IJASR International Journal of Academic and Scientific Research
ISSN: 2272-6446 Volume 2, Issue 4 (November-December 2014), PP 15-25
www.ijasrjournal.org

Web Data Record Extraction Prototype Based on Partial Tree Alignment

San San Tint[1], May Thandar Soe[2]
[1](*Research and Development II / University of Computer Studies, Mandalay, Myanmar)*
[2](*Master Student of Computer Science / University of Computer Studies, Mandalay, Myanmar)*

ABSTRACT: On *various kinds of the Webs, data records are contained a huge amount of information in structured objects today. By containing their web page, such data records are interested to be mined by the users who look for their important things in the web. Most of the data records are extracted from the web page containing the list of products and services doe to determine what are needed for people. This paper presents Data Extraction based on Partial Tree Alignment (DEPTA) which is implemented by combing Mining Data Record in the Web (MDR) and Align and Extract Data Items (AEDI) from the Identified Records. This paper especially presents a portion of example products on the web like furniture web page as a prototype. The proposed system is able to mine and extract information both contiguous and noncontiguous data records.*

Keywords - *Data extraction, Data record extraction, Furniture, Web data record, Web mining, Partial Tree Alignment, contiguous, noncontiguous*

INTRODUCTION

The World Wide Web has saturated with HTML documents containing the most important connection to various information resources [1-16]. Every web can be transformed into HTML documents. Data records in the web are structured as objects for the information. Mining Data Record is applied to be beneficial value information from various sources to achieve value added services. Data extraction plays an important role on useful results from webs [14]. Webs pages are constructed by the combing data to the templates. The templates make objectives to display different ways of information. If a user views the web pages for his or her needs, multiple pages are chosen as input for the user demands. If one page is used as a input, web data extraction will be the record level information. There are many webs data extraction tools [13]. Most business applications are critical for people daily uses. Typically automatic extraction tools are demanded to collect information on the web.

1. Related Works

The major Web data extraction approaches and compares them in which three dimensions named why an Information Extraction (IE) system fails to handle some Web sites of particular structures, classify IE systems based on the techniques used and measure the degree of automation for IE systems [5]. The Information Retrieval (IR) systematically arranged in a way for further processing, introduction of various methodologies like wrapper and the extracted information labeling are considered as the concepts [4]. In this paper construction of prototype of web data record extraction based on Partial Tree Alignment (DEPTA) is approached with small furniture web page [9-12].

They surveyed on different HTML structure based technique to scrap data from web pages. The users want to retrieve with the help of search input query in the Internet containing large amount of data. But they can get the return results with multiple dynamic output records from the web. Due to needs of flexible information extraction system web pages can be converted into machine process able structure which is essential for much application needed to be extracted & annotated automatically which is challenge in data mining [3],[5][6][8].

In order to perform efficient contemporary alignment method first pair wise and then holistically, threshold index formula has also been calculated to find the data regions and to perform clustering methods, to speed up the search engine, clustering the similar data regions and to perform efficient identification of web

pages including the concept of page ranking , similar Query result pages and applicable also in data integration and comparison shopping within the applications of this extraction of data records [2].

2. Background Theory

In order to find all data records formed by table and form related tags (i.e., table, form, tr, td, etc.) Mining Data Records (MDR) is used. A large majority of Web data records are formed by them. The algorithm is called MDR (Mining Data Records in Web pages). It currently finds all data records formed by table and form related tags, i.e., table, form, tr, td, etc. A large majority of Web data records are formed by them. Each data region in a page contains multiple data records [10]. Need to align multiple tag trees in order to produce a single database table with all the corresponding data items/fields in the same column of the table Partial Tree Alignment Algorithm use for aligning multiple trees [9-11].

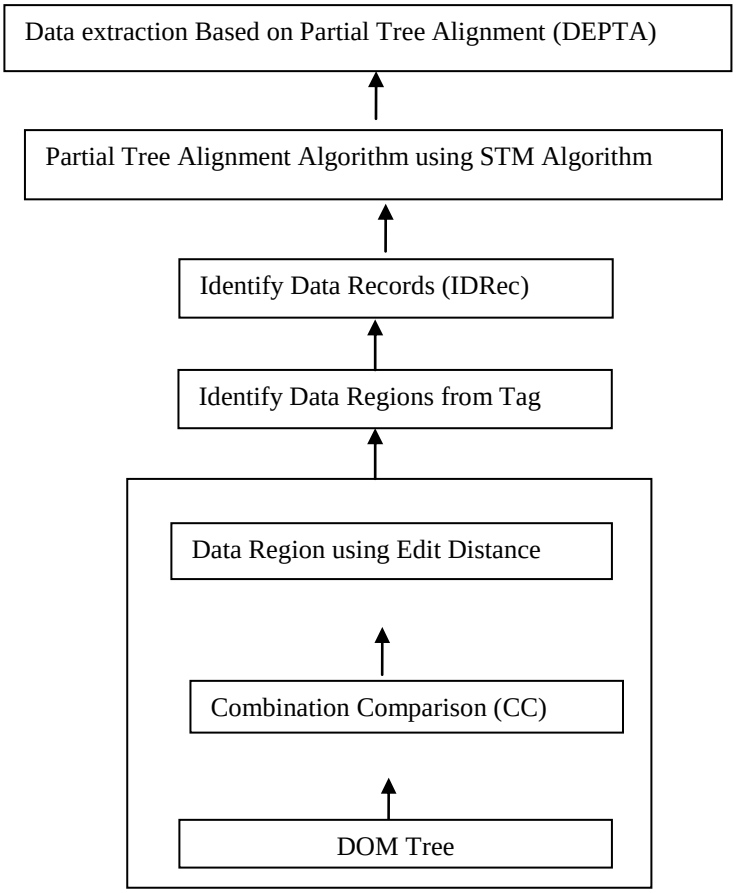


Figure 1. procedure of overview methods

Mining Data Records is useful because it allows us to integrate information from multiple sources to provide value-added services. MDR uses a string matching algorithm and it regards the HTML tags of a page as a string. These techniques perform based on data records and string matching technique.

- Data extraction based on partial tree alignment (DEPTA)

This method consists of two steps: (1) Identifying individual records in a page. And (2) Aligning and extracting data items from the Identified records [10].

- Data Record Identification

MDR: Mining Data Records

Given a single page with multiple data records, MDR extracts data records, but not data items (step1).

MDR is based on two observations about data records in a Web page and a tree matching algorithm. Consider both

- Contiguous

- non – contiguous records

Two Observations: A group of data records are presented in a contiguous region (a data region) of a page and are formatted using similar HTML tags. A set of similar data records is formed by some child sub trees of the same parent node [9].

- MDR Algorithm

Algorithm MDR (Node, K)

1. if TreeDepth(Node) >= 3 then
2 . CombComp(Node.Children, K);
3 . for each ChildNode ∈ Node.Children
4 . MDR(ChildNode, K);

(Node= any node; K= generalized node;)

- Mining Data Regions

Find every data region with similar data records.

Definition: A generalized node (or a node combination) of length r consists of r (r≥1) nodes in the HTML tag tree with the following two properties:

1. the nodes all have the same parent and
2. the nodes are adjacent.

Definition: A data region is a collection of two or more generalized nodes with the following properties:

1. The generalized nodes all have the same parent.
2. The generalized nodes are all adjacent.
3. Adjacent generalized nodes are similar.

- DEPTA: Extract Data from Data Records

Once a list of data records are identified, we can align and extract items in them

- Multiple tree alignment

It is need multiple alignments as we have multiple data records. Most multiple alignment methods work like hierarchical clustering, and require n2 pair wise matching. Optimal alignment/ matching is exponential A partial tree matching algorithm is proposed in DEPTA to perform multiple tree alignment [16].

- The partial Tree Alignment Approach

Choose a seed tree: A seed tree, denoted by Ts, is picked with the maximum number of data items.

Tree matching:

1. For each unmatched tree Ti (i≠s),
2. Match Ts and Tr
3. Each pair of matched nodes are linked (aligned)
4. For each unmatched node nj in Ti do
5. Expand Ts by inserting n into Ts if a position for insertion can be uniquely determined in Ts.
6. The expanded seed tree Ts is then used in subsequent matching.

- Sample Tree Matching (STM)
- A group of data records and a set of similar data records are considered in this system. In DEPTA, we need to apply STM algorithm for Multiple Tree Alignment.

3. System Overview

The system shows the procedure of the prototype model in which the most of parts are computed for sequential steps. It starts with loading HTML documents from Web pages in procedure 1. Constructing tag trees by using DOM trees is procedure 2. In procedure 3, it needs to mines the data region by computing the combination comparison and labeling similar nodes to denote each similar individual node and node combination. In To identify the data regions, it is needed to find Edit Distance between tag strings. There are three algorithms: (1) Combination Comparison (2) Mining Data Region (3) Edit Distance in procedure 3.Above all procedures are parts of Mining Data Region (MDR).

In procedure 4, the system identify the data records from generalize nodes in which two nodes and more than two nodes. In Multiple tag trees of multiple data records are aligned partial tree alignment algorithm and simple tree matching in procedure 5. After getting matching of the tree in fig 2, the system identifies the Data Records nodes as trees and matrixes. By applying Data extraction Partial Tree Alignment (DEPTA), it can call Simple Tree Matching (STM) then it produces the output as the tables and appears two data records as on a web page.

This system has two output data records for the furniture web page like fig. 9 and fig. 10. First Output Data Records, DataRecord1is one row and two columns and second Output Data Record 2 is two rows and two columns respectively. Our work focuses on data extraction from web page for small example web page. As the output it has information about two pieces of Dressers including their prices which are found to be particularly suitable for furniture web data extraction. It provides visible records for the users who find out the things particularly in web related their oriented environments.

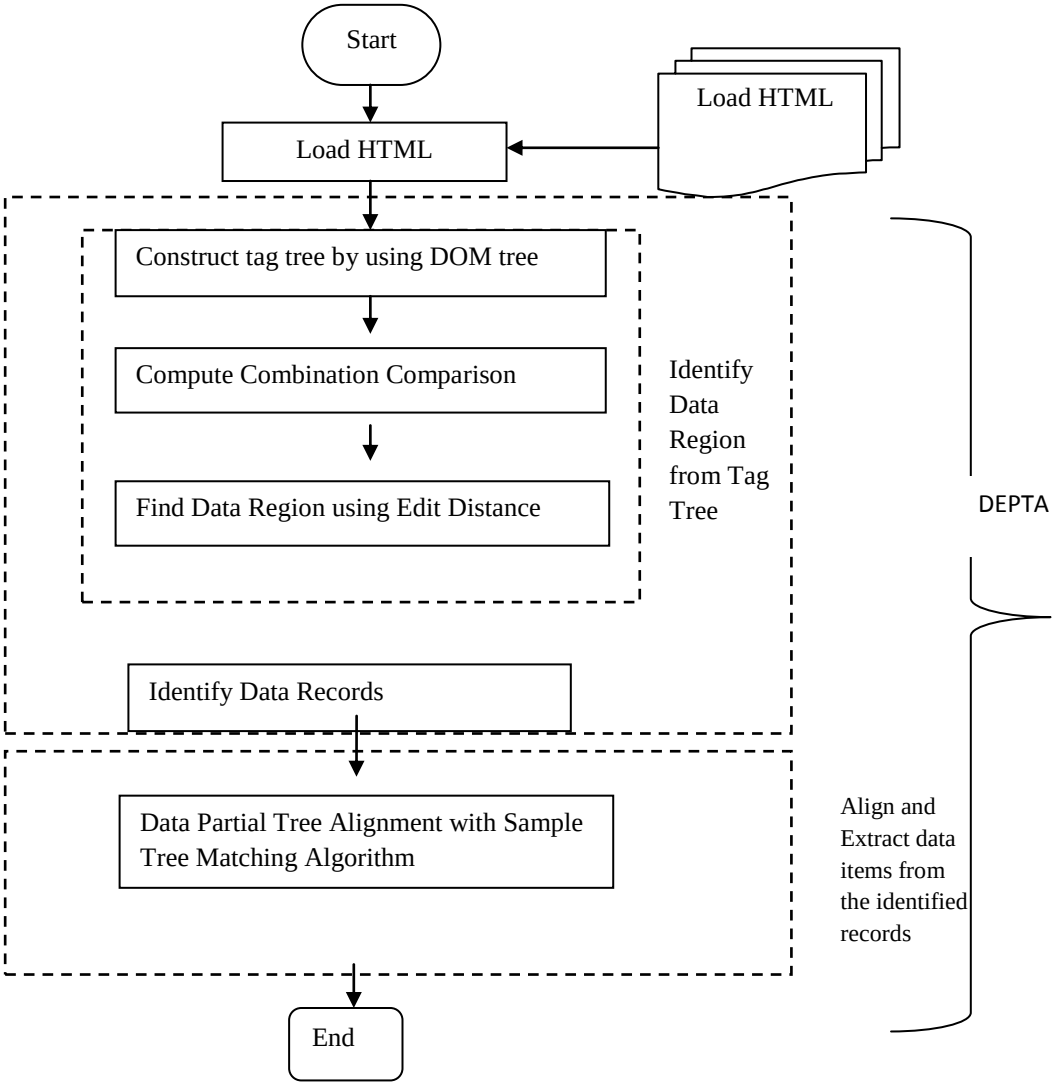


Figure 2. System overview

4. Implementation of Our Model

Procedure 1: Building HTML tag trees

Loading HTML source of Example Web Page is used for DOM tree. Below figure is an example web page for our model. This figure contains only for a prototype web page including Dresser and Mirror. It has three rows and two columns.

Example Web Page

Figure 3. Input web page for our prototype

Procedure 2: Building DOM Trees

We need to construct for DOM Tree from HTML documents in web page. In figure, tr is row and td is column. The word, img presents image and the word, br presents break line from HTML documents source.

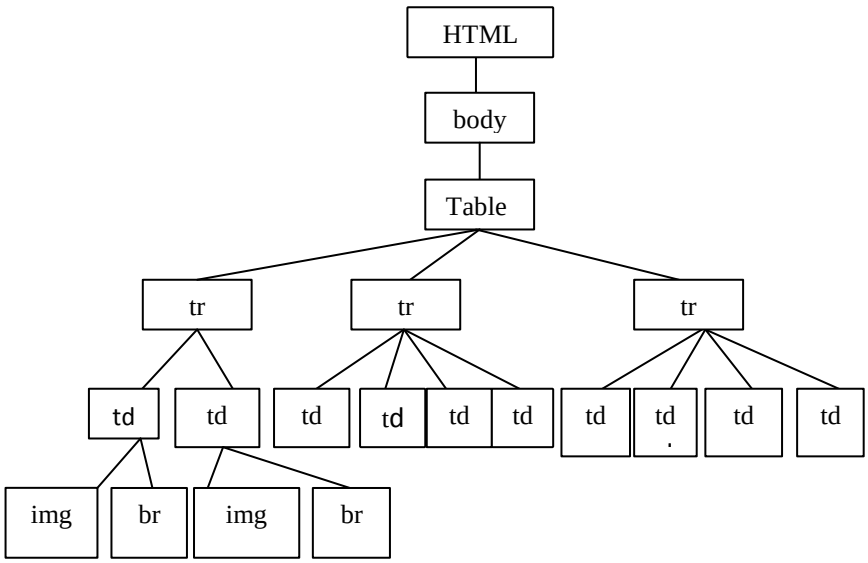


Figure 4. DOM tree of our prototype model

In this example, firstly we can describe the name of trs are tr(a), tr(b) and tr(c). As figure, each tr has nodes. Tag Nodes are

e.g <tr></tr>
<td></td>
<img>

.

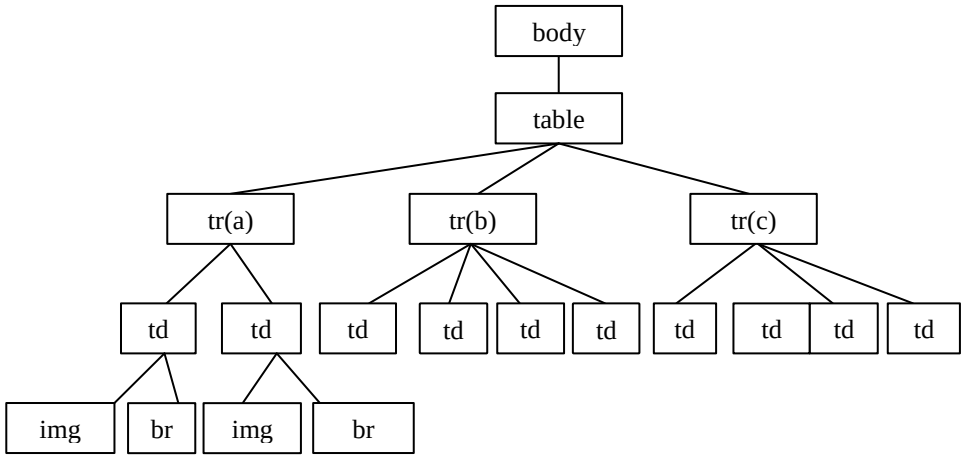


Figure 5. Name of nodes for our model

Procedure 3: Mining Data Region (MDR)

Combination Comparison, (a, b),(b, c),where

1. a=<tr td img br td img br>,
2. b=<tr td td td td>
3. c=<tr td td td td>

MDR Algorithm

1. Algorithm FindDRs(Node,K,T),Node='table',K=1,T=0.3
2. If TreeDepth(table)>=3(true)
3. Node.DRs=IdentDRs(1,table,1,0.3)
4. IdentDRs(1,table,1,0.3)
5. maxDR=[0,0,0]
6. i=1,
7. f=start=1;
8. flag=true;
9. j=f=1;
10. Distance(table,1,1)=(a,b)=10<0.3,a=<tr td img br td img br>,b=<tr td td td td>
    flag=false; and exit inner loop;
11. f=2
12. flag=true;
13. j=f=2
14. Distance(table,1,2)=(b,c)=0<0.3, b=<tr td td td td>, c=<tr td td td td>
15. curDR=[1,2,2]
16. flag=false;
17. if(maxDR[3]<curDR[3])(0<2) and (maxDR[2]=0)(true)
18. maxDR=[1,2,2]
19. where 1 means the number of nodes in a combination
20. 2 means the location of the start child node of the data region
21. second 2 means the numbers of nodes involved in the data region.
22. If(maxDR[3]!=0)(true)
23. If(maxDR[2]+maxDR[3]-1!=size(Node.Children); 2+1=3(false)
24. return maxDR=[1,2,2]
25. Node.DRs=IdentDRs(1,table,1,0.3)
26. =[1,2,2]
27. tempDRs= Ø;
28. foreach child Є Node.Children do
29. FindDRs(a,1,0.3)
30. If(TreeDepth(a)>=3)(false)

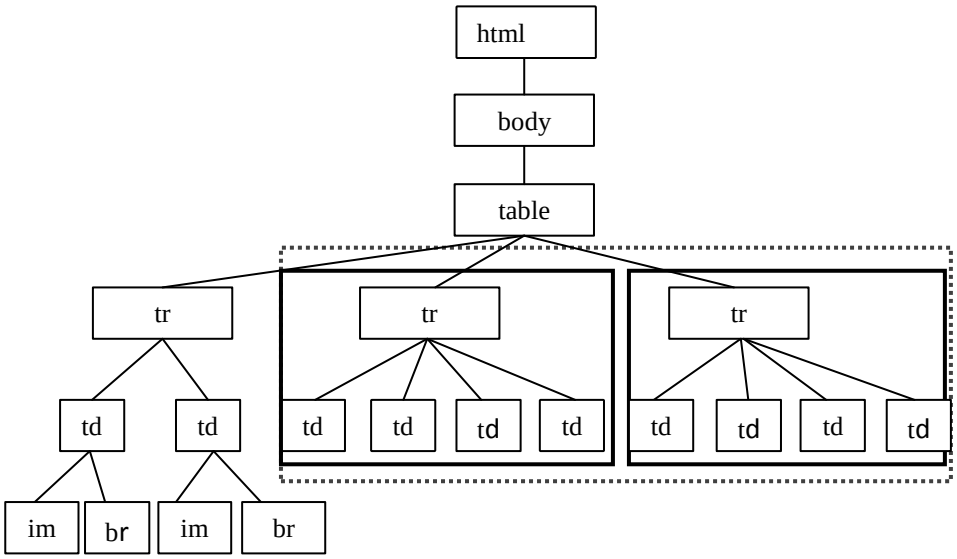


Figure 6. Data Region for Our Prototype

—— Generalized nodes
------ Data region

Procedure 4: Find Data Region (Edit Distance Algorithm)

1. Stringx= "tr", Stringy= "tr",
2. x = 'tr', y = 'tr'
3. add=ED(tr, t)+1 → x = tr, y =t
4. =1+1= 2 add=ED(tr, '')+1 → x=tr, y=''
5. = 2+1=3 return 2 (x.length)
6. remove=ED(t,t)+1→ x=t, y=t
7. = 0+1 =1 add=ED(t, '')+1 → x=t, y=''
   =1+1= **2** =1
   (x .length)
8. remove=ED('',t)+1 → x='', y=t
   = 1+1=**2** =1 (y.length)
9. change=ED('','')+0 → x='', y=''
   =0+0= **0** =0 (x.length)
   min(2, 2, 0) =0
10. change=ED(t, '')+1 → x = t, y = ''
    =1+1= **2** return 1 (x.length)
    min(3, 1, 2) = 1
11. remove=ED(t, tr)+1 → x=t, y =tr
    =0+1= 1 add=ED(t, t) + 1 → x=t, y=t
12. =0+1= 1 add=ED(t, '')+1 → x = t, y=''
    = 1+1=2 return 1 (x.length)
13. remove=ED('', t)+1 → x='', y=t
    =1+1=2 return 1 (y.length)
14. change=ED('', '')+0 →x='', y=''

=0+0=0 return 0
min(2, 2, 0) = 0

15. remove=ED(‘’, tr)+1 → x=’’, y=tr
= 2+1=3 return 2 (y.length)

16. change=ED(‘’,t)+ → x=’’, y=t
= 1+1=2 return 1 (y.length)
min(1, 3, 2) = 1
17. change=ED(t, t)+0 → x = t, y=t
= 0+0=0 add=ED(t, ‘’)+1 → x = t, y = ‘’
18. = 1+1=2 return 1 (x.length)
19. remove=ED(‘’, t)+1 → x = ‘’, y =t
= 1+1=2 return 1 (y.length)

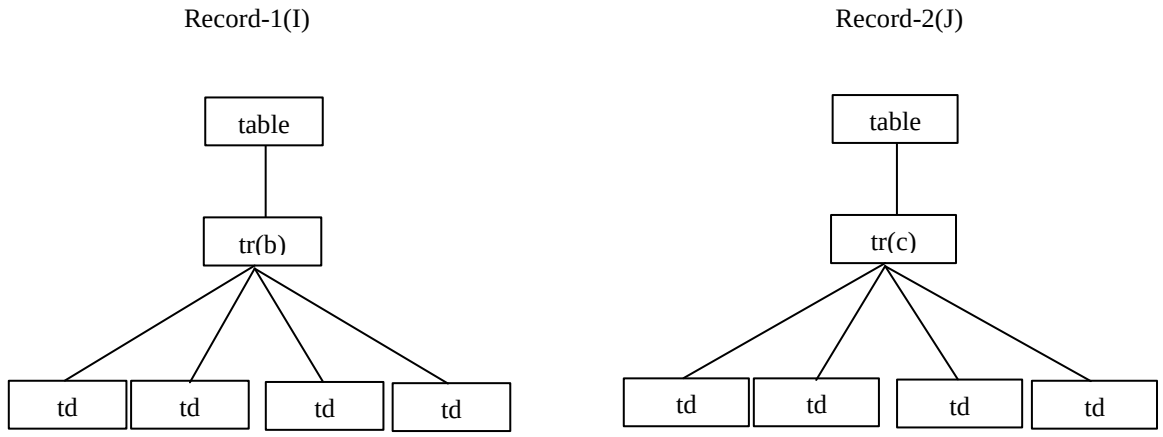
20. change=ED(‘’, ‘’)+0 → x=’’, y=’’
= 0 return 0 (y.length)
21. min(2, 2, 0) =0
22. min(2, 1, 0) = 0
23. Result is 0 (edit distance)
24. If G consists of only one tag node

Table 1. Possible Configuration of Data Records of Our Prototype

<td>		<td>	
<td>	<td>	<td>	<td>
<td>	<td>	<td>	<td>

Procedure 5: Identifying Data Records

1. Find DataRecords-1(G)
2. G is a data table row
3. G itself is a data record-(2)
4. DataRecord not in Data Region
5. Match each tag string of the children of the sibling nodes of tr



Simple_Tree_Matching (I, J)

M_{I-J}

	0	1(c)
0	0	0
1(a)	0	5

W_{I-J}

	1(c)
1(a)	5

Figure 7. Trees and Matrixes

Procedure 6: DEPTA

PatrialTreeAlignment(S),S={I,J}

1. Ts=I,S={J}
2. flag=false;R=Φ;I=false;
3. While(S≠ Φ)
4. Ti=J;S= Φ ;
5. Simple_Tree_Matching(I,J)
6. L=alignTrees(I,J);
7. J is not completely aligned with I (false)
8. If(L has new alignment) or (I is false)(true)
9. flag=true;
10. If(S= Φ) and flag=true(true)
11. S=R= Φ, R= Φ, flag=false; I=false;

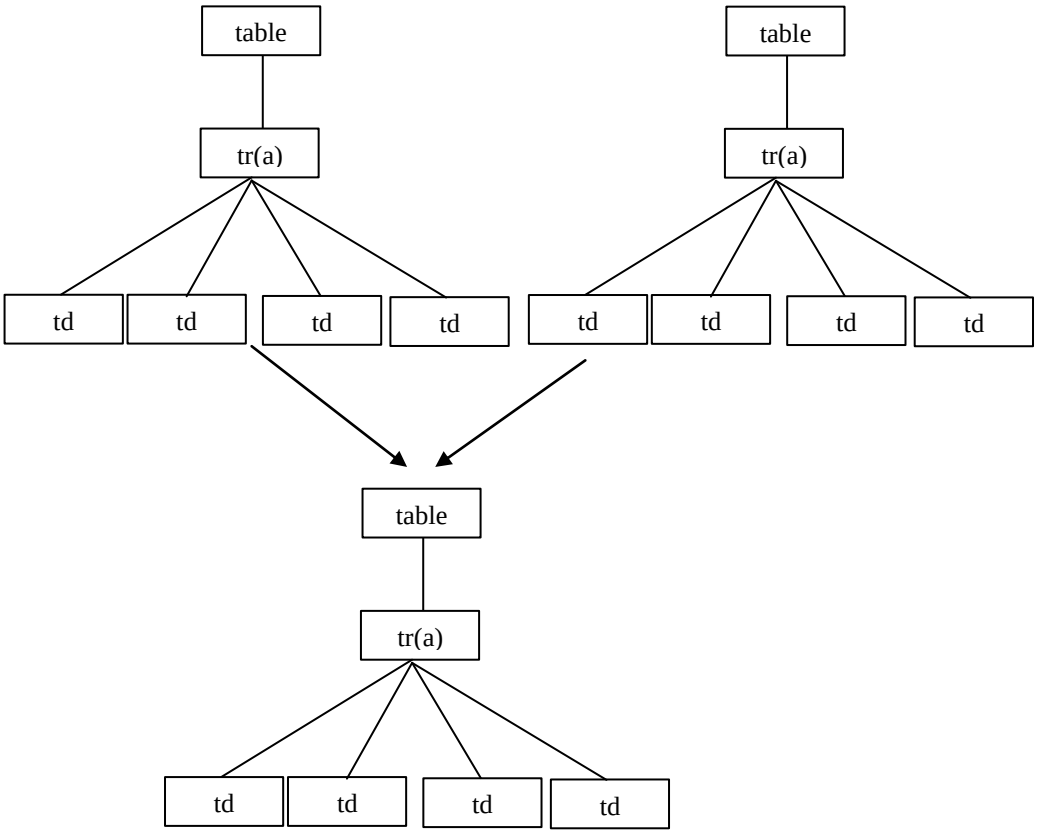


Figure 8. Tree for output table

Table 2. Output Data Table (1 indicates a data item)

	td	img	br
T1	1		
T2	1		
T3	1	1	1

Output Data Records

DataRecord1


Highland Merlot Dresser & Mirror Magnolia White Dresser & Mirror

Figure 9. First record with one row and two columns

Data Record 2

Was	Today	Was	Today
$499.99	$474.99	$354.55	$355.55

Figure 10. Second data record with two rows two columns

CONCLUSION

In this system, data record extraction in web pages is implemented with Mining Data Records (MDR) algorithm and Partial Tree Alignment Algorithm (DEPTA) for aligning multiple trees based on small web page like furniture web page. This design with effective techniques is to mine data records in any web pages. These algorithms are able to discover not only contiguous data records but also non-contiguous data records. Many people want to collect the records including interested things in the web. This system provides the users who need to find their wants in the web without spending more time in daily life. Mostly well known structured applications are demanded to be more feasible as visible views for the users who have to access everyday needs their lives.

Acknowledgement

Our heartfelt thanks go to all people, who support us at the University of Computer Studies, Mandalay, Myanmar. This paper is dedicated to our parents. Our special thanks go to all respectable persons who support for valuable suggestion in this paper.

REFERENCES

[1] V. B. Kadam, V. B. Kadam , and G. K. Pakle et al, Data Extraction Using DOM Tree and Selectors, *International Journal of Computer Science and Information Technologies, Vol. 5 (2)* , 2014, 1403-1410.
[2] S. M. E. Suganya, R. M. E. Devi, and T. M. E. Thangam, Data Extraction of a Novel Method for Clustering Alignment based on Combining Tag and Value Similarity, *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2,* February 2014.
[3] V. D. Mohod, and J. V. Megha, A Survey on Data Extraction of Web Pages Using Tag Tree Structure, *International Journal of Computer Science and Information Technologies, Vol. 5 (3)* , 2014, 4361-4363.
[4] Y. W. Wanjari, D. B. Gaikwad, V. D. Mohod,and S. N. Deshmukh, Data Extraction and Annotation for Web Databases using Multiple Annotators Approach- A Review, *International Journal of Computer Applications (0975 – 8887) Volume 88 – No.18,* February 2014.
[5] C. H. Chang, M. Kayed, M. R. Girgis, and K. Shaalan, A Survey of Web Information Extraction Systems, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* TKDE-0475-1104.R3 1.
[6] D. P. Krishna, T. S. Latha, and T. R. Reddy, Extracting Web Data Based On Partial Tree Alignment Using Fivatech, *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3*, March 2012.
[7] V. B. Kadam , and G.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, 1655-1658.
[8] G. V. R.Lakshmi ,and B. N. Swamy, Web Data Identification and Extraction, *International Journal of Electronics and Computer Science Engineering 1862,* www.ijecse.org ISSN- 2277-1956.
[9] P.V.P. Sundar, Towards Automatic Data Extraction Using Tag and Value Similarity Based on Structural -Semantic Entropy, *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4,* April 2013.
[10] K .Devika, and S. Surendran, An Overview of Web Data Extraction Techniques, *International Journal of Scientific Engineering and Technology (ISSN : 2277-1581),Volume 2 Issue 4,* pp : 278-287, 1 April 2013.
[11] V.K. Deepak, and N. V. R. Kumar, Retrieve Records from Web Database Using Data, *International Journal of Computer Science and Information Technologies, Vol. 5 (2)* , 2014, 1552-1554.
[12] Y. Zhai, and R. G. B. Liu, Mining Data Records in Web Pages, *ACM Conference,00,Month 1-2, 2000*, City, State.
[13] Y. ZHAI , *STRUCTURED DATA EXTRACTION FROM THE WEB*, the degree of Doctor of Philosophy in Computer Science in the Graduate College of the University of Illinois at Chicago, 2005, Chicago, Illinois.
[14] R. Rapp, and S. Sharoff, Extracting Multiword Translations from Aligned Comparable Documents, *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)* @ EACL 2014, pages 87–95, Gothenburg, Sweden, April 27, 2014.
[15] Y. Zhai, and B. Liu, Web Data Extraction Based on Partial Tree Alignment, *the International World Wide Web Conference Committee (IW3C2),* WWW 2005, May 10-14, 2005, Chiba, Japan.
[16] Y. Zhai and B. Liu, Automatic Wrapper Generation Using Tree Matching and Partial Tree Alignment, *American Association for Artificial Intelligence (www.aaai.org).*