International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 5, May - 2014

Website Cleaner Using Site Style Tree (SST)

Pratiksha U. Jadhav
Department of Computer Engineering,
University of Pune,
KKWIEER,
Nashik, India.

Prof. Rupali D. Kulkarni
Department of Computer Engineering,
University of Pune,
KKWIEER
Nashik, India

Abstract— **A webpage generally contains data along with navigation panels, advertisements, copyright and privacy notices. Except data these other things does not contain any important information. These blocks can be called as non-informative blocks. As these blocks are non-informative, they can affect the result of web data mining. To avoid this it is important to separate the main data i.e.. informative blocks and non-informative blocks from the web page. In a website these non-informative blocks are generally present in different web pages and have same format. Also the data contained in these blocks is also same. In case of informative blocks, data contained by the block and their format are different. We need a structure at site level to capture the same format of the blocks and the data present in the blocks. DOM Tree structure is available at page level. Many tools are available to construct a DOM Tree of a webpage. But DOM Tree structure is not useful at site level. So we need to construct a Site Style Tree(SST) for a website. After analyzing this SST we can identify which part of SST is informative and which is non-informative. There is no tool available to construct a style tree for a given website. This work aims at constructing a style tree for given website and separating informative and non-informative blocks from the website.**

Keywords— Informative blocks, Non-informative blocks, Web mining.

I. INTRODUCTION

hese days we can get any information on the world wide web. It is the main source of information. As large amount of information is available on web, it is very important that one should get useful or required information from the web. To provide user friendly environment; navigation bars, copyright notices are present along with main data. Also different types of advertisements are also included on website. To make the website attractive many decorative images are also included.

These all items are useful for viewers and necessary for website. Due to these items, retrieving required information from the web becomes very difficult. To improve the process of web data mining it is necessary to remove non-informative blocks from the web pages. For this we first need to identify such blocks from the website. This work aims at identifying such non-informative blocks from the website. These non-informative blocks can be removed to make the data mining more efficient.

Non-informative blocks generally share same contents and presentation style in multiple web pages. So capture this at site level a structure called Site Style Tree (SST) is needed. Once the SST for a website is built, informative and non-informative parts can be easily identified. The example SST formation from [1] is shown below.

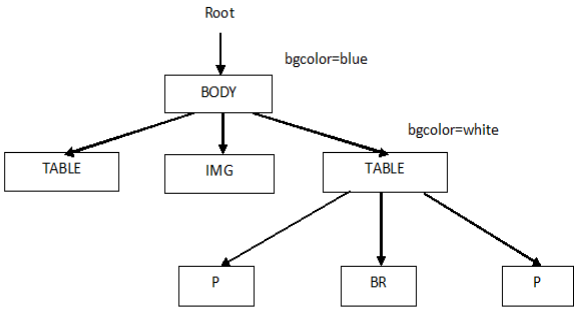

Fig. 1. DOM Tree D_1

In Fig. 1. an example DOM tree D_1 of a web page is shown. Intermediate nodes in the DOM Tree represents different HTML tags from corresponding web page and leaf node contains the actual content from the web page.

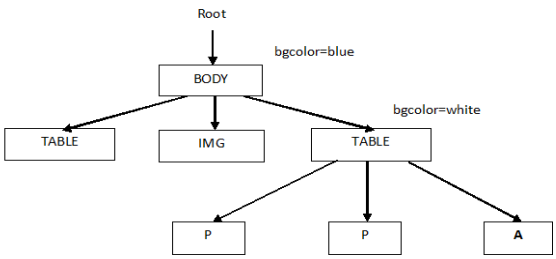

Fig. 2. DOM Tree D_2

Fig. 2. shows another DOM tree D_2. All tags in D_1 has its corresponding tags in D_2 except the bottom level tags. So

these two DOM trees can be combined to generate a Style tree.
The resultant Style tree is shown on Fig. 3.

This style tree contains all the nodes from D_1 and D_2. There are two types of nodes in style tree, element nodes and style nodes. In Fig. 3. P-BR-P and P-P-A are two style nodes. Tag nodes in the style node are called element nodes. A count is maintained which indicates how many pages have same presentation style at that level.

From this style tree we can observe that two presentation styles are present under rightmost table tag.
Thus by applying informative measure we can identify non-informative blocks. To clean the website we can remove these non-informative parts. Also when the new pages are added in the website, that page can be mapped on the SST of that site.

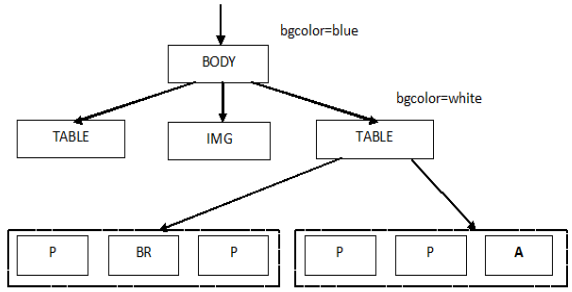

Fig. 3. Style Tree

II. LITERATURE SURVEY

A method is proposed in [4] to detect informative blocks from the news website. This work assumes that system already knows how webpage is partitioned and blocks containing similar information from different pages. But partitioning webpage and identifying corresponding blocks in different web pages are big issues. Also in [4] web page is considered as collection of blocks and each block as collection of words.
This is true in case of news website. Generally these assumptions are very strong.

Web page cleaning is considered as frequent template detection problem in [3]. In [3] webpage partitioning depends on the number of hyperlinks of an HTML element. This partitioning method is not useful in case of web pages from the same website.

In [8] some learning mechanisms are proposed. These helps to identify banner advertisements, redundant links of web pages. But these techniques require large training data set. These also require domain knowledge to generate classification rules. Some work includes duplicate records detection and cleaning of data for data mining .

In [6] one approach called block analyzer is mentioned. Using this approach blocks are preclustered. An entropy based value is assigned to each cluster. Using this value the blocks in the cluster can be classified as informative or non-informative blocks. But it may cluster some informative blocks into a noisy cluster.

III. IMPLEMENTATION DETAILS

A. *Mathematical Model*

Here consider WP as web page, DT as DOM tree, EDTS as evaluated DOM Tree Structure, SST is Site Style Tree, SA as Structure Analysis.
Also IB is Informative Blocks and NIB is non-informative Blocks, IH is input HTML web page.

Let S:{WP, DTG, EDTS, SST, IB, SA, NIB, IH}

Let WP:{wp_1, wp_2, ..., wp_n} where WP consist of number of web pages.

Let DT:{dt_1, dt_2, ..., dt_n} where DT consist of generated DOM Trees.

Let IH:{ih_1, ih_2, ..., ih_n} where IH is input HTML pages.
Let EDTS:{$edts_1$, $edts_2$, ..., $edts_n$} where EDTS consist of evaluated DOM Tree structures.

Let IB:{ib_1, ib_2, ..., ib_n} where IB consist of Informative Blocks.

Let NIB:{nib_1, nib_2, ..., nib_n} where NIB consist of Non-informative Blocks.

Function F_1 returns the DOM Trees of input web pages.
$F_1(WP) \rightarrow$ DT for input web page.
e.g. $F_1(WP) \rightarrow \{dt_1, dt_2, ..., dt_n\}$ DT
Function F_2 evaluates the generated DOM Trees.

$F_2(DT) \rightarrow$ EDTS for generated DOM Trees.
e.g. $F_2(DT) \rightarrow \{edts_1, edts_2, ..., edts_n\}$ EDTS

Function F_3 returns the Site level tree structure.
$F_3(IH, EDTS) \rightarrow$ SST

Function F_4 returns structure analysis results.
$F_4(SST) \rightarrow$ SA

Function F_5 returns final results.
$F_5(SA) \rightarrow$ IB, NIB for *generated* SST of a website.
e.g. $F_5(SA) \rightarrow \{ib_1, ib_2, ..., ib_n, nib_1, nib_2, ..., nib_n\}$ IB, NIB

B. *Functional Dependency*

	F_1	F_2	F_3	F_4	F_5
F_1	1	0	0	0	0
F_2	1	1	0	0	0
F_3	0	1	1	0	0
F_4	0	0	1	1	0
F_5	0	0	0	1	1

C. Process block Diagram

As shown in the diagram below, web page from a website is downloaded and provided to HTML Parser as a input. HTML Parser then generates DOM tree for that web page by parsing it. From this generated DOM tree, tags which does not contain any data or information, e.g. script tag, gets filtered using filter tags database. The result of this filtration process is the final DOM tree for that web page. This DOM Tree structure gets evaluated and resulting structure is called as Site Style Tree (SST). Now input from remaining HTML pages of a website is provided to generate the SST for a website. Leaf nodes in the DOM tree contain actual content of the page like text, image. So based on the information in this actual contents importance of leaf node is decided. For this leaf nodes from the DOM tree are extracted and information analysis is done. So final SST gets analyzed by applying these informative measures. After analyzing this SST informative and non-informative blocks are identified.

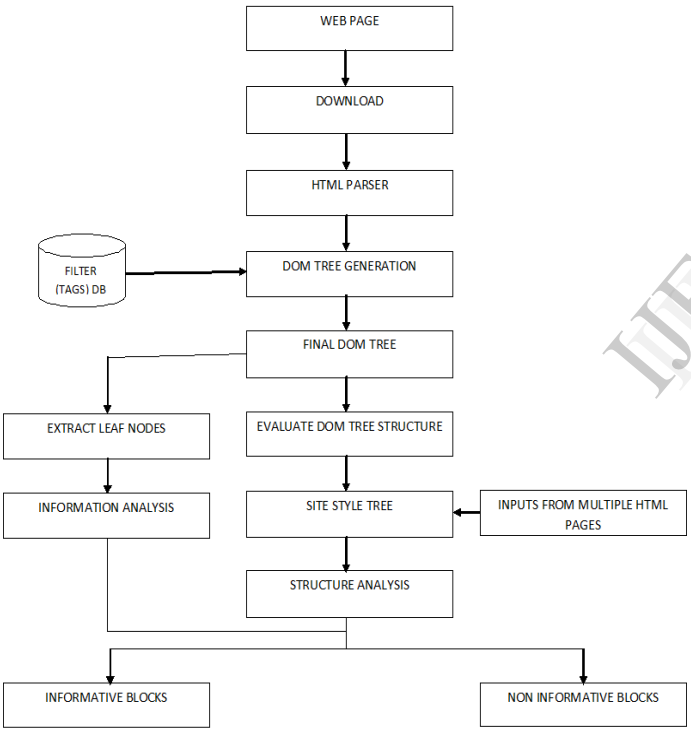

Fig. 4. Block Diagram

IV. SIMULATION

Aim of this system is to clean web pages to improve web data mining. Clustering and classification, these two tasks can be performed to evaluate the performance of the system. We can compare the results of the data mining before and after removing non-informative blocks.

A. Data set

Here data set from [2] is shown for the performance evaluation. This data set contains web pages from five websites. These sites contain information and overview pages of different products. As the sites shown in the table below are commercial websites, they contain large amount of non-informative blocks.

Websites shown in the data set contains variety of products. But web pages focusing the products shown in the Table 1 are chosen.

TABLE I
NUMBER OF WEB PAGES AND THEIR CLASSES

Web sites	Amazon	CNet	J&R	PCMag	ZDnet
Notebook	434	480	51	144	143
Camera	402	219	80	137	151
Mobile	45	109	9	43	97
Printer	767	500	104	107	80
TV	719	449	199	0	0

Table 1 shows the number of documents downloaded for the particular class of product from the corresponding website. We are using some live websites for our system testing.

Using this data set, clustering and classification is performed. Results of these tasks shows that removing non-informative blocks helps in improving data mining results.

V. CONCLUSION

To improve web data mining, web pages should be clean. For cleaning the web pages non-informative blocks must be identified. As non-informative blocks share common presentation style and content, a site level structure is used called Site Style Tree (SST). SST captures these common presentation styles. An information based measure is used to evaluate SST element nodes. When new pages are added to the website, these pages from the site are mapped to the SST. After removing non-informative blocks the storage space and time for a webpage can be saved.

REFERENCES

[1] R.Gunasundari, S.Karthikeyan, "Removing Non-informative Blocks from the Web Pages", Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference.
[2] B. Liu, K. Zhao, and L. Yi, "Eliminating Noisy Information in Web Pages for Data Mining" , Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery an Data Mining, pp. 296-305, 2003.
[3] Bar-Yossef, Z. and Rajagopalan, S., "Template Detection via Data Mining and its Applications", WWW 2002, 2002.
[4] Shian-Hua Lin and Jan-Ming Ho., " Discovering Informative Content Blocks from Web Documents", KDD-02, 2002.
[5] S. Debnath, P. Mitra, and C.L. Giles, N.Pal "Automatic Identification of informative sections of Web Pages" , IEEE Transaction on Knowledge and Data Engineering , 2005.
[6] Chia-Hsin Huang, Po-Yi Yen, Yi-Chan Hung, Tyng-Ruey Chuang, and Hahn-Ming Lee, "Enhancing Entropy-based Informative Block Identification Using Block Preclustering Technology ", 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan
[7] Hung-Yu Kao, Jan-Ming Ho, Member, IEEE, and Ming-Syan Chen, Fellow, IEEE, " WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model", IEEE transaction on Knowledge and Data Engineering, VOL. 17, NO. 5, MAY 2005.
[8] Jushmerick, N., "Learning to remove Internet advertisements", AGENT-99, 1999.
[9] Yao, Z. and Choi, B., 2007. "Clustering Web Pages into Hierarchical Categories," International Journal of Intelligent Information Technologies, Special Issue on Web Mining, Vol. 3, No. 2, pp.17-35.

[10] Peng, X. and Choi, B., 2005. "Document Classifications Based on Word Semantic Hierarchies," The IASTED International Conference on Artificial Intelligence and Applications, pp.362-367.