INFROMATION EXTRACTION IN WEBDOCUMENT USING Clustering TECHNIQUE

D. Saravanan

Faculty of Operations & IT
ICFAI Business School (IBS), Hyderabad,
The ICFAI Foundation for Higher Education (IFHE)
(Deemed to be university u/s 3 of the UGC Act 1956)
Hyderabad-India

Abstract: *The problem of extracting a template from the web documents conforming to a common template has been studied. Due to the assumption of all documents being generated from a single common template, solutions for this problem are applicable only when all documents are guaranteed to conform to a common template. However, in real applications, it is not trivial to classify massively crawled documents into homogeneous partitions in order to use these techniques. Since subtle changes in scripts or CGI parameters may result in a significant difference, we cannot simply group the web documents by URL and apply these methods for each group separately. In this problem, clustering of web documents such that the documents in the same group belong to the same template is required, and thus, the correctness of extracted templates depends on the quality of clustering. To overcome this in this paper we propose a Hyper graph based clustering mechanism for extracting HTML tags and templates from a large number of web documents.*

Keyterms: *Clustering, Web documents, Web pages, Tags, Hyper graph.*

1. INTRODUCTION

The World Wide Web (WWW) is widely used to publish and access information on the Internet. In order to achieve high productivity of publishing, the web pages in many websites are automatically populated by using common templates with contents. For human beings, the templates provide readers easy access to the contents guided by consistent structures even though the templates are
not explicitly announced. However, for machines, the unknown templates are considered harmful
because they degrade the accuracy and performance due to the irrelevant terms in templates. Thus, template detection and extraction techniques have received a lot of attention recently to improve the performance of web applications, such as data integration, search engines, classification of web documents .For example, bio gene data are published on the Internet by many organizations with different formats and scientists want to integrate these data into a unified database. For price comparison services, the price information is gathered from various Internet marketplaces. Good template extraction technologies can significantly improve the performance of these applications.

To overcome the limitation of the techniques with the assumption that the web documents are from a single template, the problem of extracting the templates from a collection of heterogeneous web documents, which are generated from multiple templates, was also studied.

1.1 Existing systems

1. The primary problem is of forming a common template from extracted web documents templates.
2. Next problem is of grouping the web documents by URL and apply various methods for each group separately is harder.

3. Further problem is the correctness of extracted templates depends on the quality of clustering.
4. However, clustering is very expensive with tree-related distance measures.

1.2 Issues

Downloading of 1000 documents from the web, and then sent them to be the summarizer, and select the best one as template is difficult without clustering.Inefficient when pages are not processed in site order

Eg: in a web crawler pipeline(Most template detection methods process web pages in batches that a newly crawled page can not be processed until enough pages have been collected. This results in large storage consumption and a huge delay of data refreshing.)

- Need to maintain hashes and counts for all sites
- Marking site-level templates for new websites
- Not all templates are site-level in nature
- Low recall

1.3 Proposed system

- Using heterogeneous extraction (multiple web sites template extraction)
- Provide a faster and efficient way of clustering.
- Clustering on web documents is used practically to handle large number of web documents.
- We cluster only documents not paths, and moreover, the numbers of clusters of columns and rows are dynamic

1.3.1 Advatages

- Information extraction is attempting to find some of the structure and meaning in the hopefully template driven web pages.
- Used no manually labeled training data.
- Very high precision
- Extract general structural and content cues from the DOM nodes

2. SYSTEM OVERVIEW

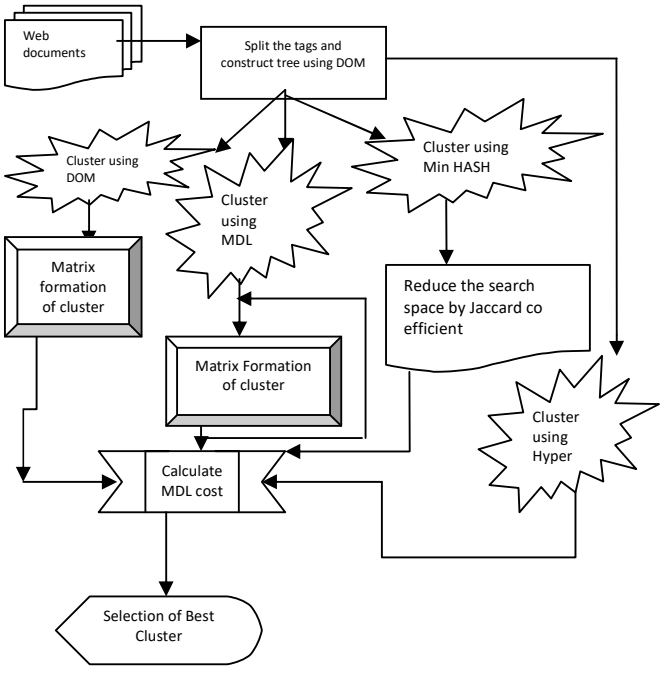

Figure 1 Overall Architecture

3. Experimental Setup

1. Construction of DOM tree
2. Clustering and matrix formation using DOM (Document Object Model)
3 Applying Cluster using MDL (Minimum Description Length).
4 Cluster using Min HASH
5 Cluster using proposed approach (Hypergraphs)

3.1. Construction of DOM tree

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

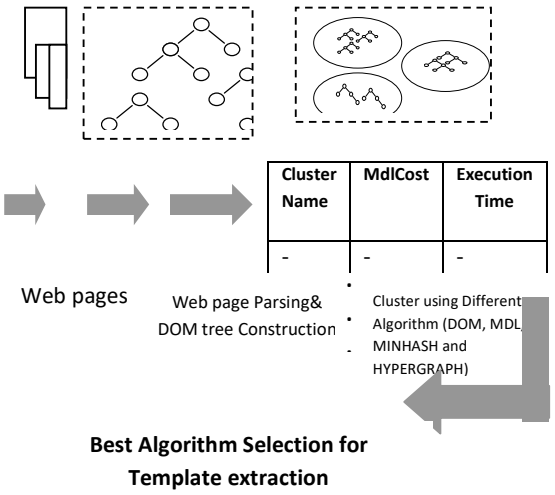

Figure 2. System Architecture

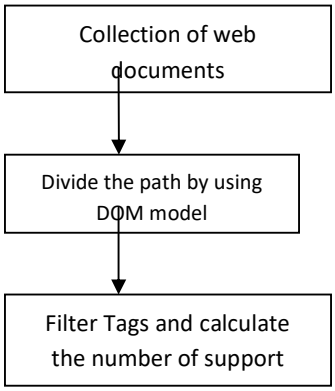

Figure3. DOM Tree Construction

3.2. Clustering and matrix formation using DOM

Here we use DOM (Document Object Model)based clustering mechanism to cluster the html tags that are extracted. In this clustering mechanism we are providing a support threshold value and this threshold value depends upon the document minimum path support value (Di). For the formation of the matrix value we take the considerations as web document set D with its path set PD, we use a |PD| x |Di| matrix ME with 0/1 values to represent the documents with their essential paths. The value in the matrix ME is 1 if a path is an essential path of a document di. Otherwise, it is 0.The MDL cost is identified in order to find the efficiency of the individual clustering algorithm and is given as Cost(M,D) = Cost(D|M) + Cost(M) where Cost(M) - cost of the path and Cost(D|M) - cost of the data D if path M is given. Thus we do not need any additional template extraction process after clustering.

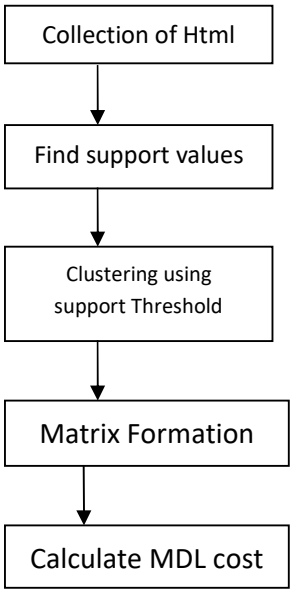

Figure 4.Clustering and Matrix Formation using DOM

3.3.Applying cluster using MDA

In order to manage the unknown number of clusters and to select a good partitioning of cluster from all possible partitions of HTML documents, we employ MDL principle. TEXT-MDL is an agglomerative hierarchical clustering algorithm which starts with each input document as an individual cluster. When a pair of clusters is merged, the MDL cost of the clustering model can be reduced or increased. The procedure GetBestPair finds a pair of clusters whose reduction of the MDL cost is maximal in each step of merging and the pair is repeatedly merged until any reduction is not possible.

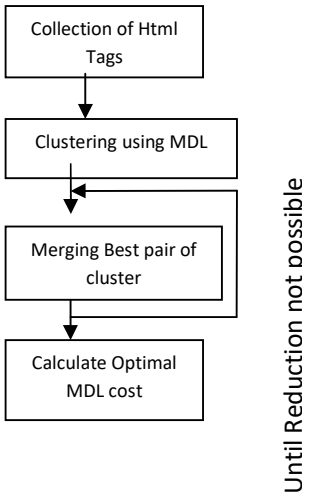

Figure 5. Minimum Description Length

3.4. CLUSTER USING MIN HASH

A way to consistently sample words from bags and which is a technique for quickly estimating how similar two sets are. This Clustering algorithm uses hash intersections to probabilistically cluster similar user data. In order to find the duplications in the web page we utilize the jaccard coefficient for similarity measurement.

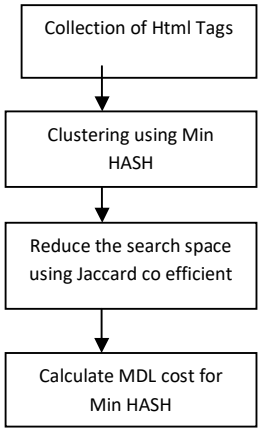

Figure 6.Cluster Using Min HASH

3.5. CLUSTER USING HYPERGRAPH

A hypergraph is a generalization of a graph where in edges can connect more than two vertices and are called hyperedges.

Hypergraph (H) = (V, E) V::a set of vertices; E::a set of hyperedges.

The clustering problem is then formulated as of finding the minimum-cut of a hypergraph. A minimum-cut is the removal of the set of hyperedges (with minimum edge weight) that separates the hypergraph into k unconnected components.

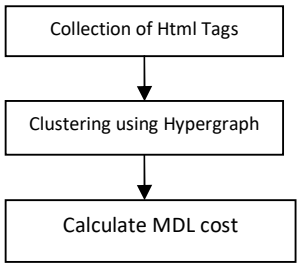

Figure 7.Cluster Using Hypergraph

4. ALGORITHM USED

4.1 AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

```
C={c1,c2…..cn}
(ci,cj,ck)=GetBestPair(C)
Let ci and cj be the best pair of merging
Let ck be a new cluster made by merging ci and cj
While(ci,cj,ck) is not empty do
{
 C=C-{ci,cj}U{ck}
  (ci,cj,ck)=GetBestPair( C)
 }
return C
 end
```

4.2 minHASH CLUSTERING ALGORITHM

This Clustering algorithm uses hash intersections to probabilistically cluster similar user data. In order to find the duplications in the web page we utilize the jaccard coefficient for similarity measurement.

The jaccard's coefficient assigning random ranks to the universal set and comparing the minimum values from the ranks of each set.

The jaccard coefficient between two sets s1 and s2 is defined as $\gamma(s1,s2)=|s1 \cap s2|$
$|s1 U s2|$

5. CONCLUSION

We introduced a novel approach of the template detection from heterogeneous web documents. We employed the MDL principle to manage the unknown number of clusters and to select good partitioning from all possible partitions of documents, and then, introduced our extended MinHash technique to speed up the clustering process.

Experimental results with real life data sets confirmed the effectiveness of our algorithms. Also here we conclude that our clustering proposed approach is well designed and suitable for identify the web page template and to extract its unstructured data. Also our proposed Hypergraph based clustering algorithm is fine tuned for accuracy and efficiency and faster time response. Experimental results on different web pages are established to check the feasibility of the proposed algorithm.

6. FUTURE ENHANCEMENT

In our future work to avoid the risk of missing critical information that is scripts. Another issue is handling images with descriptive functions in the web pages. It is not good to simply remove any image from the page, as we all know that a picture sometimes worth one thousand words. For example, fashion site like elle.com prefers image text rather than a plain text. Image text certainly facilitates readers' understanding; on the other hand it prohibits data extraction. Recently Google enables the search for images by the image name. From our observation, the "ALT" parameter of an image link sometimes also describes the image content, if edited by a human editor. We include both image name and "ALT" value in our data extraction process, while getting rid of worthless values such as "img9", "spacer" or "graphic". We are working on evaluating the effect of including informational image name and image name parsing rules

7. Experimental Results

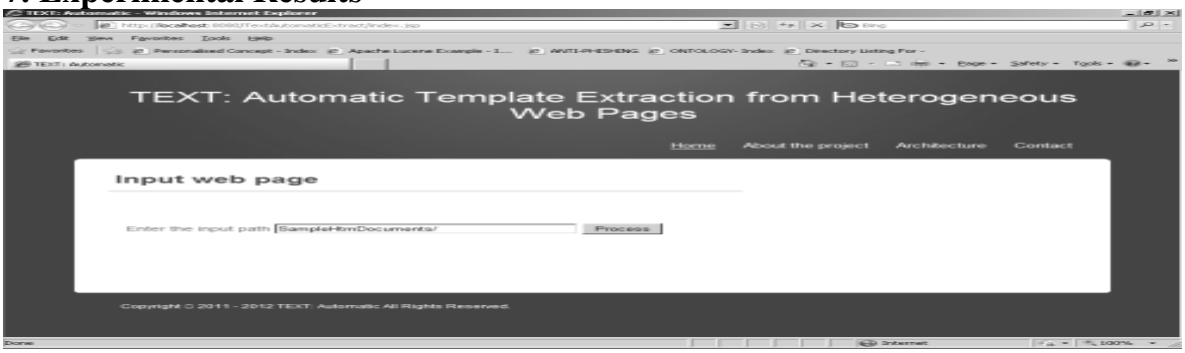

Figure 8. Input Web Page

Figure 9. File Processing

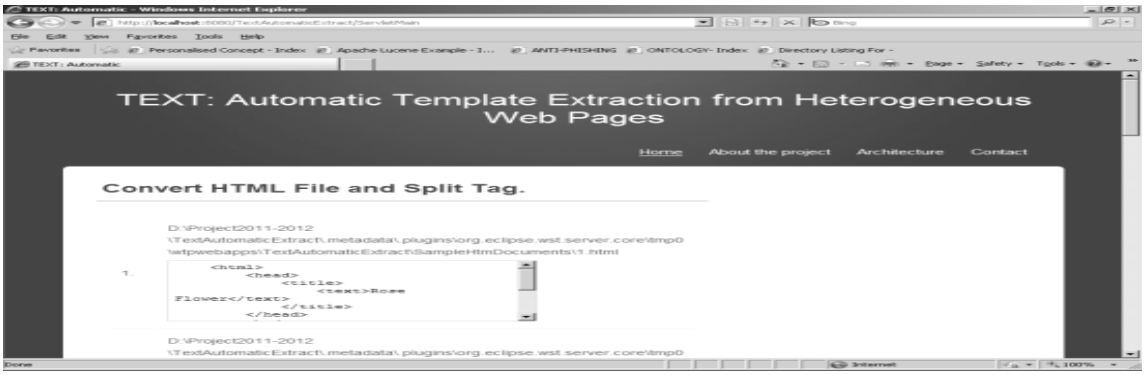

Figure 10. Split Tag

Figure 11. Path Token

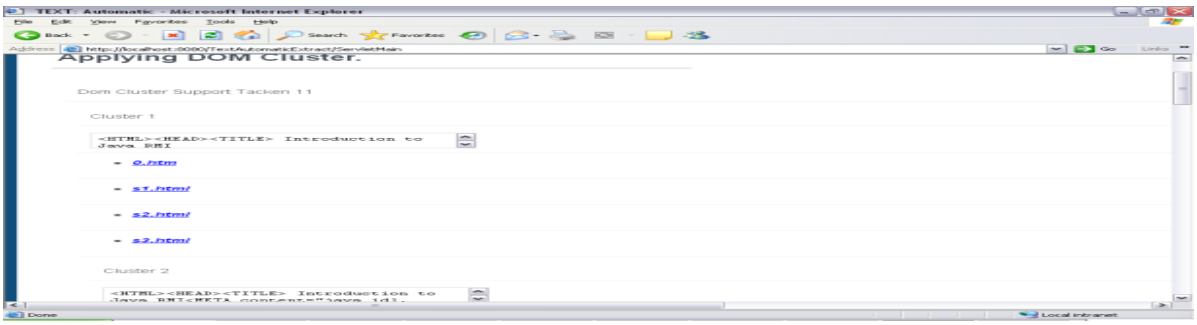

Figure 12 Dom Cluster

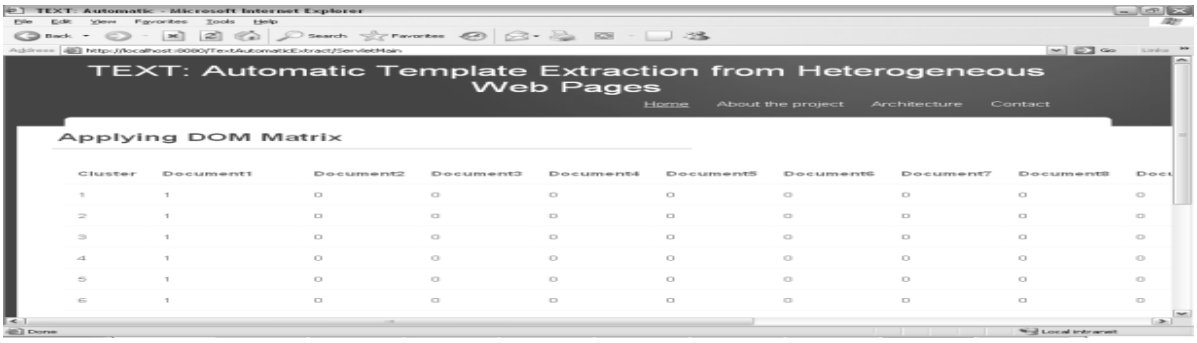

Figure 13 Dom Matrix

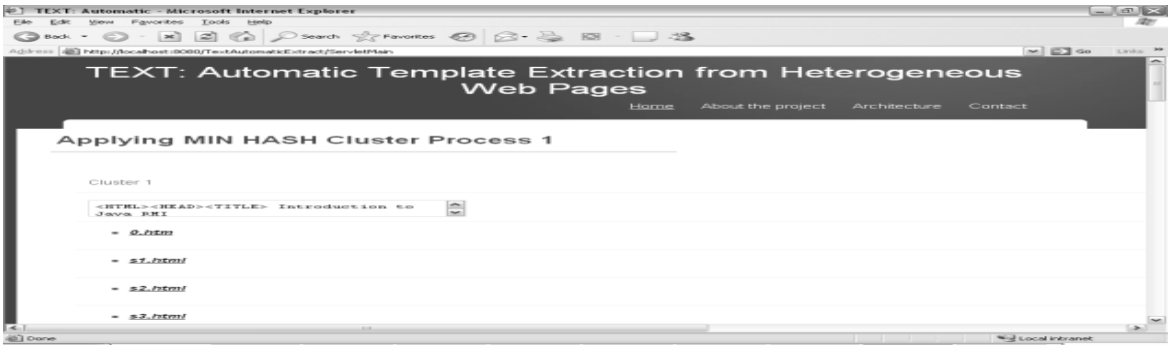

Figure14 MIN NASH Cluster Process

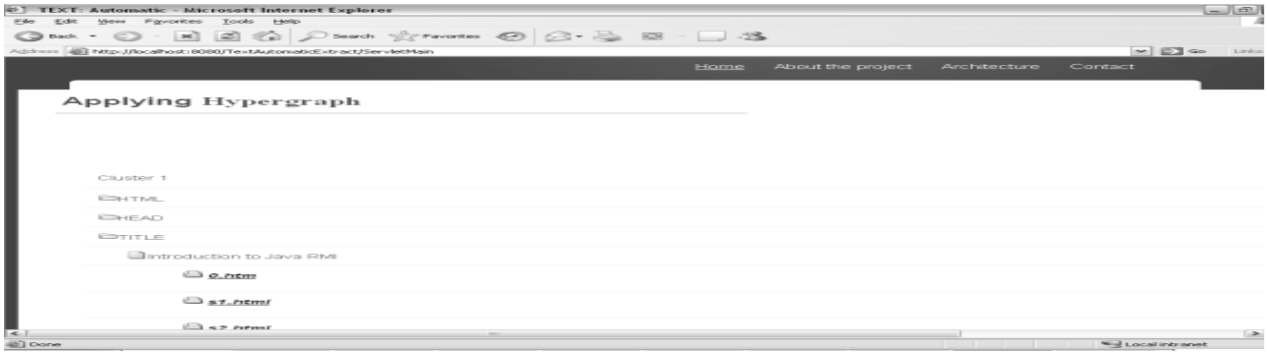

Figure 15Hypergraph

Figure 16 Analysis of Cluster Process

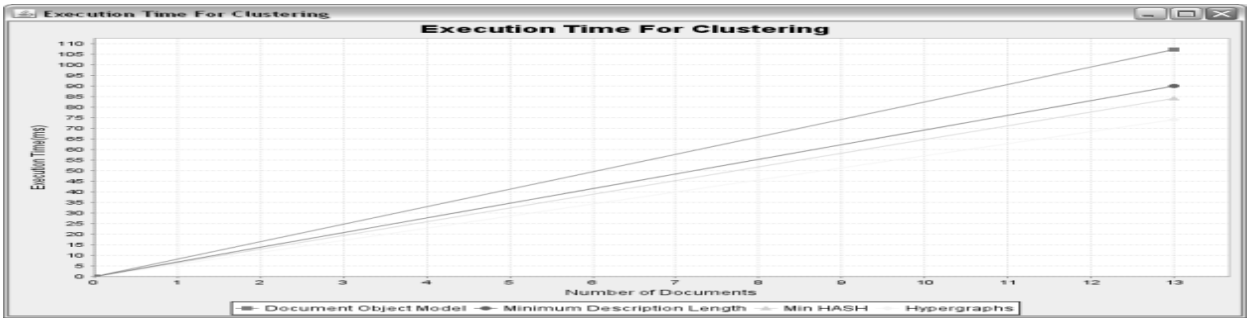

Figure 17 Execution Performance

REFERENCES

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

[2] *Z. Bar-Yossef and S. Rajagopalan, "Template Detection via Data Mining and Its Applications," Proc. 11th Int'l Conf. World Wide Web (WWW), 2002.*

[3] *A.Z. Broder, M. Charikar, A.M. Frieze, and M. Mitzenmacher, "Min-Wise Independent Permutations," J. Computer and System Sciences, vol. 60, no. 3, pp. 630-659, 2000.*

[4] *D. Chakrabarti, R. Kumar, and K. Punera, "Page-Level Template Detection via Isotonic Smoothing," Proc. 16th Int'l Conf. World Wide Web (WWW), 2007.*

[5] *Z. Chen, F. Korn, N. Koudas, and S. Muithukrishnan, "Selectivity Estimation for Boolean Queries," Proc. ACM SIGMOD-SIGACTSIGART Symp. Principles of Database Systems (PODS), 2000.*

[6] *Saravanan", Information extraction using user opinion procedure", International journal of recent technology and engineering (IJRTE), Pages 10403-10407, Volume 8, Issues-4, Nov 2019*

[7] *D.Saravanan," Efficient Video indexing and retrieval using hierarchical clustering techniques", Advances in Intelligence systems and computing, Volume 712, Pages 1-8, ISBN:978-981-10-8227-6, Nov-2018*

[8] *V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," Proc. 27th Int'l Conf. Very Large Data Bases (VLDB), 2001.*

[9] *D.Saravanan, Multimedia data Retrieval Data mining image pixel comparison technique", Lecture notes on Data Engineering and communications Technology 31,Aug 2019, Pages 483-489, ISBN 978-3-030-24642-6, Chapter 57*

[10] *M. 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'l Conf. World Wide Web (WWW), 2004.*

[11] *D.Saravanan, Dr. Dennis Joseph, " A study on hierarchical clustering algorithms", American International journal of research in science , technology, engineering &*

Mathematics'(AIJRSTEM), March-May 2019, Vol. 1, issue 1, Pages 87-89, May2019,(ISSN2328-3580)

[12]*Vinod Kumar Raav, Sathya p kumarsomayajula,"Automatic template extraction from heterogeneous webpages", International journal of advanced research in computer science and software engineeri9ng, volume 2, issue 8, aug 2012, pages 408-418.*

[13] *D.Saravanan, "Information retrieval using image attribute possessions" Soft computing and signal processing , Advances in Intelligence systems and computing 898, Springer. DOI:10.1007/978-981-13-3393-4_77, Pages 759-767. March 2019.*