Template Extraction from Heterogeneous Web Pages

[1]Mrs. Harshal H. Kulkarni, [2]Mrs. Manasi k. Kulkarni

Asst. Professor, Pune University, (PESMCOE, Pune), Pune, India

Abstract: **Templates are used by many websites for increasing the productivity of publishing the Web pages. Common templates are used with its contents. The templates provide users to easily access the contents due to its consistent structures. However, for machines, the templates are considered harmful because of its irrelevant terms in the template, so it will degrade the performance of web applications. Thus, template detection and extraction techniques have received a lot of attention to improve the performance of search engines, web application, clustering and classification of web documents. The objective of this paper is to cluster the web documents based on the similarity of underlying template structures in the documents so that the templates for each cluster are extracted simultaneously. While extracting the templates, here consider the web page structure along with its contents. To effectively manage an unknown number of clusters (templates) Minimum Description Length (MDL) Principle is use and MinHash technique to estimate the MDL cost quickly. So that it will form a qualified cluster.**

Keywords: **Template Extraction, HTML Documents, MDL Principle, TEXT-MDL, TEXT-MAX.**

I. INTRODUCTION

As we know, Templates have different meanings depending on the application. For web application the templates are nothing but the structural information of the web pages. In web applications templates are nothing but Master page of that web application. Normally A web template has standards and conditions for their common elements of the web pages. The common elements of these template (i.e. master pages) pages are going to focus on their linked page. Examples of these elements are main page of Yahoo, any government web sites, any college or school web sites etc. [8] [9]. The following fig.1.1 shows a simple example of the school web template. In this example, the website contains 4 central areas of information. (1) Menu system, (2) News,(3) Miscellaneous information about the school and (4) Information of the Education Agency [8].

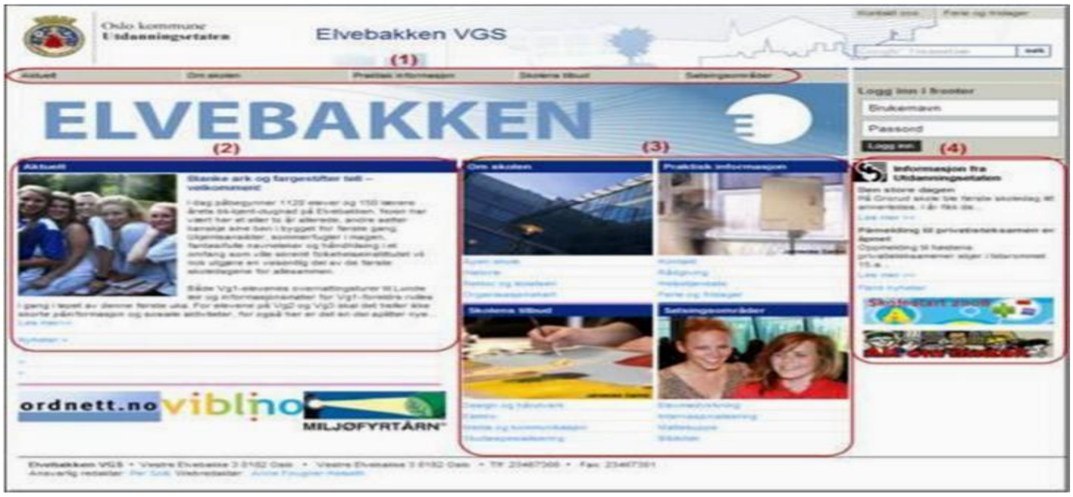

Fig.1.1 shows a simple example of the school web template.

As we know the main purpose of internet is to publish or to access the various type of information with the help of web sites. To achieve the high productivity of the web pages, the web pages are published on many web sites with its common template. The templates provide users easy access to the information with its common contents and consistent web page structures.

When unknown templates are generated due to their irrelevant term then, they look harmful for machines because they degrade the accuracy and the performance of the system. Thus template detection and extraction techniques have received a lot of attention to improve the performance of the web applications such as data integration, search engines, classification of the web documents and so on [1][2][3]. The focuses of the paper is on the problem of extracting and detecting the templates from heterogeneous web documents and for that purpose here propose an algorithm. This algorithm manages an unknown number of the templates, to improve the efficiency and scalability of template detection and extraction, and also generate the qualified clusters.

This paper represents a web document as a set of paths and a template as a set of paths in a Document object model (DOM) tree [5]. To deal with the unknown number of templates and select good partitioning from all possible partitions of web documents, here use Rissanens Minimum Description Length (MDL) principle [4]. The model of each cluster is the template itself of the web documents belonging to the cluster. Thus, there is no need to process an additional template extraction after clustering web pages. The MDL cost is nothing but the number of bits required to describe data with a model and the model is the description of clusters represented by templates. Clustering is ranked according to two methods,

a) The number of bits required to describe a clustering model.

b) Partitioning with the minimum number of bits which is selected as the best one.

In order to improve efficiency and scalability for handling a large number of web documents for clustering, MinHash technique is going to be introduced [7]. In that, Jaccord Coefficient technique is going to be used. The clustering of web documents is performed such that the documents in the same group belong to the same template and thus, the correctness of extracted templates depends on the quality of clustering.

2. LITERATURE SURVEY

Before going to start the actual work of the paper here is a literature review of this paper.

- **Paper Discussion:**

1. Arasu and H. Garcia-Molina, “Extracting Structured Data from Web Pages”.

The goal is to solve the extraction problem for real pages that produces the values that a human would consider semantically correct. The observation is that, for real pages, the words that are part of the template have a high correlation of occurrence in input pages with other words in the template. The words that are part of encoded data, on the other hand, do not have high occurrence correlation with other words. This observation can be used to design an algorithm for deducing the template from input pages. The full version of the paper contains formal definition of high occurrence correlation and algorithm. This paper gives an idea how to generate the template with the help of words in web pages.

2. Chulyun Kim and Kyuseok Shim,” TEXT: Automatic Template Extraction from Heterogeneous Web Pages.”

This is the base paper referred by me for my dissertation work. This paper gives a brief idea, how to extract and detect the template from heterogeneous web pages. For that they generate the novel algorithms such as TEXT MDL, TEXT-HASH, and TEXT-MAX, which refer in this paper. The main objective of this paper is to find the unknown number of the template to improve the performance, scalability and efficiency of the template detection and extraction of the algorithms. For that they represent the web document and template as a set of path in DOM Structure and apply the MDL principle to manage the unknown number of the template. After that a MinHash technique is going to be used to estimate the MDL cost quickly so that a large number of the document can be processed.

3. F. Pan, X. Zhang, and W. Wang, Crd: "Fast CoClustering on Large Data Sets Utilizing Sampling Based Matrix Decomposition".

This paper introduces a Co-Clustering algorithm. Co clustering algorithm simultaneously clusters both columns and rows of the data matrix. Co-clustering takes advantage of the duality between rows and columns to effectively deal with the high dimensional data. It has successful applications in gene expression data analysis and text mining. The existing algorithm requires, the whole data matrix which is stored in the main memory. To address these limitations of existing work, this paper proposes a general co-clustering framework, Co-clustering random Dataset (CRD**),** for large datasets. This framework is based on sampling based matrix decomposition method. And this matrix representation method is used in my paper for clustering web pages.

4. Jorma Rissanen, "Fisher Information, Stochastic Complexity and Universal Modeling"

The main problem in the implementation of the principle is how to estimate the shortest code length for the data, given a suggested model class. This can be difficult requiring ingenuity and hard work if the class of models is complex. In this paper they are going to introduce MDL Principle for complexity model class which built up of simpler ones, so that code length can be composed of the stochastic complexities of the component, and this again makes a formula useful in my paper to decrease the length of the code.

5. M. de Castro Reis, P.B. Golgher, A.S. da Silva, and A.H.F. Laender," Automatic Web News Extraction Using Tree Edit Distance".

This paper presents a domain-oriented approach to Web data extraction and its application to automatically extracting news from Web sites. This approach is based on the concept of tree-edit distance and allows not only the extraction of relevant text passages from the pages of a given Web site, but also the fetching of the entire Web site content, the identification of the pages of interest and the extraction of the relevant text passages discarding no useful material such as banners, menus and links. This paper introduced a Tree Edit Distance method and RTDM algorithm.

This paper offers an alternative and uniform solution for three important problems in automatic Web data extraction: structure based page classification, extraction and data labeling. And the RTDM algorithm is flexible with the complex derivations of the problem.

6. Sruthi Kamban, K.S, M.Sindhuja," Extraction of html document from heterogeneous web pages for clustering Tech".

In this paper the web pages are going to be clustered with the help of URL. This paper uses the DOM clustering method for clustering the web documents. Cluster techniques are used in clustering those templates as the single template. Novel algorithms are used for extracting templates from a large number of web documents which are generated from heterogeneous templates.

7. Valter Crescenzi, Paolo Merialdo, Paolo Missier," Clustering Web Pages Based on Their Structure".

This paper focused on document clustering without template extraction, they crawled web pages from different sites and compare with their links. For comparison they are going to develop their own MinHash Technique

3. PROPOSED WORK

The concept of this paper is based on the base paper [2]. The concept of detecting and extracting the template from heterogeneous web document is going to be proposed in "Template Extraction from Heterogeneous Web Pages" paper. The Goal of the paper is to generate the template from heterogeneous web pages with its common contents and also to give a qualified cluster with a common content of the template.

- In this system the HTML Parser is used for parsing the web documents and generating Paths. After parsing, Support values for distinct paths are calculated. By using paths and tokens Matrix representation of document set is carried out.
- MDL Principle will be useful to generate the clusters with minimum MDL Cost. And this Cluster will be represented as Template.
- Text-MDL algorithm uses Min-Hash technique. It will reduce search space by merging the clusters.

- Text-MAX algorithm uses Min-Hash technique. It will reduce search space by merging the clusters having same Min-Hash signature.

- **Interpret inputs for Project:**

Input for this system is web document set. The Web documents can be, .html or .html file

Here use a HTML parser which parse the only HTML web documents and generate the Paths and Tokens for each document. By using paths and tokens, Threshold values and essential paths for documents are calculated which will useful for the algorithm calculation.

- **Input: Web Documents*:***

The set of web documents with extension .html is the input to the system which is shown in fig.3.1.Simple Web Documents (a) Document *d*1 (b) Document *d*2(c) Document *d*3 (d) Document *d*4

```
<html>            <html>            <html>           <html>

<body>            <body>            <body>           <body>

<h1>Tech</h1>  <h1>World</h1>   <h1>Local</h1>     List

<br>              <br>              <br>             </body>

</body>           List              List             </html>

</html>           </body>           </body>

                  </html>           </html>

(a)               (b)               (c)              (d)
```

Fig.3.1. Simple Web Documents (a) Document *d*1 (b) Document *d*2 (c) Document *d*3 (d) Document *d*4

Here consider the first document d2 and from that design the DOM Model for the first document,

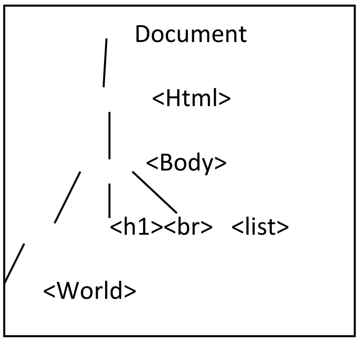

Fig. 3.2 DOM tree for d2.

- **Steps to carry out the Project work**

To increase the performance of the system by avoiding traditional way of visiting different web pages for the similar contents. The system combines the contents of heterogeneous web pages into templates according to structure similarity with contents. So that required information can be retrieved with minimum search space. Objective of the system is to generate the qualified clusters. For that there is need to generate the template from heterogeneous web pages by following steps.

1. Parsing the Web documents using HTML Parser.

To distinguish between Html tags and contents the input web documents are parsed with Html parser. Since an HTML document can be naturally represented with a Document Object Model (DOM) tree, web documents are considered as trees and many existing similarity measures for trees have been investigated for clustering.

2. Generating Distinct Paths & Support Values for Each Document.

Consider the example documents in Fig.3.1. To extract the template, documents *d*1**,** *d*2**,** *d*3**,** *d*4 has to be represented in DOM tree. Set of distinct paths for all documents are represented in Table 3.1.First column in Table. 3.1 State the path ID i.e. the serial number assigned to particular token. Second Column represent the path of node in DOM tree as we have calculated earlier with the help of fig 3.1 & fig 3.2.Last column is the support value i.e. Number of Documents that contain the particular path, means consider a path "Document\<Html>" this path is occurred in all four documents therefore support value becomes 4 same we can calculate the remaining support values.

Table . 3.1 Paths of Tokens and Their Support

ID	Path	Support
P_1	Document\<html>	4
P_2	Document\<html>\<body>	4
P_3	Document\<html>\<body>\<h1>	3
P_4	Document\<html>\<body>\ 	3
P_5	Document\<html>\<body>\<List>	3
P_6	Document\<html>\<body>\<h1>\Tech	1
P_7	Document\<html>\<body>\<h1>\World	1
P_8	Document\<html>\<body>\<h1>\Local	1

3. Calculating Threshold values for each document in document set.

Threshold value for individual document is calculated on the basis of support values of the paths. Threshold value is calculated on MODE (i.e. Most Frequent Support Value) of support values. Essential paths are the paths of documents who's support value is at least or greater than threshold of that document

4. Matrix representation of document set and paths.

- Essential path Matrix(M_E)

Row represents Path Number and Column represents Document Number.

$$M_{E=} \quad \begin{array}{ccccc} & d_1 & d2 & d3 & d_4 \\ P_1 & 1 & 1 & 1 & 1 \\ P_2 & 1 & 1 & 1 & 1 \\ P_3 & 1 & 1 & 1 & 0 \\ P_4 & 1 & 1 & 1 & 0 \\ P_5 & 0 & 1 & 1 & 0 \\ P_6 & 0 & 0 & 0 & 0 \\ P_7 & 0 & 0 & 0 & 0 \\ P_8 & 0 & 0 & 0 & 0 \end{array}$$

- Template path Matrix(M_T)

Row represents Path Number and Column represents Cluster Number.

$$M_T = \begin{matrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}$$

- Document Matrix(M_D)

Row represents Cluster Number and Column represents Document Number.

$$M_D = \begin{matrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}$$

- Deference Matrix(MΔ)

Row represents Path Number and Column represents Cluster Number.

$$M_\Delta = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}$$

5. Apply minimum description length (MDL) principle.

The MDL principle states that the best model inferred from a given set of data is the one which minimizes the sum of 1) the length of the model, in bits, and 2) the length of encoding of the data, in bits, when described with the help of the model. The MDL costs of a clustering model C and a matrix M are denoted as $L(C)$ and $L\,(M)$, respectively. Considering the values in a matrix as a random variable X, p_r (1) and p_r (−1) are the probabilities of 1 s and -1 s in the matrix and *pr* (0) is that of zeros. Then, the entropy H(X) of the random variable X,

$H(X) = \sum -p_r(x)\log_2 p_r(x)$ and

$x \in \{1,0,-1\}$

$L(M) = M.H(X)$

The MDL cost of clustering Model C is calculated as,

$L(C) = L\,(M_T) + L\,(M_D) + L\,(M_\Delta)$

6. Calculating MDL cost with MinHash technique. Then implement algorithms such as Text-MDL & Text-Max for finding the best pair of clusters calculating MDL Cost for generating the qualified clusters.

- Calculating signature of initial cluster set (document) Merging of clusters which are having same signature.

- Calculating MDL Cost of each cluster pair. And selecting the cluster pair having Minimum MDL Cost as Init best pair.

- Finding succussesive best pair by calculating nearest cluster to the current best Pair with minimum MDL Cost.

○ ***Requirements:***

To implement the algorithms the following software and hardware are required.

Software Requirement

 Windows XP or Windows 7.

 Java Language (JDK Tool 1.3 onwards).

Hardware Requirement

 Pentium-Core2 or later version Processor, Hard disk space: 20GB.

 2GB of RAM

4. RESULT AND DISCUSSION

The aim of the project is to extract a template from heterogeneous web pages. Different extracting algorithms are implemented like TEXT-MDL, TEXT-MAX. Result analysis is carried out in terms of time required to extract template and number of compact clusters are generated by previous implemented algorithms. After successfully execution of Text-MDL & Text-Max, here conclude that Text-MDL generate better clustering output as compare to Text-MAX, and there is slit difference in execution of time.

1) The output of Text-MDL is,

<html> <body> <h1> </h1> List </body> </html>	<html> <body> </body> </html>
Template1	Tempalte2

Fig.4.1. out Put of TEXT_MDL

2) The Output of Text-Max,

<html> <body> <h1> </h1> </body> </html>	<html> <body> </body> </html>
Tempalte0	Template1

Fig.4.2. out Put of TEXT_MAX

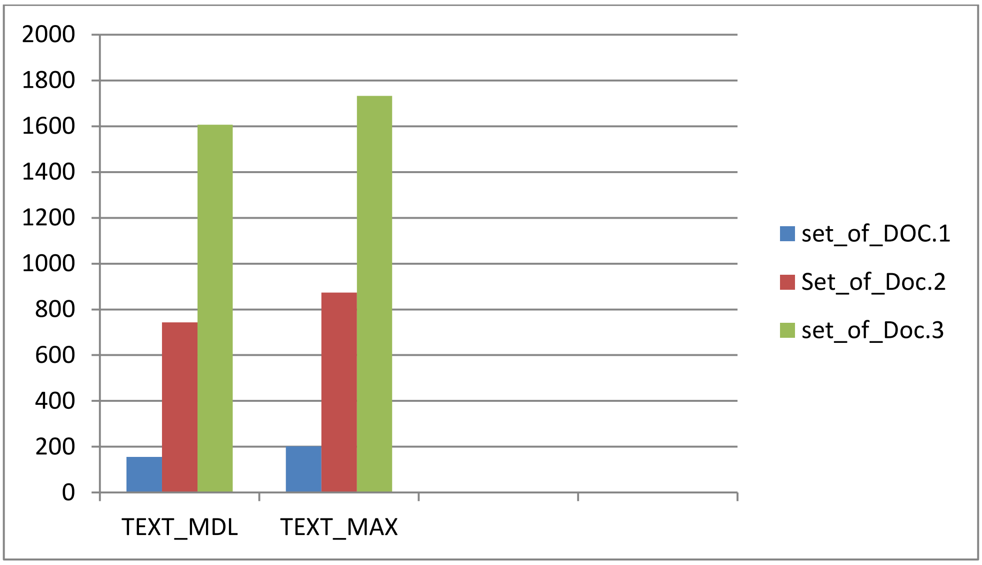

Fig.4.3. Bar chart for Time Analysis

5. CONCLUSION

Automatic template extraction system combines the contents of heterogeneous web pages into templates according to structure similarity of documents with its contents. So that required information can be retrieved with minimum search space. Extraction is depending on the underlying structure of web documents. While extracting the template from collection of documents this technique not only considered the tags available in web designing but also considers contents as a part of the template along with tags. Web documents are represented into DOM structure to simplify the further clustering calculation. Employed MDL principle manages the unknown number of clusters.

A . FUTURE SCOPE:

The two basic directions are encountered for future work.

- More attention should be given on design and structure of templates such as considering attributes of tags, images etc.
- Explore the simplicity, efficiency and effectiveness of Dice-Coefficient approach with some standard data set.

REFERENCES

[1] Arasu and H. Garcia-Molina, “Extracting Structured Data from Web Pages”, Proc. ACM SIGMOD, 2003..

[2] Chulyun Kim and Kyuseok Shim,” TEXT: Automatic Template Extraction from Heterogeneous Web Pages”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011.

[3] F. Pan, X. Zhang, and W. Wang, Crd: “Fast CoClustering on Large Data Sets Utilizing Sampling-Based Matrix Decomposition”.Proc.ACMSIGMOD,2008.

[4] Jorma Rissanen, "Fisher Information, Stochastic Complexity and Universal Modeling” ,IBM Research Division, Almaden research center , 650 Harry Road, San Jose Ca 95120-6099.

[5] de Castro Reis, P.B. Golgher, A.S. da Silva, and A.H.F. Laender,” Automatic Web News Extraction Using Tree Edit Distance”, proc.13th Int’1 conf. World wide web (www), 2004

[6]]Sruthi Kamban, K.S, M.Sindhuja,” Extraction of html document from heterogeneous web pages for clustering Tech”, International Journal Of Engineering and Computer Science ISSN:2319-7242,4 April 2013.

[7] Valter Crescenzi, Paolo Merialdo, Paolo Missier,” Clustering Web Pages Based on Their Structure”, Data and knowledgeEng, Vol.54, pp.279, 299, 2005.