Design and Development of Efficient Methodology for Template Detection and Extraction from Heterogeneous Web Pages

[1]Harish Kumar
[1]Ph.D CSE Scholar, Glocal University, Saharanpur, India

[2]Dr. Praveen Kumar
[2]Supervisor, Glocal University, Saharanpur, India

[3]Dr. Kamal Upreti
[3]Co-Supervisor, Inderprastha Engineering College, Ghaziabad, India

Abstract – In recent years, many researchers have tried to enhance the performance of template detection and extraction methodology due to increase the performance of web applications. In web applications templates are created for many web sites to increase the productivity and search time of web pages. The templates have consistent structure so that users can easily access the information on the web sites. Templates will be utilized for characterizing auxiliary data of various zones, for example, web applications zones, biometric regions, computerized system, Programming Languages. Template extraction from heterogeneous web pages can be done by constructing Document Object Model (DOM) tree of HTML document and finding essential paths of document. But due to large variety of web documents, there is a need to manage unknown number of templates. In this paper, authors present a novel approach for detection and extracting templates from a large number of web applications by using classification techniques.

Keywords– **Template Extraction, DOM Tree, Classification, Web Application**

I.Introduction

Now days, template detection and extraction have been a lot of attention to improve the performance of web application and such as data integration, search engines, classification of documents etc. A good template extraction methodology can significantly improve the performance of such application. Templates can be used to effective separation between presentation logic and business logic. The unknown templates are considered harmful because they degrade accuracy and performance due to irrelevant terms in template. Web pages in the Websites are constructed in such a way that almost 50% of the data contains templates. Web readers easily access to contents by using templates. Templates are a foundation on which actual content is built. The common elements of template pages are going to focus on their linked page. Examples of these elements are any education related web sites, any business related web site, any online shopping web sites etc. [3, 4].

Data mining is the process of extracting patterns from data. It is the process of discovering knowledge from large amounts of data stored either in databases or warehouses [1]. Many techniques are used in data mining to extract patterns from large amount of database [2]. Classification is a data mining (machine learning) technique used to predict group membership for data instances.

Paper is organized into different sections as follows: Section-II covers literature review, section-III presents a background of web template, presentation of template and web documents, section-IV discusses about various methods of template detection and extraction, section-V describes flow diagram of the proposed methodology and section-VI shows the conclusion while references are mentioned at the last.

II.Literature Review

In recent years, many researchers have tried to improve performance of web template detection methodology. Template detection improves the performance of web application. Some of the works are summarized here.

Vieira *et al.*[5] says templates are detected by finding identical nodes in Document Object Model (DOM) trees and sub trees by performing mapping between the tree structures of web pages.

Crescenzi *et al.*[6] also assumes that every HTML tag is generated by the template. It extracts template by analyzing a pair of web pages based on Matching technique called ACME (Align, Collapse under Mismatch, Extraction). Roadrunner does not rely on any prior knowledge about the about page contents.

Roadrunner-Towards Automatic Data Extraction from Large Web Sites [6], introduced extracting data from HTML sites through the use of automatically generated wrappers. This paper develops a novel technique to compare HTML pages and generate a wrapper based on similarity and differences. Goal is automatic generation of wrapper that is without any prior knowledge of target pages and human interaction. Matching technique is used to compare the HTML codes of two pages and to infer a common structure and a wrapper.

R. Henzinger *et al.*[7] says the templates present in web pages are considered to be great challenge in search engine since, it degrades the performance of the search engines.

Gibson, K. Punera, and A. Tomkins [9] proposed the volume and evolution of web page templates where templates represent 40-50% of the total bytes on the web, and this fraction continues to grow at a rate of approximately 6% per year.

The "Template-Extraction" problem is based on Data Mining Concept. Initially template extraction problem is studied by Yossef & Rajagopalan [3]. In that, they use html tags of the web pages for template extraction system. But they analyzed that the word is also part of web pages so it can be used for template extraction.

Arasu and H. Garcia-Molina [10] proposed paper on extracting structured data from web pages which describes the automatic extraction of database values from templates without any human input.

Tak-Lam Wong and Wai Lam [11] proposed an unsupervised learning framework which can jointly extract information and conduct feature mining from a set of Web pages across different sites. Important characteristic of this model is that it allows tight interactions between the tasks of information extraction and feature mining.

Xiangwen Ji et al. [12] devised a new method based on Tag tree template. Web pages from different Web sites are parsed into Tag trees, and then templates of each site are generated from the trees by using a cost-based tree similarity measurement. The exclusive content in each page is then extracted by using the templates to parse the page.

Gilles Nachouki, Mohamed Quafafou [13] explained the process for mashing heterogeneous data sources based on the Multi-data source Fusion Approach (MFA). The aim of MFA is to facilitate the fusion of heterogeneous data sources in dynamic contexts such as the Web.

Hua Wang, Yang Zhang [14] presented a Web data extraction method based on simple tree matching by analyzing the structure and content of Web documents.

Jellouli I., Mohajir M.E [15] devised an approach that does not make any prior assumption on the design and the format of web pages, it is totally independent and it is able to achieve semantic extraction from a single web page with a single instance.

Hao Han, Tokuda T [16] proposed a method for Web information extraction to generate the virtual Web service functions from Web applications at client side. They show that the general Web applications can be also integrated easily.

III.Web Template– A Background

A template is set of common layouts and format features that appear in a set of HTML pages that is produced by a single program or script that dynamically generates the HTML page content. Template of a document cluster is a

set of paths which commonly appear in the documents of the cluster. Following figure-1 shows a presentation of web template below.

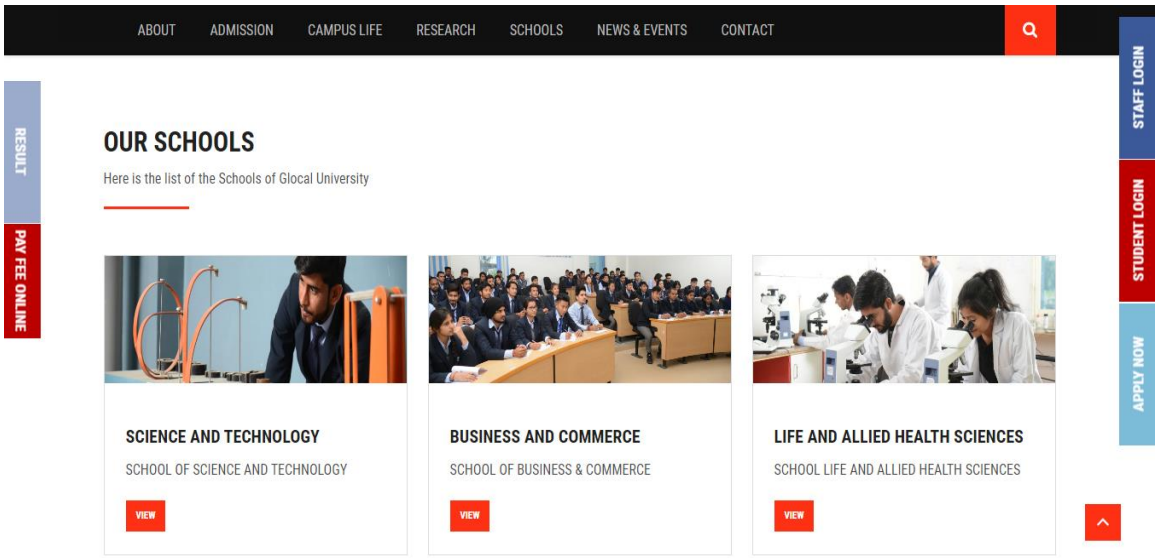

Fig.1. Presentation of Web Template

Web page usually contains various contents such as navigation, decoration, interaction and contact information, which are not related to the topic of the web-page. Furthermore, a web page often contains multiple topics that are not necessarily relevant to each other. Therefore, detecting the semantic content structure of a web page could potentially improve the performance of web information retrieval. The templates stored in database can be further used by the web designer to develop Web pages which does not corrupt the performance of the search engines. We can access this data (information) with the help of web sites in terms of web pages. To access this huge data from the web sites we need to achieve the efficient access and less search time 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. It also enables the web designer to develop web pages fast and easy.

3.1 Representation of Web Document

The sample representation of a web document is shown in figure-2.

```
<html>
<head>
<title> Glocal University </title>
</head>
<body>
<p> glocaluniversity.edu.in </p>
</body>
</html>
```

Fig.2. Sample Presentation of Web Document

3.2 DOM Tree Construction

Document Object Model (DOM) tree is used to represent HTML documents, web documents, and web pages are considered as trees. It is an application Programming Interface (API) for valid HTML documents. It defines the

logical structure of web documents. Document Object Model (DOM) is platform and language neutral interface that would allow scripts to dynamically access and update the content, structure, and style of documents. It splits a tree into many small sub trees and nodes, DOM tree of web documents, find out two types of paths content paths and template paths. This approach is based on highly efficient tree structure analysis. Grouping of pages is done to form cluster and find generic representation of structure of pages within a clusters. Structure of web page could be described by a tree. Tree-edit distance has been used to evaluate the structural similarity between web pages.

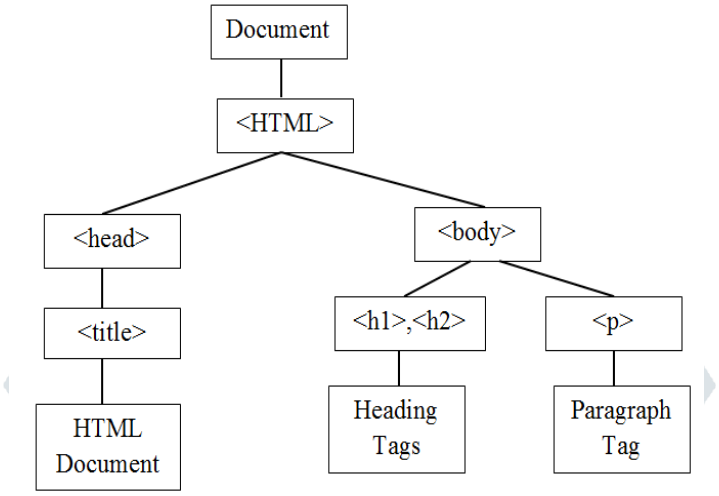

Fig.3. Construction Process of DOM Tree

Input: HTML document are extracted from different Web pages which is taken for preprocessing. In the html document the text information and html tags are spitted separately. The separated HTML tags have been constructed into HTML DOM tree and have been investigated for clustering. Then the path is discovered by the DOM model and also used to calculate the number of support values in the individual tags.

IV.Template Detection & Extraction Methods

The problems of template detection and web page cleaning have received considerable attention in the recent literatures.

The extraction method proposes two algorithms known as Content Extractor and Feature Extractor.

(i) Content Extractor Algorithm-It distinguishes between content blocks and non-content blocks based on repetition of same blocks in several pages.

(ii) Feature Extractor Algorithm- It extracts contents based on features supplied externally. The algorithms deal with web pages having similar underlying template structure.

Template detection and extraction methods are mainly classified into two main parts.

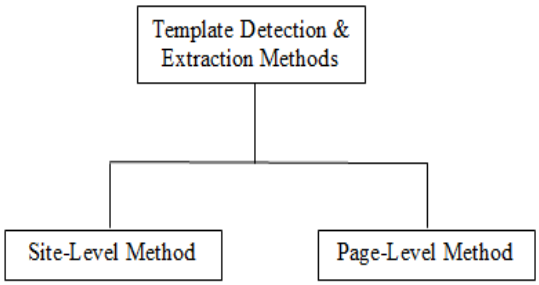

Fig.4. Classification of Template Detection and Extraction Methods

Each of method is described in this section.

(a) Site-level method- In Site-level methods templates based on several pages from a site. A set of sample pages will be collected as input and templates are detected based on various factors like similarity criteria or a threshold value depending on the technique used. Site-level type of template detection and extraction in which templates are found out from a set of heterogeneous web pages from a site. The method works automatically with the help of algorithms and it performs a type of grouping known as clustering based on similarities existing in the input pages. This method contains four algorithms Extract Sub Tree, RTDM-TD, Retrieve Template and finds Template. It uses a principle known as Minimum Description Length (MDL) for detecting templates and calculates a value termed as MDL cost to identify the best cluster. The method proposes an algorithm known as TEXT-MDL and extracts the detected templates [17].

(b) Page-level method- In Page-level method templates based on a single page. A page is taken as input and decision on templates is made based on a certain similarity criteria or a threshold value based on the technique used. Page-level type of template detection that detects templates on a page by page basis. The method proposes a framework that detects templates based on historical information stored about web pages. As soon as a page is collected, it is passed through the process of template detection. Detection is done based on the repetition of text portions termed as text segments in pages [8].

V.Proposed Methodology

The complete methodology of the proposed work is shown below in figure-5.

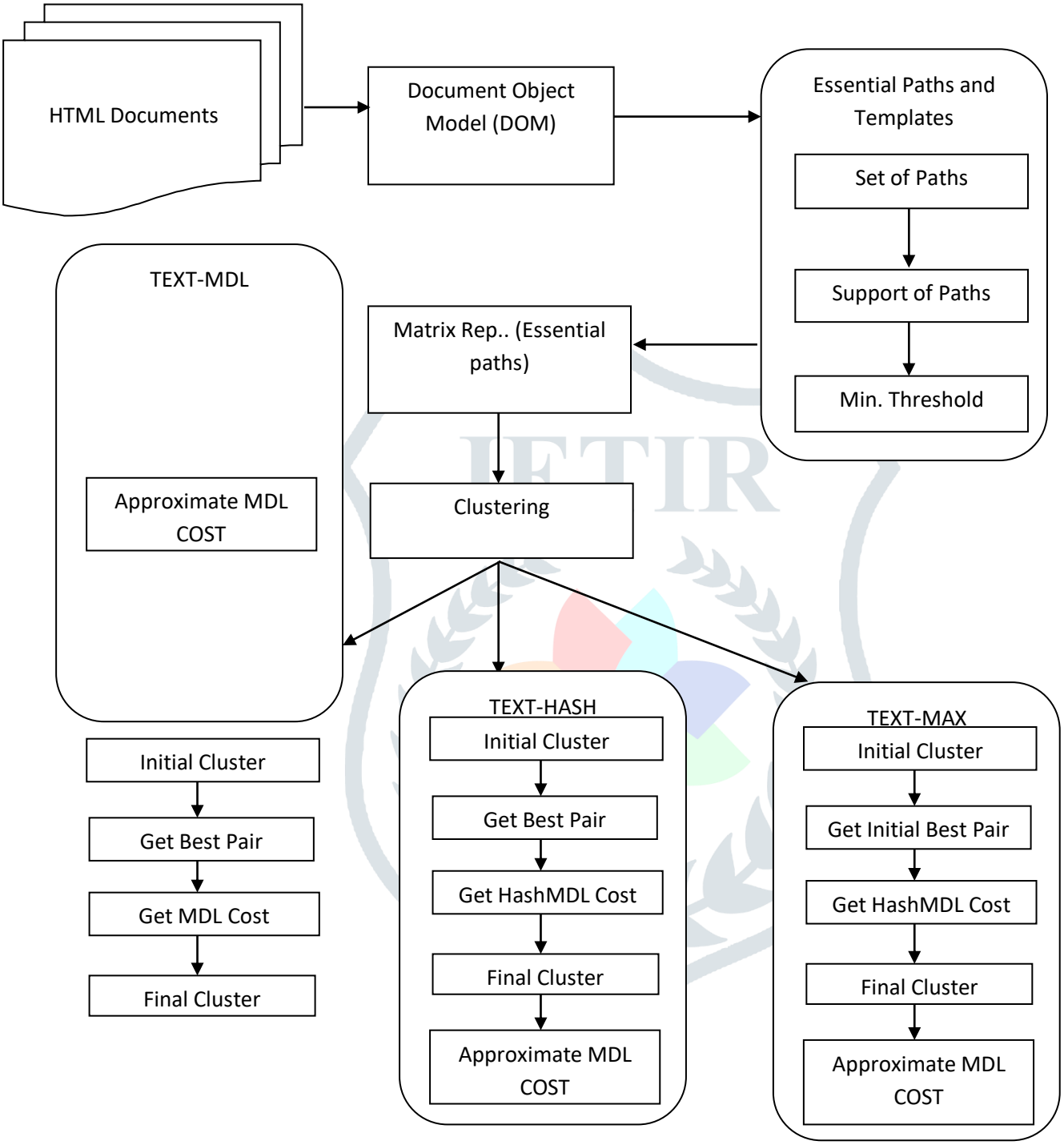

Fig.5. Flow Diagram of the Proposed Methodology

VI.Conclusion

In this paper, we presented a novel methodology for detecting and extracting templates in heterogeneous web pages for web application performance improvement. We used the MDL method to handle the unknown number of clusters and select suitable partitioning from all possible web documents, and then we used our extended MinHash

technique to accelerate the clustering process. Experiment results using real-world data sets have verified the efficacy of our methods.

References

[1] Jiawei Han, Michelin Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 2001, 70-181.
[2] Lyman, Peter; Hal R. Varian, "How Much Information"2003.
[3] Z. Bar Yossef and S. Rajagopalan, (2007), "Template Detection via Data Mining and It'sApplications",Proceedings 11th International Conference, World Wide Web(WWW), 2002.
[4] S. Zheng, D. Wu, R. Song and J.R. Wen, "JointOptimization of Wrapper Generation and TemplateDetection", Proceeding of ACM.
[5]Mohammed Kayed and Chia-Hui Chang, "FiVaTech: Page-Level Web Data Extraction from Template Pages", IEEE transactions on knowledge and data engineering, Vol. 22, No. 2, 2010.
[6] V. Crescenzi, G. Mecca, and P.Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites", Proceeding. 27th International Conference Very Large Data Bases (VLDB), 2001.
[7] Monika R. Henzinger Rajeev MotwaniCraig Silverstein, "Challenges in Web Search Engines" ACM SIGIR Forum, Volume 36 Issue 2, 2002.
[8] Yu Wang, Bingxing Fang, Xueqi Cheng, Li Guo, Hongvo Xu, "Incremental Web Page Template Detection", Proceeding17th International Conference World Wide Web(WWW), Pages 1247- 1248, 2008.
[9] R. Song, H. Liu, J.-R. Wen, "Learning block importance models for web pages", In Proceedings of the 13th International Conference on World Wide Web, Pages 203–211. ACM Press, 2004.
[10] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages", Proceeding ACM SIGMOD, 2003.
[11] Tak-Lam Wong, Wai Lam, "An unsupervised method for joint information extraction and feature mining across different web sites", Data & Knowledge Engineering, Volume 68, Issue 1, Pages 107-125, 2009.
[12] Xiangwen Ji, Jianping Zeng, Shiyong Zhang, Chengrong Wu, "Tag tree template for Web information and schema extraction" , Expert systems with Applications, Volume 37, Issue 12, Pages 8492-8498, 2010.
[13] Gilles Nachouki, Mohamed Quafafou, "MashUp web data sources and services based onsemantic queries", Information Systems, Volume 36, Issue 2, Pages 151-173, 2011.
[14] Hua Wang, Yang Zhang, "Web Data Extraction Based on Simple Tree Matching", IEEE Conference Proceedings on International Conference on Information Engineering, Volume 2, Pages 15-18,2010.
[15] Jellouli I., Mohajir M.E, "An ontology-based approach for Web Information extraction", IEEE conference Proceedings on Information Science and Technology, Pages 5, 2011.
[16] Hao Han, Tokuda T, "A method for Integration of Web Applications Based on Information Extraction", IEEE Conference Proceedings on Eighth International Conference on Web Engineering, Pages 189-195, 2008.
[17] Chulyun Kim and Kyuseok Shim, "Text:Automatic Template Extraction from Heterogeneous Web Pages", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 4, 2011.