www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 4 April, 2013 Page No. 1106-1110

Extraction of Html Documents From Heterogeneous WebPages Using Cluster Techniques

Sruthi Kamban K.S, M.Sindhuja
Department of Information Technology, Rajalakshmi Engineering College, Chennai
sruthikamban@gmail.com

Department of Information Technology,Rajalakshmi Engineering College, Chennai
sindhuja.m@rajalakshmi.edu.in

Abstract: The World Wide Web is a vast and rapidly growing source of information. Most of this information is in the form of unstructured text which makes the information hard to query. To make the queries easy and to provide the result accurately, template extraction technique is used .In the existing system the techniques which are used to extract the data is not efficient and causes the factors such as delay, accuracy, and duplicate data. The proposed system is presented with Hyper Graph technique for extracting the templates from a large number of web documents which are generated from heterogeneous templates for making the web search more efficient in cost wise, performance and time wise. In addition the proposed approach make use of a clustering technique to retrieve the web documents based on the similarity of underlying template structures in the documents so that the template for each cluster is extracted simultaneously providing goodness measure with its fast approximation for clustering.

Key terms: Document Object Model, Min Hash, Minimum description length, Jaccard coefficient, Template Extraction.

I. INTRODUCTION

Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. Data mining allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases

The 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, and so on. Templates contains a considerable number of common terms and hyperlinks which are replicated in large number of pages, a relevance assessment that does not take templates into account may turn out to be inaccurate, leading to incorrect results. Template detection and removal techniques follows the trend of recent works which aim to detect unimportant (or less important) information of web pages and web collections in an attempt to save computational resources, to reduce the cost of the web searches and to speed up processing.. The main aim is to find the best algorithm that suits for the web searches in terms of speed and costs.

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 significant difference. Next is to group the documents with the means of URLs. Documents in the cluster are generated by identical URLs. In this case their URLs are identical except the value of the layout parameter. If only URLs are used to group pages, these pages from the different templates will be included in the same cluster. To overcome this limitation the web documents are treated as single template.

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. We cluster only documents not paths. So, none of the websites are omitted during clustering and there are no chances of occurrence of document duplication.

II. RELATED WORKS

The problem of template detection in large scale search engines has been studied by L. Chen and X. Li [3].Thomas Gorton [6], proposed the difference between content extraction algorithm and template detection algorithm. 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. Arasu and H. Garcia-Molina [1] proposed paper on extracting structured data from web pages which describes the automatic extraction of database values from templates without any human input. V. Crescenzi, P. Merialdo, and P. Missier [4] clustering studied on web Pages based on their structure in which they determined the homogeneity properties of entire collection of WebPages. F. Pan, X. Zhang, and W. Wang [8] proposed their paper on fast co-clustering on large datasets using matrix decomposition. Co-clustering of Rows and columns Decomposition (CRD) is used to cluster the rows and columns and Iterative Single Side Clustering algorithm is used to iterate the values of rows and columns.

Chuan Yuen Mengjun Xie Haining Wang [2] studied on automatic cookie usage setting with cookie picker. In this paper, Cookie picker is used to validate the usefulness of the cookies and acts behalf of users. Meng X F, Wang H Y, Hu D [10] SG-WRAM: Schema Guided Wrapper Maintenance for Web Extraction. The wrapper generation and wrapper maintenance are provided here.

R. Song, H. Liu, J.-R. Wen [7] learned the block importance models for web pages which extract the noisy portions in the websites using blocks. The VIPS (Vision - based Page segmentation) algorithm is used here. Lan Yi, Bing Liu, Xiao Li [5] In their paper about elimination of noisy Information in Web Pages for Data Mining proposed that web pages in a given Website usually share some common layout or presentation styles, a new tree structure, called Style Tree (ST) is proposed .

III. PROPOSED METHOD

A. Architecture

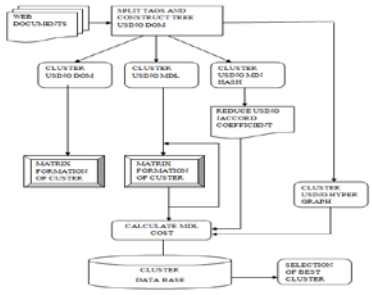

Fig.1 Proposed architecture diagram

Fig.1 represents the architecture diagram of the system. The architecture describes the extraction of web document is done in efficient manner. The documents are extracted and Dom tree is constructed by splitting the tags. Our proposed technique Hyper graph is also used as one of the algorithm to cluster. The cluster techniques are then used to cluster the documents and matrix formation is done. At last the values are stored in the database and they are compared. This comparison is done to find out the best cluster algorithm.

IV. CLUSTER TECHNIQUES USED IN TEMPLATE CLUSTERING

Following steps are involved in clustering the templates:

A. 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.

B. Clustering and matrix formation using DOM (Document Object Model)

DOM based clustering mechanism is used 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, |PD| x |Di| matrix ME is used 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.

C. Applying Cluster using MDL (Minimum Description Length)

In order to manage the unknown number of clusters and to select a good partitioning of cluster from all possible partitions of HTML documents, MDL principle is employed. 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

D. 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.

E. Cluster using proposed approach (Hyper graphs)

A hyper graph is a generalization of a graph where in edges can connect more than two vertices and are called hyper edges. Hyper graph (H) = (V, E) V::a set of vertices; E::a set of hyper edges. The clustering problem is then formulated as of finding the minimum-cut of a hyper graph. A minimum-cut is the removal of the set of hyper edges (with minimum edge weight) that separates the hyper graph into K unconnected components

V. EXPERIMENTAL RESULTS

All algorithms in this section were implemented in jsp. For back end SQL server was used. A collection of 13 html pages are collected and the algorithms were used in it. Examination of the algorithm performances was done.

A. Implemented Algorithms

The related work and the proposed work algorithms are been used:

DOM: This algorithm uses the DOM structure of the pages on a website by searching for nodes of the DOM tree that are repeated across multiple pages on the website.DOM tree is constructed using the structures. The html tags are separated. 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. At last, the matrix is formed. 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. 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. Dom matrix is formed as shown in fig.2

Cluster	Document1	Document2	Document3	Document4	Document5	Document6	Document7
1	1	0	0	0	0	0	0
2	1	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Fig.2 Dom Matrix

MDL: 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. To select a good partitioning of cluster from all possible partitions of HTML documents MDL cluster is used. Clustering the documents using MDL does not provide accurate result. So, for this purpose, the MDL matrix is evaluated. Thus, mdl matrix is provided using the mdl cluster. The matrix is represented in fig.3.

Cluster	Document1	Document2	Document3	Document4	Document5	Document6	Document7
1	1	1	1	1	1	1	1
2	1	0	0	0	0	0	0

Fig.3 MDL Matrix

Min Hash: MinHash (or the min-wise independent permutations 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.

Jaccard Coefficient: To make the min hash results effective jaccard coefficient is used. 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. It is represented as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Using the jaccard coefficient the smallest cluster formed.

Hyper Graph: A hyper graph is a generalization of a graph in which an edge can connect any number of vertices. Formally, a hyper graph H is a pair H=(X,E) where X is a set of elements called nodes or vertices, and E is a set of non-empty subsets of X called hyper edges or

edges.Clustering the documents using this algorithm makes the no of documents fewer .So that the searches will be made easy and the execution time will be lower compared to other algorithms.

By comparing algorithms such as Dom, MDL, Min Hash, Jaccard Coefficient and Hypergraph which is the proposed system, the best cluster identified as Hypergraph both in cost and time wise. In fig.4 the best cluster is evaluated as Hyper Graph by comparing it with other algorithms.

Fig.4 Finding Best Cluster

VI. PERFORMANCE EVALUVATION

The performance evaluation in the project is done for the purpose of cost and execution time. The graphs for cost and the execution is the real time graph in which the performance changes is done according to the changes made in the performance if the algorithm

A. Cost Evaluation

The cost analysis is done for the algorithms DOM, MDL, MinHash and hyper graph. In the evaluation of cost made so far, the cost of hyper graph is lower than the other algorithms. This indicates that hyper graph is the best algorithm to be used. The graph of cost analysis is given in figure 5.

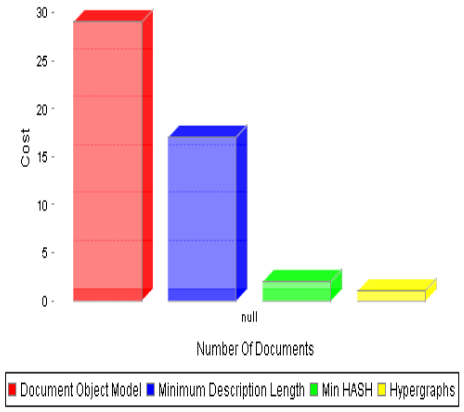

Fig.5 Cost analysis

B. Execution Time

The execution time for each algorithm is evaluated along the cost evaluations. Thus in our work the execution time for hyper graph is lower than other algorithms. The fig.6 represents that hyper graph algorithm is the best algorithm of all the other algorithms used.

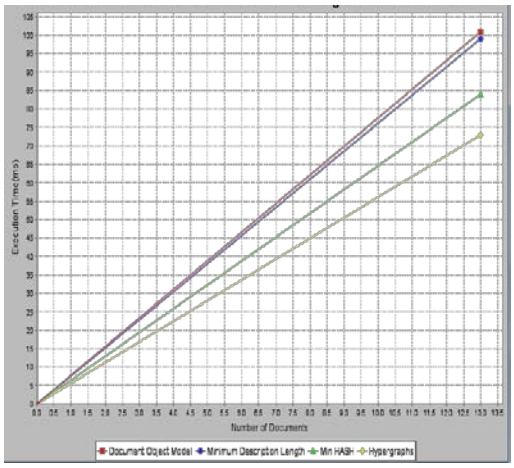

Fig.6 Execution Time

VII. CONCLUSION

A novel approach of the template detection from heterogeneous web documents is introduced. The system proposed provides selection of algorithm which could further be applied in the web pages to extract the document in efficient, accurate manner. The proposed algorithm Hyper graph is added as a module with the existing system. This work is done in order to reduce the duplications and to increase the speed in which the document is loaded. Due to the algorithm hyper graph, the speediness and the cost is reduced.By using the proposed algorithm the web searches will be made easy and the time consumption will be less for searching the web documents.

VIII. ACKNOWLEDGEMENT

I am very thankful to my guide Mrs.M.Sindhuja of Rajalakshmi Engineering College ,who has rendered her whole-hearted support at all times for the successful completion of the work Extraction of html documents from heterogeneous WebPages using cluster techniques

REFERENCES

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

[2] Chuan Yuen Mengjun Xie Haining Wang, Automatic Cookie Usage Setting With Cookie Picker

[3] L. Chen, S. Ye, and X. Li. Template detection for large scale search engines. In SAC, pages 1094–1098. ACM, 2006.

[4] V. Crescenzi, P. Merialdo, "Clustering Web Pages Based on Their Structure," Data and Knowledge Eng., vol. 54, pp. 279-299, 2005

[5] D. Gibson, K. Punera, "The Volume and Evolution of Web Page Templates," Proc. 14th Int'l Conf. World Wide Web (WWW), 2005.

[6] Lan Yi, Bing Liu, Xiao Li. Eliminating Noisy Information in Web Pages for Data Mining. Of the SIGKDD'03 Conf., pages 296305, 2003

[7] Meng X F, Wang H Y,SG-WRAM: Schema Guided Wrapper Maintenance for We Extraction, Demonstration Proceedings of ICDE, 2003, 750-752.

[8]]F. Pan, X. Zhang, A General Framework for Fast Co-clustering on Large Datasets Using Matrix Decomposition Proc. ACM SIGMOD, 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] Thomas Gorton, Bridging the Gap: From Multi Document Template Detection to Single Document Content Extraction.