatic Method)

DEPTA is an automatic method to extract content of the webpage. It works on the basis of tree edit distance method instead of string matching method used in MDR for extracting the content from the webpage. It works better than the MDR because DEPTA can also extract the data even when multiple records are contained in a single record.

Algorithm for DEPTA

Step 1: Build the tag tree of the webpage.
Step 2: Apply Tree matching method to find similarity score of the subtrees of the tag tree.
Step 3: Find data regions using the similarity score in the step 2.
Step 4: Extract the content from the data regions found in step 3.

4. EXPERIMENTAL RESULTS

For comparing the results of different techniques three websites named website1 (consisting of 105 WebPages (Dataset1)), website2 (consisting of 85 WebPages (Dataset2)) and website3 (consisting of 95 WebPages (Dataset3)) were developed. The experiment extracts the noise free data from the WebPages of these websites. The parameter used for comparison among the above mentioned techniques are

i) scalability i.e. on how many types of WebPages the particular technique can be applied
ii) Recall, precision and F-measure i.e. accuracy of data extracted or how much data can be extracted.
iii) Extraction time i.e. how much time is taken to extract a particular data

It evaluates the performance using parameters recall, precision, F-measure which are defined as following:-

$$Precision\ (P) = (LEC - (LEC - LEP + LM))/LEC$$

$$Recall\ (R) = (LEC - (LEC - LEP + LM))/LEP$$

$$F - measure\ (F) = 2 * ((P * R)/(P + R)$$

$$Data\ extraction\ Time = Time\ to\ extract\ data$$

Whereby · LEC refers to extracted content. · LEP refers to expected content · LM refers to missing content

Scenario 1: in this Scenario the Dataset1 uses the template nX1. The data was extracted using all the four methods. The results are as shown below:

Table 1: Recall in scenario 1

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	100	10	12
MDR	68	59	62
DEPTA	83	85	78

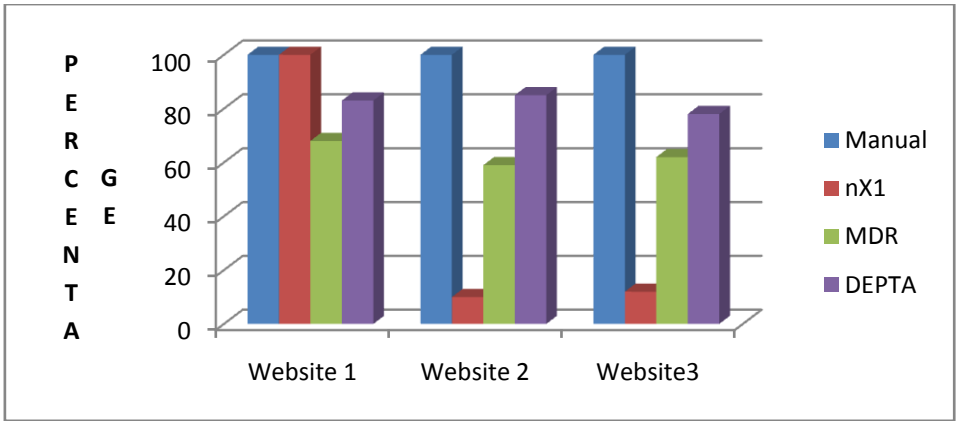

Figure 1: Recall in Scenario 1

Table 2: **Precision in scenario 1**

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	98	14	13
MDR	86	88	88
DEPTA	92	94	92

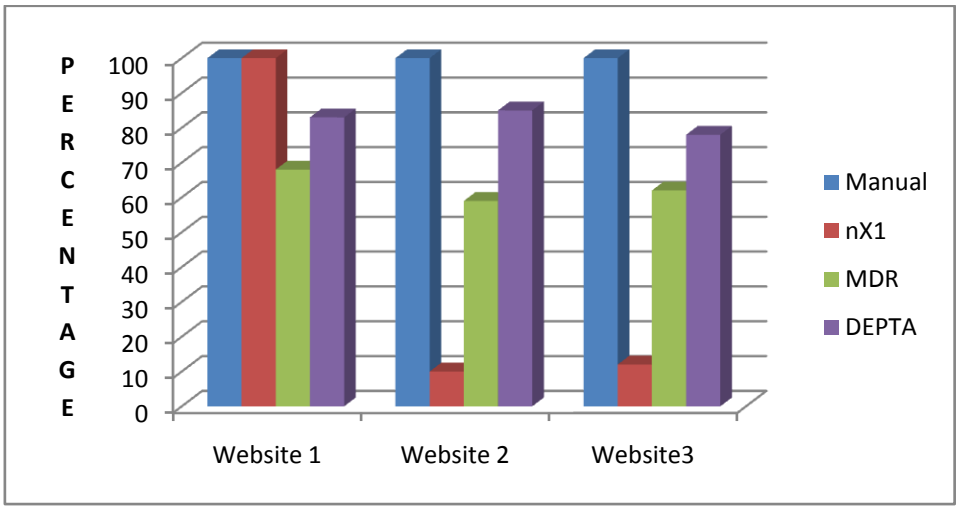

Figure 2: Precision in Scenario 1

Table 3: **F-measure in scenario 1**

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	98.98	11.66	12.48
MDR	75.94	70.63	72.74
DEPTA	87.268	89.27	84.42

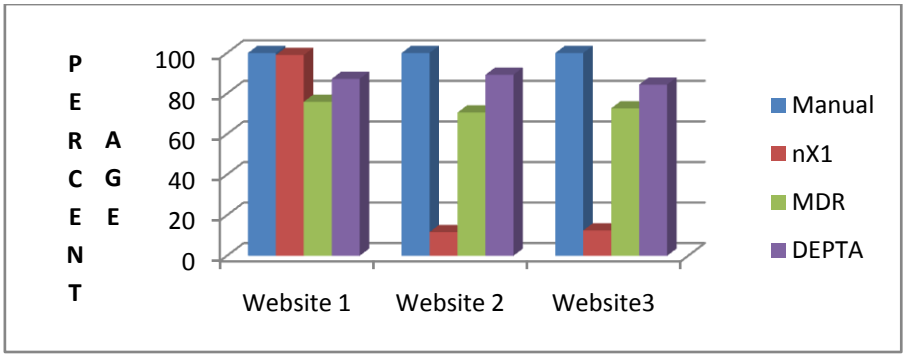

Figure 3: F-measure in Scenario 1

Scenario 2: in this Scenario the Dataset2 uses the template nX1. The data was extracted using all the four methods. The results are as shown below:

Table 4: **Recall in scenario 2**

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	13	100	18
MDR	66	59	62
DEPTA	83	85	78

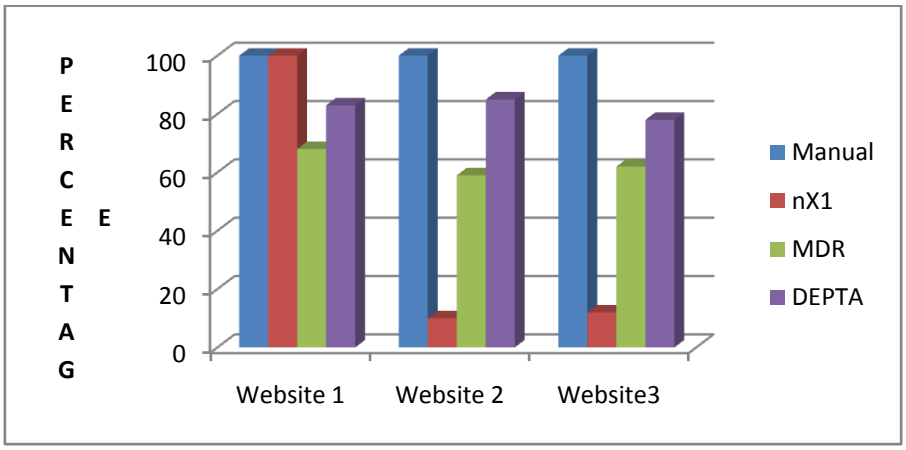

Figure 4: Recall in Scenario 2

Table 5: Precision in scenario 2

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	16	96	18
MDR	86	88	88
DEPTA	92	94	92

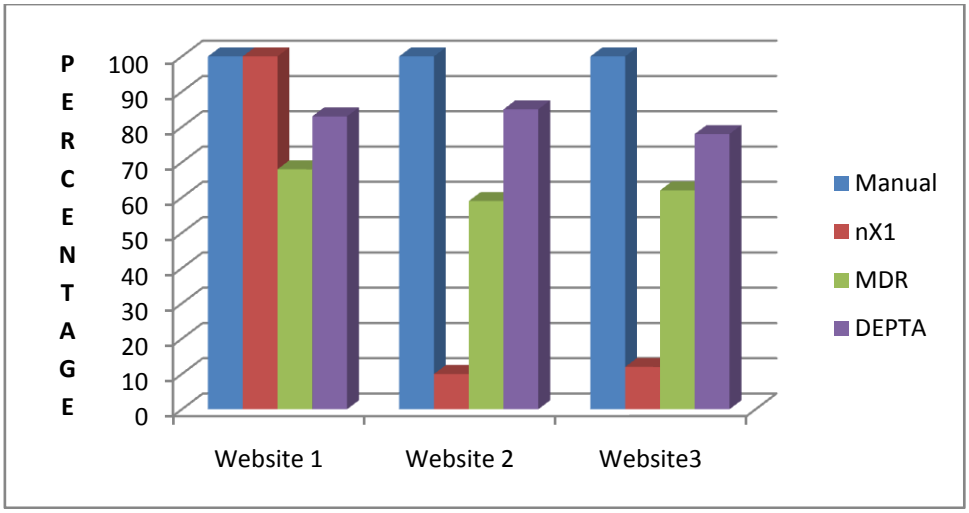

Figure 5: Precision in Scenario 2

Table 6: **F-measure in scenario 2**

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	14.34483	97.95918	18
MDR	74.68421	70.63946	72.74667
DEPTA	87.26857	89.27374	84.42353

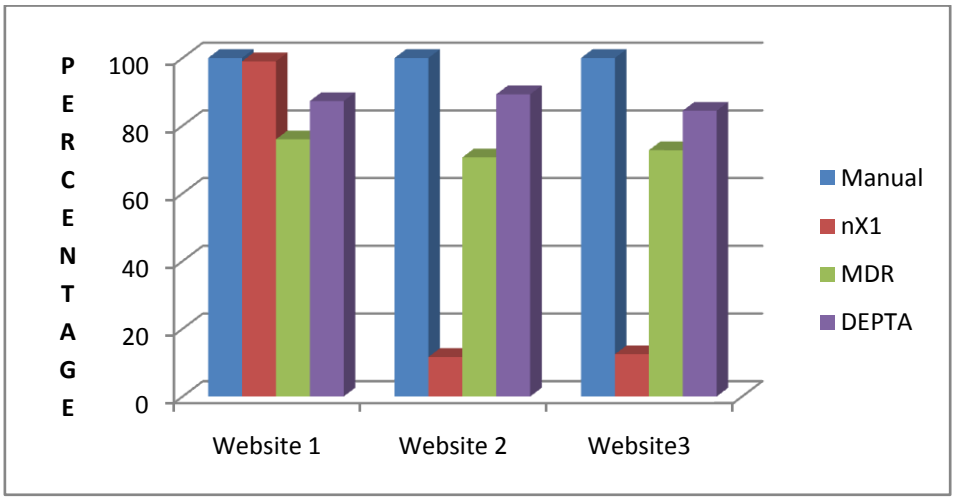

Figure 6: F-measure in Scenario 2

Scenario 3: in this Scenario the Dataset3 uses the template nX1. The data was extracted using all the four methods. The results are as shown below:

Table 7: **Recall in scenario 3**

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	15	22	100
MDR	65	59	63
DEPTA	83	85	78

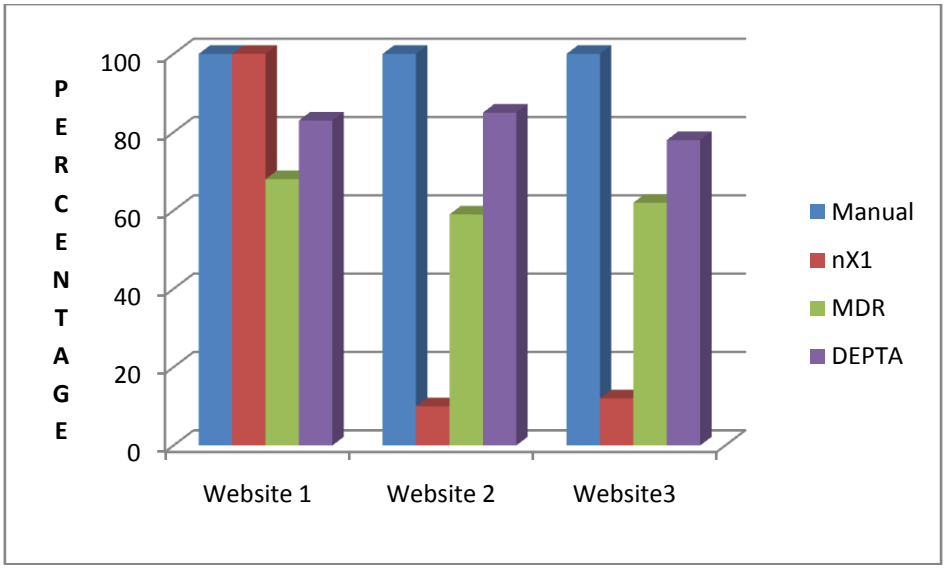

Figure 7: Recall in Scenario 3

Table 8: **Precision in scenario 3**

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	16	20	96
MDR	86	88	88
DEPTA	92	94	92

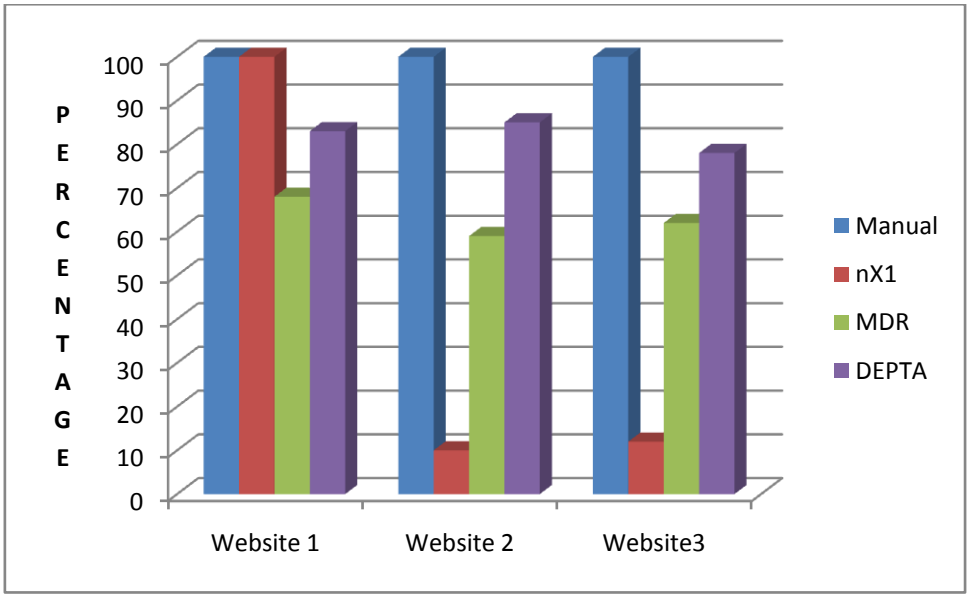

Figure 8: Precision in Scenario 3

Table 9: **F-measure in scenario 3**

	Website 1	Website 2	Website3
Manual	100	100	100
nX1	15.48387	20.95238	97.95918
MDR	74.03974	70.63946	73.43046
DEPTA	87.26857	89.27374	84.42353

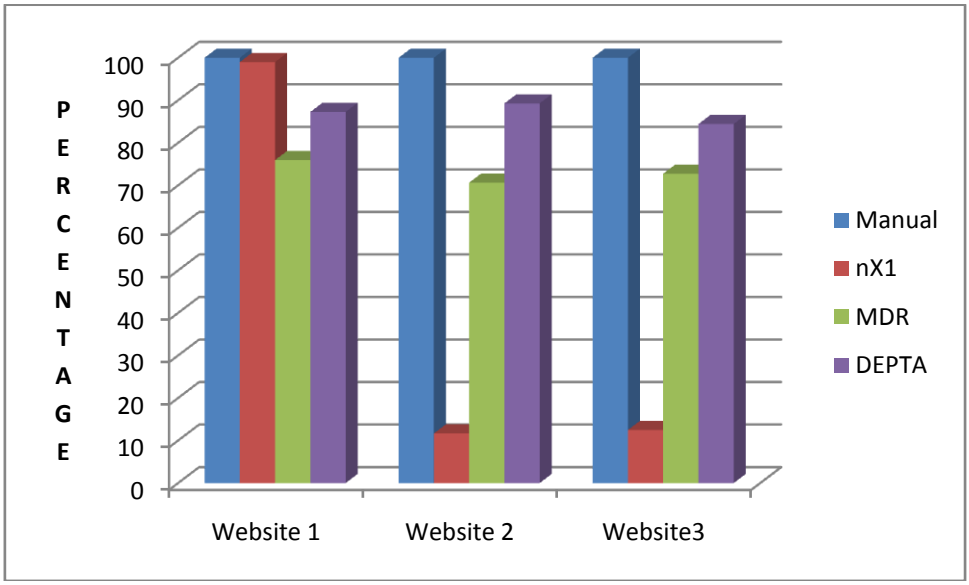

Figure 9: F-measure in Scenario 3

Table 10 : Average time(in ms) to extract the content of a webpage

	Website 1	Website 2	Website3
nX1	1076	1192	1185
MDR	576	652	498
DEPTA	1030	1085	1154

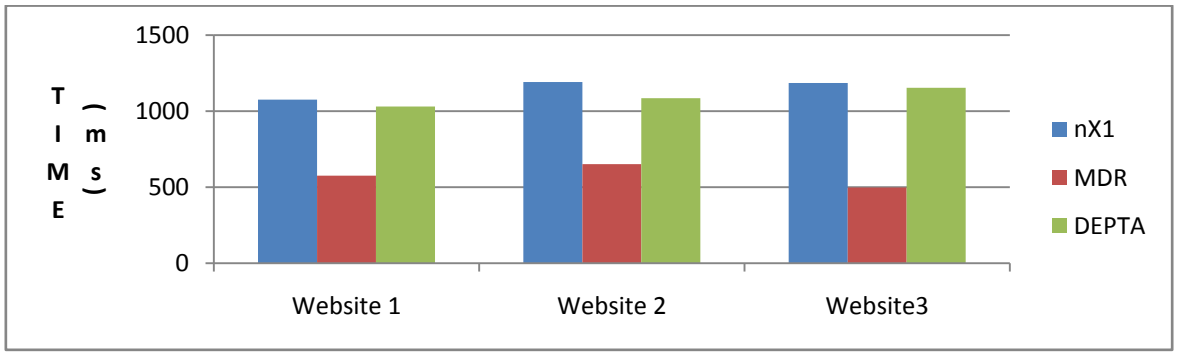

Figure 10: Average time to extract the content of a webpage

5. CONCLUSION

The performance comparison based on the techniques used for web data extraction is shown in this research work. The results show that manual technique is very time consuming but accuracy level is very high. The web wrapper induction technique (nX1) performs better if the websites template matches with the program developed for it, because web wrapper induction programs are developed by looking at the website templates. Finally two techniques for automatic web wrapper (DEPTA and MDR) were applied to the websites and DEPTA performs better than the MDR in terms of accuracy. But it takes some more time to extract the data because of tree matching instead of string matching. These techniques takes less time than the other techniques and can be applied to any type of website without the limitation of a particular template as in case of web wrapper induction.

6. REFERENCES

[1] C. Chang, S. Lui.," IEPAD: Information extraction based on pattern discovery", in WWW, pp. 681-688, 2001.

[2] Neeraj Raheja, V.K.Katiyar, " A Noise Reduction Approach based on n x 1 table and XSL display method for efficient web data extraction" , IJCA International Journal of Computer Applications (0975 – 8887) Vol. 64, No.11, pp. 12-17, February 2013.

[3] Y. Zhai, B. Liu. , "Web data extraction based on partial tree alignment", in WWW, pp. 76-85, 2005.

[4] B. Liu, R. L. Grossman, Yanhong Zhai, "Mining data records in Web pages", in KDD, pp. 601-606, 2003.

[5] Bing Liu and Yanhong Zhai, "NET - A System for Extracting Web Data from Flat and Nested Data

Records", proceedings of 6th International Conference on Web Information Systems Engineering(WISE-05), 2005.

[6] Valter Crescenzi et.al. "ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites", 2001.

[7] Zhai Y and Liu B, "Extracting Web data using instance-based learning" in *WISE-05*, 2005.

[8] Arasu A and Garcia-Molina H., "Extracting Structured Data from Web Pages", in *SIGMOD-03*, 2003.

[9] Lerman K., Getoor L., Minton, S. and Knoblock C, "Using the Structure of Web Sites for Automatic segmentation of Tables", *SIGMOD-04*, 2004.

[10] J. Hammer, H. Garcia Molina, J. Cho, and A. Crespo. , "Extracting semi structured information from the web", i*n proceedings of the Workshop on the Management of Semi-structured Data, 1997.*

[11] Chang C-H., Lui, S-L., "IEPAD: Information Extraction Based on Pattern Discovery", *WWW-01*, 2001.

[12] Kushmerick N., "Wrapper Induction: Efficiency and Expressiveness.*Artificial Intelligence"*, 2000.

[13] P. S. Hiremath, Siddu P. Algur, "Extraction of Data from Web Pages: A Vision Based Approach", International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol 3, No 3, pp. 623-632, 2009.

[14] Faustina Johnson, Santosh Kumar, "Web Content Mining Using Genetic Algorithm", *in Advances in Computing, Communication, and Control Communications in Computer and Information Science (Springer),* Vol. 361, pp. 82-93, 2013.