*International Journal of Latest Trends in Engineering and Technology*
*Special Issue SACAIM 2016, pp. 216-220*
*e-ISSN:2278-621X*

# GENERATING DATA USING TREE PATTERN MATCHING-A SURVEY

Nagarathna[1], Nikhila Yuvaraj[2] and Ms. Suchetha VijayaKumar[3]

**Abstract – Data extraction is the process of retrieving data from thedata sources for data processing or data storage. Automatic wrapper is used to extract data records from search engine results pages which contain important information for computers users and also focuses on HTML tag based data extraction. These techniques consist of a series of data filter techniques to remove irrelevant data from the webpages. The algorithms can be used to check the similarity of data records and to detect and extract the correct data region based on their component sizes. Experimental and deletion test can be used to evaluate the performance of our algorithms. The Experimental test describes the performance of the existing state of the wrapper tools such as ViNT and DEPTA. Deletion test replaces the novel techniques with state of the art conventional techniques to indicate that the wrapper design can extract data records from Search engine results pages easily.**

**KEYWORDS: Information Extraction, HTML Tag, Automatic Wrapper, Search Engine, Tree Matching Algorithm**

## I. INTRODUCTION

The data extraction from anobject source is Information extraction. The target source can be a natural language source or structured records. To check the similarity of data records by comparing the position and identify of each node in the trees to remove data which are not related with different structure, Non visual wrappers use Tree matching algorithm [2].

The web data extraction using the HTML tag is a huge process and consists of less errors. The working of a Tree matching algorithm is it checks of two tree structures and it analyses whether the first tree is transformed into the second tree. The samples of web data extraction are 1) Extracting the list of the competitors price list from a web page that stays consistently ahead in the competition. 2) The transfer of data from one web page to another application.3) The data is stored in the database by extracting the information of the user from the web page. The results in the search engines are retrieved from the HTML pages which consists of the extracted results from the user’s data, it is done through an automated tool known as Wrapper.Database are generated by predefined templates [5].

Tree matching algorithm is based on the frequency measures of a tree structure as apart of the filtering stages to check the similarity of data records. This algorithm doesn’t match two tree structures and find their similarity by checking the identify of each node, but uses the number of nodes in a tree to determine the similarity of two trees.

The steps following Introduction are Existing Work, Proposed Algorithm, Conclusion, and References.

## II. EXISTING WORK

[1] *Department of Information Technology AIMIT (St. Aloysius College),Mangalore, Karnataka, India*
[2] *Department of Information Technology AIMIT (St. Aloysius College),Mangalore, Karnataka, India*
[3] *Department of Information Technology AIMIT (St. Aloysius College),Mangalore, Karnataka, India*

In existing system, various tools are used to extract the similar data such as

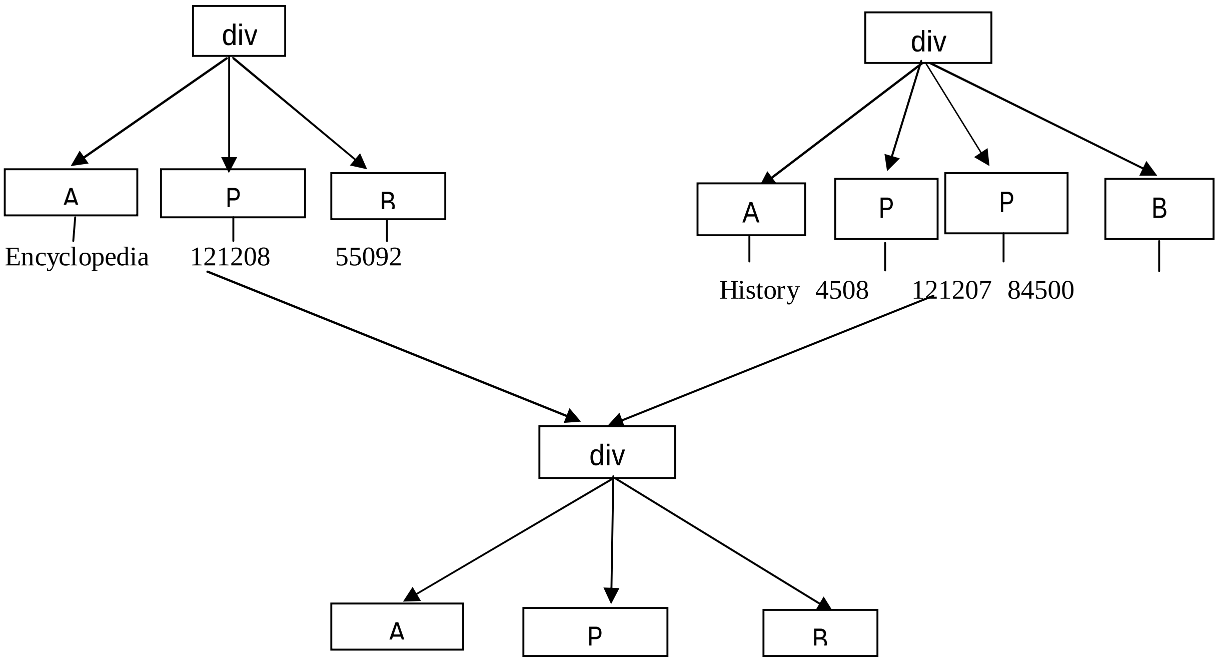
- Data Extraction Partial Tree Alignment(DEPTA),
- ViNT((Visual Informationand Tag Structure Based Wrapper Generator),
- DeLa(Data Extraction and Label Assignment for Web Databases),
- ViPER(Visual Perception based Extraction of Record),
- NET(Nested data Extraction using Tree matching and Visual cues)

### *A. DEPTA*

To check the similarity of the structure of records DEPTA uses String Edit Distance and Tree Edit Distance techniques.

In String Edit Distance algorithms it matches two strings and checks whether the first string is transformed into the second string. These algorithms are fast and run in complexity of O(m), where m is the number of tags in a data records**.** There are several variants of String Edit Distance Algorithm such as Hamming Distance, Episode Distance and Longest common subsequence Distance.

Tree Edit Distance techniques uses two tree structures and matches them by comparing the node identity and the position. Types of Tree Edit Distance Algorithms are, Tree Edit Distance, Alignment Distance, Isolated-Sub Tree Distance, Top Down Distance and Bottom Up Distance. Tree matching algorithm contains tree nodes matching which is quite similar to String Edit Distance Algorithm. DEPTA uses Bottom Up Tree matching algorithm to match tree structures of data records. DEPTA's tree matching algorithm consists of identifying the nodes and finding the topmost matches between two trees by estimating their location [5].



### B. *ViNT*

This tool is used for producing wrapper automatically that extracts the search results from the dynamically generated HTML result pages which is returned by any search engine. The visual data similarity is used instead of tag structure to identify the data. This techniques is used for producing the wrappers for the search engine. One of the disadvantage of ViNT is that there will be changes in the format of the query result page, it usually fails in the pre-learned wrapper. The main usage of ViNT is to monitor the query result pages to monitor continuously to format the changes which are the most difficult problem [5].

### C. *DeLa*

DeLa, the data objects that are extracted from the retrieved web pages by sending the queries through the HTML forms. It fits the generated data into a table and the labels are allocated to the attributes of the data objects that is the columns of the table. DeLa consists of 4 components.

- Web Crawler
- Wrapper Generator
- Data Alignment
- Label Assigner

Web Crawler identifies the data and gathers the labels of the website form elements and sent queries to the website. Wrapper Generator extracts regular expression from the data contained in webpages. Then new data structure is aligned to record the data extracted by the wrappers to generate algorithm to fill the data table. Label assigner explores the feasibility of rule based automatic data annotation for web databases and validates the effectiveness of the proposed rules [9].

*D. ViPER*

It is one of the fully automatic information extraction tool. Extracting and differentiating the relevance of repetitive information contents with respect to the user's visual perception of the webpage is the major function of ViPER. The relevant data is aligned using Multiple Sequence Alignment. ViPER uses both visual data value similarity features and the HTML tag structureto find the potential repetitive patterns [1].

*E. NET*

NET is an effective method used where data is extracted from web pages. It is based on the tree edit distance method. There are two stages, first stage is to build the tag tree of the page and in second stage data records are identified and extracted from them. One of the advantage of using this tool is that it is useful in enabling accurate alignment and the extraction of both flat as well as nested data records where as it doesn't identify correctly whether a flat structure is nested one [5].

## III. PROPOSED ALGORITHM

### *3.1 OVERVIEW*

The Proposed algorithm is Tree Wrap algorithm. The pages used to extract data should be obtained from the query given in search engine. There should be at least three data records in these sample pages. This algorithm doesn't require conversion from HTML page into XHTML because the parser can recognize HTML. There are two components in this algorithm. In first component, the HTML page is parsed and organized into DOM tree representation. In second component, we use Dummy tree matching algorithm and scoring function for extracting the data records.

### *3.2 COMPONENTS OF TREE WRAP*

*1. Parser for Tree Wrap*

To parse the webpage, search engine result page is required as the input. HTML parser will divide the pages into tokens by reading them. There are two types of tokens, HTML command tags and text tokens. Once the pages are parsed it is then stored and arranged the contents in a DOM tree which is then used for the further processing.

2. *Extracting the data at record level*

*2.1 Breadth First Search Technique*

Breadth First Search technique is used to detect and label the data regions which are identified by traversing the parsed webpages through DOM tree. Data regions are the set of data records. A data records can contain repetitive sequence of HTML tags and are located at the same level of DOM tree.

2.2 *Filtering Stage*

Tree wrap will have the lists of data records after the completion of the BFS technique. There are four stages of filtering.

*2.2.1HTML Tag Structures*

HTML Tag Structures are used to remove the data records that have less HTML tags which are obtained from the BFS Extraction.

*2.2.2 Similarity*

Dummy Tree Algorithm is used to check the similarity between the data records

ALGORITHM

```
For (int i:1 to no.ofDATARECORDS){
    //n nodes in the data records
    //total no of distinct tags(Step 1)
    int firstDistinctTags=get the number of distinct tags
        (record(i));
    int secondDistinctTags=get the number of distinct tags
        (record(i+1));
    //comparing the left and right tree
    if(abs(firstDistinctTags-secondDistinctTags)>1){
      //remove the record
      record(i)
      //if not similar delete the left tree
    }//endif
    //Calculating the total no of distinct tags in every level
    int AllLevelFirstDistinctTags=getNoAllLevelDistinctTags(record(i));
    int AllLevelSecondDistinctTags=getNoAllLevelDistinctTags(record(i+1));
    //comparing the left and right tree
    if(abs(AllLevelFirstDistinctTags-AllLevelSecondDistinctTags)>1){
      //remove the record
      record(i)
      //if not similar delete the left tree
    }
  }
```

There are two stages of screening procedure to check the similarity of group of trees. When a number of trees are given, this algorithm first examines the distinct tags of first as well as second tree. If the distinct tags occur concurrently in both the trees, then the similarity test of the first stage will be in favor of the tree. Then we calculate total number of distinct tags available of all levels of first and the second tree is done in Stage two. If the number of these distinct tags are almost equal, then two trees are considered equal.

The two trees are said to be similar only if they meet the screening procedure of both the stages. If first two trees are similar, then first tree is retained and then next tree is compared with the third tree of the group in order to check their similarity. If both trees are not similar, the first tree will be removed and the second tree will be compared with the third tree. These steps will be repeated until the last tree is used for similarity checking. The filtering stage in Dummy Tree Matching algorithm is used to detect the data regions that contains structurally similar data records that are normally exists in the result pages of search engine.

*2.2.3Number of Nodes*

In this stage, irrelevant data records are filtered out for further processing.

*4. Scoring Function*

Once the irrelevant data are removed, only one data region is chosen from the list of available data records. This is the important component of data extraction because in order to differentiate the correct data region form incorrect data region we need a good scoring function.

The value of scoring function is calculated as following:

Score(x) = (a*150) + ((b + (c*15) + (d*50)) *5 )

Where a denotes number of Parents Nodes Level

b denotes Total Text Length

c denotes Number of Images

d denotes Number of Separator Tags

x denotes Data Region

## IV. CONCLUSION

Most of the tools to extract the data from the webpages are inaccurate. Tree wrap algorithm is used to extract the data records from the web pages. Here filtering methods and dummy tree algorithm is used that simplifies the process of comparing each node of the trees to check their similarity. Further the algorithm can be implemented and tested.

## REFERENCES

[1] Pranali Nikam, Yogita Gote, Vidhya Ghogare, Jyothi Rapalli, ”Web Data Extraction and Alignment Tools: A Survey”**,** 2015.

[2] Jer Lang Hong, Fariza Fauzi,” Tree Wrap-data Extraction Using Tree Matching Algorithm”, 2010.

[3] Jiying Wang, Fred H. Lochovsky,”Data Extraction and Label Assignment for Web Databases”**,** 2003.

[4] J. Wang and F. Lochovsky, “Data-rich section extraction from HTML page”Proc.3rd Conf. on Web Information Systems Engineering, 2002, 313-322.

[5] Appukutti Chandrashekhar, Dr.P.Venkata Subba Reddy,”HTML Tag Based Web Data Extraction and Tree Merging From Template Page”**,** 2014.

[6] Vidya V.L,”A Survey of Web Data Extraction Techniques”**,** 2014.

[7] Ivarez M, Pan A, Raposo J, ”Extracting List of Data records from Semi Structured Web Pages”, 2008.

[8] Hong J.L, Siew E and Egerton S,”DTM, Extracting Data Records from Search Engine Results Pages using Tree Matching Algorithm ”**,** 2010.

[9] Alberto H.F.Laender, Berthier A.Riberio-Neto,”A Brief survey of Web data Extraction Tools”**,** 2002.

[10] Zhai Y, Liu B,”Web Data Extraction Based on Partial Tree Alignment”**,** 2005.