A Noise Reduction Approach based on n x 1 Table and XSL Display Method for Efficient Web Data Extraction

Neeraj Raheja[1,] V.K.Katiyar, PhD.[2]
Assistant Professor1, Professor & Head2
1, 2 Department of Computer Science and Engineering,
M.M.University, Mullana (Ambala), Haryana, India

ABSTRACT

A web page which is a source of information consist lots of parts among which only a part of the information is useful for a particular application and the remaining information are noises. An effective technique for users to extract the useful information from the total information is urgently required. Hence by removing those noise patterns from the web page, the efficiency of the web data extraction can be improved. This research work propose an approach for removing the local noise from a given web page based on n x 1 table and XSL display method with filter feature for improving the efficiency of web data extraction.

Keywords- noise reduction, web data mining, data extraction, XML, XSL.

1. INTRODUCTION:

As the web pages are unstructured or semi-structured in nature so to extract the required data directly is not an easy task as in databases which are structured in nature. Still the extraction of data from web pages is becoming a requirement because the information over the internet is growing at an exponential rate and users get simply lost in the linking structure on the web [1] [2], [3], [4].

In addition to main content, web pages commonly consist of have image-maps, logos, advertisements, search boxes, footers and headers, navigational links, related links and copyright information along with the primary content. Though these items are required by web site owner (for marketing purpose) and by the users (to navigate through the website) but all these content affect the web data mining and reduce the performance of the search engines and web data extraction [4][5]. Hence to remove these noise patterns from web page and extract the main content has become a major concern.

Types of Noises

Web noise can be grouped into two categories namely global noise and local noise.
Global noises are large objects which are in no means smaller than the individual pages. Global noise comprises of mirror sites, replica Web pages and antiquated web pages that are to be deleted.
Local Noise also known as intra page noise consists of extra items inside a web page except the main content (intra-page) noise. Such noise comprises banner commercials, navigational guides, garnishing images, etc [6]. Removing local noise from web page is the prime concern for efficient data extraction.
Most of the webpage consist of main content in the middle block hence its location, occupied area of webpage, subject, topic etc also plays a significant role while differentiating it from noisy patterns [7], [8].

Manual noise elimination is very expensive, hence automatic noise elimination has becomes necessary in web data mining. The approaches [10][11][12] on web data extraction uses DOM tree as the mining basis by parsing the DOM of web page and then group the data into objects by tags such as TABLE and DIV.

2. LITERATURE REVIEW:

A technique which was employed to eliminate noise was introduced by Haitao YAO, Zhiyi YIN, Fuxi ZHU and Changsheng GONG [12]. In this method, web pages were processed as images. And then, all of image features were used as criteria to measure noise blocks. As a result, noise blocks and information blocks were distinguished after measuring similarity, and the reduction of noise is realized.

A technique to remove noise was introduced by P. Siva Kumar and R. M. S Parvathi [4] In this approach first, the web page information was divided into various blocks. From which, the duplicate blocks were removed using Simhash. For each block, using three parameters Keyword Redundancy, Linkword Percentage and Titleword Relevancy; the importance of the block was calculated.

Lan Yi et al. [6]. Proposed an approach capturing the general structure and comparable blocks in a group of Web pages, the technique built a compacted structure tree. Then an information based measure was employed to assess the significance of every node in the compacted structure tree. On the basis of the tree and its node significance values, their technique allotted a weight to each word characteristic in its content block. The resultant weights were employed in Web mining. Using two Web mining tasks, namely Web page clustering and Web page classification, the proposed technique was assessed. Experimental outcome revealed that their weighting technique was able to considerably develop the mining results.

A technique which was employed to eliminate noise was introduced by A. K.
For fully-automatically extracting the contents from Web news page, a simple but effective approach, named ECON was proposed by Yan Guo et al. [13]. A Document Object Model (DOM) tree was employed by ECON for representing the Web news page and leverage the extensive characteristics

of the DOM tree. A part of the news content was wrapped firstly by the snippet-node which was found by ECON, and until a summary-node was found it backtracks from the snippet-node, and the summary-node wraps up the entire content of news. Noise was removed by ECON during the process of backtracking. ECON fully satisfied the requirements for scalable extraction and it also attained high accuracy as per the experimental results. Several accepted languages namely Chinese, English, French, German, Italian, Japanese, Portuguese, Russian, Spanish, Arabic Web news pages applied ECON. ECON were employed without difficulty.

3. Proposed Approach

The proposed approach works on the basis of following three phases

Phase 1: Development of webpage in the form of n x 1 (n rows and single column) table.
Phase 2: Convert the webpage into XML format.
Phase 3: Use XSL (with filter feature) display method of XML to extract the data.

Phase 1: Development of webpage in the form of n x 1 (n rows and single column) table

This phase consists of following three steps for development of webpage:

Step 1: In this step every web page to be developed by the web developer uses the format of starting with table (main table) of size n x 1 (n rows and single column) as shown in Fig 1:

```
<Table> (Main Table)
<tr>
<td>
Row 1 DATA (e.g HEADER
Part)
</td>
</tr>
<tr>
<td>
Row 2 DATA (e.g. links and
content of web page Part)
</td>
</tr>
<tr>
<td>
Row 3 DATA (e.g. Footer Part)
</td>
</tr>
</table>
```

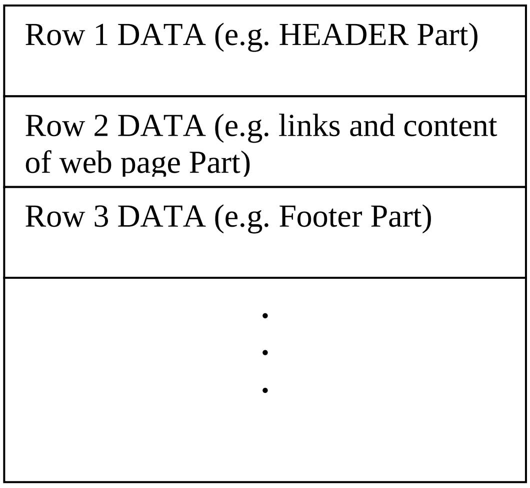

Row 1 DATA (e.g. HEADER Part)
Row 2 DATA (e.g. links and content of web page Part)
Row 3 DATA (e.g. Footer Part)

Fig 1: Format of Webpage having n x 1 Structure

Step 2: Insert the data **required** in the corresponding row of main table created in step 1 in the form of internal tables

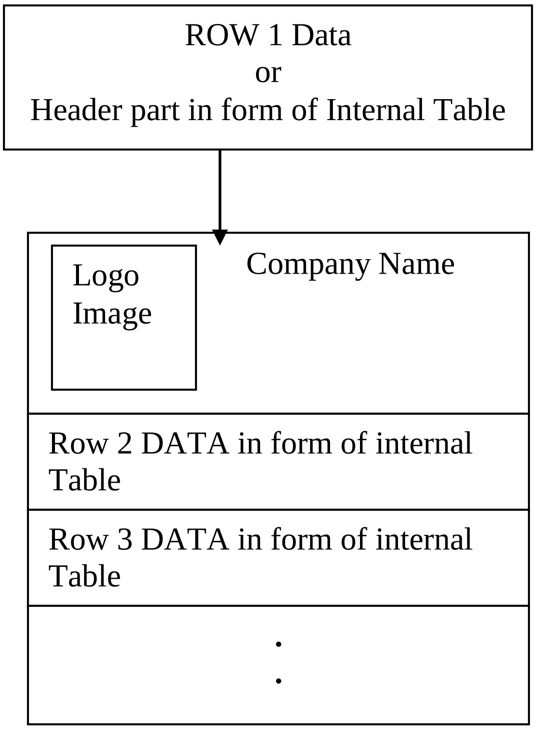

ROW 1 Data
or
Header part in form of Internal Table
Logo Image
Company Name
Row 2 DATA in form of internal Table
Row 3 DATA in form of internal Table

Fig 2: Format of Webpage having n x 1 Structure and Data in form of internal tables

Step 3: In this Step each internal Table is provided an ID (attribute) likewise each website has its first page named index similarly table having main content of webpage will be provided an attribute value of 'content' e.g. <table id=content> and other tables may have <table id=links or footer> .

By using the above approach we have divided the webpage into block structure as shown in fig. 3

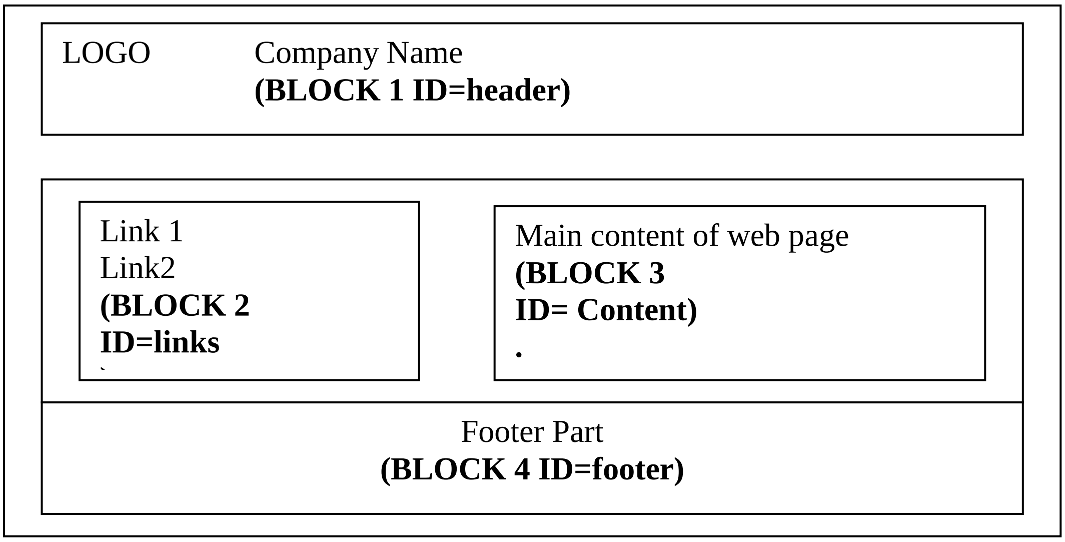


Fig. 3 Block Structure of a webpage

Benefits of using n x 1 approach

1. As the web pages are unstructured in nature, so by using the above approach we can provide at least outer structure to the webpage.
2. No need to create DOM tree as it will be automatically created through this approach hence that time will be saved out.
3. Updating the webpage becomes easy because every row becomes independent of each other i.e. whatever change is required will be made inside the internal tables.
4. Can be easily converted to XML format as XML required a proper format as discussed in Phase 2.

Phase 2: Convert the webpage into XML format.

In this phase the web page structure created through phase 1 is converted to XML (extensible Markup Language) document. As XML required a proper structure the above mentioned structure can be converted to proper structure easily as shown in fig 4.

Phase 3: Use XSLT display method of XML to extract the data.

XSL (eXtensible Stylesheet Language) is used to display XML document. We will use its feature named filter to filter the content of the document from XML document on the basis of ID attribute as shown below in fig 5. This style sheet document will remain same i.e. can be applied to every page.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl"
href="name_of_xsl.xsl"?>
<table> or (maintable)
<tr> or (mtr-main table row)
<td> or (mtd- main table data)
Ist row Data
</td>
</tr>
<tr>
<td>
2nd row Data
</td>
</tr>
. . .
</table> or (/maintable)
```

```
Ist Row Data replaces by

<table><tr><td>(Internal Table)
<id>header or content</id>
<data>
<![CDATA<Table>
<TR>
</td>
<img src=”logo.jpg”>
</td>
<Td>
<font size=”10”>Company
Name</font>
</td>
</TR>
]]></data>
</td></tr>
</table>
```

Fig 4: XML structure of the webpage

```
<xsl: template match="/">
<xsl:apply-templates select="maintable/mtr/mtd/table/tr/td"/>
</xsl:template>
<xsl:template match="mtr/mtd/table/tr/td[id='content']">
<xsl:value-of select="data"/>
</xsl:template>
```

Fig 5: XSL with filter feature

The overall working is shown through the following flowchart:

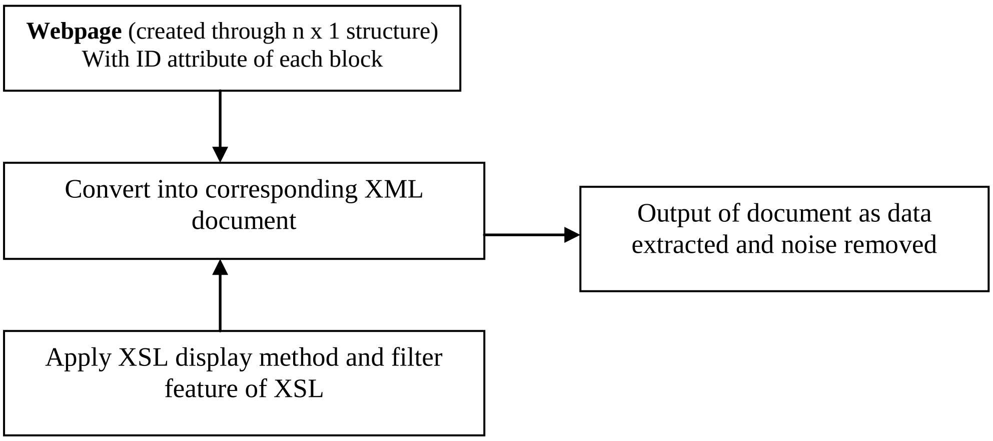
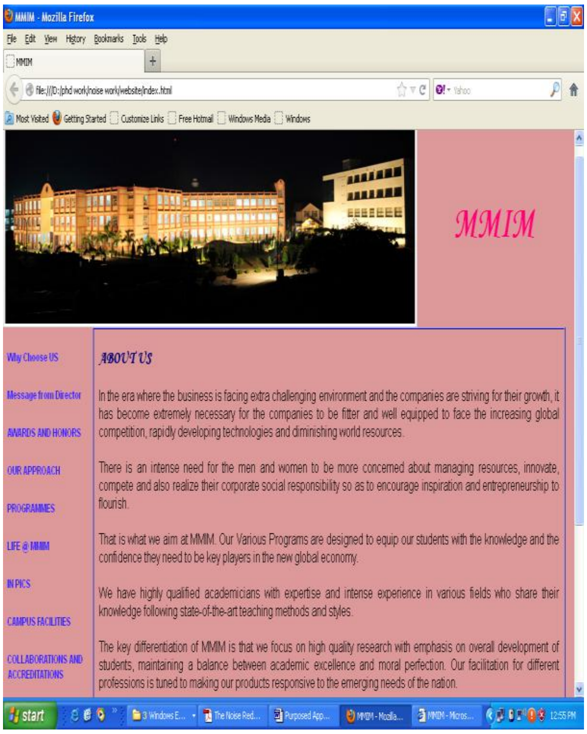

Webpage (created through n x 1 structure)
With ID attribute of each block
Convert into corresponding XML document
Output of document as data extracted and noise removed
Apply XSL display method and filter feature of XSL

Fig 6: Flowchart of proposed Approach

4. Experimental Results and Discussions:

To show the results of the proposed approach we have created a website of M.M.Institute of Management and apply our technique. The main page (index page) of the website is shown in fig 7. Data extracted (i.e. after noise removed) from this web page are shown in fig 8



ABOUT US

In the era where the business is facing extra challenging environment and the companies are striving for their growth, it has become extremely necessary for the companies to be fitter and well equipped to face the increasing global competition, rapidly developing technologies and diminishing world resources.

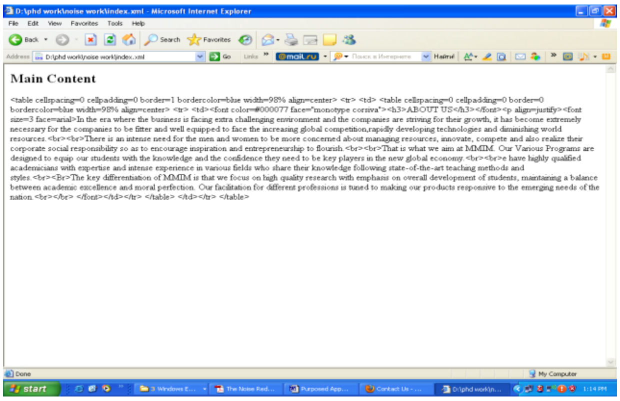
There is an intense need for the men and women to be more concerned about managing resources, innovate, compete and also realize their corporate social responsibility so as to encourage inspiration and entrepreneurship to flourish.

That is what we aim at MMIM. Our Various Programs are designed to equip our students with the knowledge and the confidence they need to be key players in the new global economy.

We have highly qualified academicians with expertise and intense experience in various fields who share their knowledge following state-of-the-art teaching methods and styles.

The key differentiation of MMIM is that we focus on high quality research with emphasis on overall development of students, maintaining a balance between academic excellence and moral perfection. Our facilitation for different professions is tuned to making our products responsive to the emerging needs of the nation.

Fig 7: index page of website


D:\phd work\noise work\index.xml - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address D:\phd work\noise work\index.xml
Go
Links
Done
My Computer
start

Main Content

```
<table cellspacing=0 cellpadding=0 border=1 bordercolor=blue width=98% align=center> <tr> <td> <table cellspacing=0 cellpadding=0 border=0 bordercolor=blue width=98% align=center> <tr> <td><font color=#000077 face="monotype corsiva"><h3>ABOUT US</h3></font><p align=justify><font size=3 face=arial>In the era where the business is facing extra challenging environment and the companies are striving for their growth, it has become extremely necessary for the companies to be fitter and well equipped to face the increasing global competition,rapidly developing technologies and diminishing world resources.<br><br>There is an intense need for the men and women to be more concerned about managing resources, innovate, compete and also realize their corporate social responsibility so as to encourage inspiration and entrepreneurship to flourish.<br><br>That is what we aim at MMIM. Our Various Programs are designed to equip our students with the knowledge and the confidence they need to be key players in the new global economy.<br><br>e have highly qualified academicians with expertise and intense experience in various fields who share their knowledge following state-of-the-art teaching methods and styles.<br><Br>The key differentiation of MMIM is that we focus on high quality research with emphasis on overall development of students, maintaining a balance between academic excellence and moral perfection. Our facilitation for different professions is tuned to making our products responsive to the emerging needs of the nation.<br></br> </font></td></tr> </table> </td></tr> </table>
```

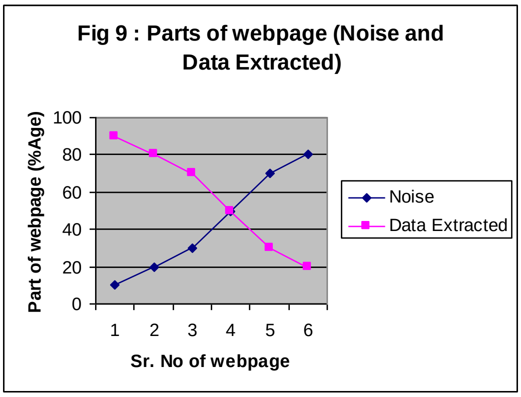
Fig 8: Data extracted from index page

The results for data extraction and noise removed from various pages of website are as shown in Table 1

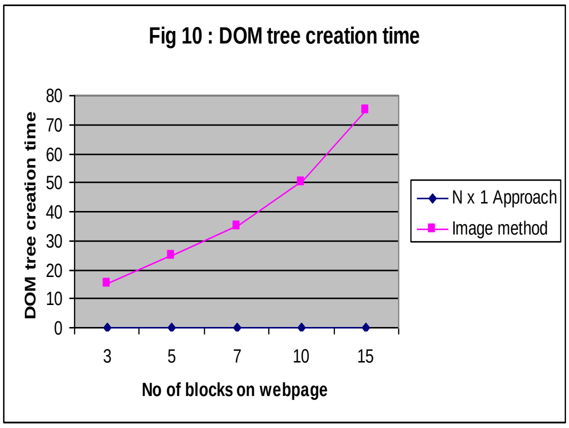
Page Name (.html extension)	**Part of web page removed as noise (%age)**	**Part of webpage extracted as main data(%age)**
Index	52%	48%
why_choose	45%	55%
message	61%	59%
approach	35%	65%
Life	35%	65%
Facilities	32%	68%
collaboration	80%	20%
Search	80%	20%
knowledge	40%	60%
contact	80%	20%

TABLE 1: Noise and Data Extracted from website using n x 1 method

Hence the results of n x 1 approach may be summarize that some area on web page is main content rest is the noise as shown in fig. 9.



As already mention that according to proposed approach there is no need to create DOM tree because this method can provide outer structure to the web page. It means that DOM tree creation time will be saved as was required in noise reduction approach based on image method. To show the results DOM tree creation time for a webpage with n block is n*5 ms (average time for 1 block creation=5ms taken as assumption).
The results are shown below in fig. 10



5. Conclusion:

This paper proposes an approach for noise reduction based on the web page structure. Local noise reduction improves the efficiency of web data extraction. By using this approach we can save time for DOM tree creation which was included in data extraction time in previous approaches. The limitation of this approach is that if some noise is hidden in the content block that noise cannot be automatically detected.

6. References

[1] Deng Cai1, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma, "Extracting Content Structure for Web Pages based on Visual Representation", In Proceedings of the 5th Asia-Pacific Web Conference on Web Technologies and Applications, pp. 406-417, Xian, China, 2003.

[2] G. Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V.Uma, "Signed Approach for Mining Web Content Outliers", World Academy of Science, Engineering and Technology, Vol.56, pp. 820-824, 2009.

[3] Malik Agyemang, Ken Barker and Rada S. Alhajj, "Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams", In Proceedings of the ACM Annual Symposium on Applied Computing, pp. 482-487, New Mexico, March 2005.

[4] P. Sivakumar, R. M. S Parvathi , "An Efficient Approach of Noise Removal from Web Page for Effectual Web Content Mining" European Journal of Scientific Research ISSN 1450-216X Vol.50 No.3 , pp.340-351,2011.

[4] Manisha Marathe, S. H. Patil, G. V. Garje and M. S. Bewoor, "Extracting Content Blocks from Web Pages", International Journal of Recent Trends in Engineering (IJRTE), Vol.2, No.4, pp.62-64, November 2009.

[5] Sandip Debnath, Prasenjit Mitra and C. Lee Giles, "Automatic Extraction of Informative Blocks from Web Pages", In Proceedings of the ACM symposium on applied computing, pp.1722 – 1726, Santa Fe, New Mexico, 2005.

[6] Lan Yi and Bing Liu, "Web Page Cleaning for Web Mining Through Feature Weighting", In Proceedings of the 18th International Joint Conference on Artificial Intelligence,Vol.18, pp.43-50, August 09 - 15, Acapulco, Mexico, 2003.

[7] Ruihua Song, Haifeng Liu, Ji-Rong Wen and Wei-Ying Ma, "Learning Important Models for Web Page Blocks based on Layout and Content Analysis", ACM SIGKDD Explorations Newsletter, Vol. 6, No. 2, pp. 14 - 23, 2004.

[8] Byeong Ho Kang and Yang Sok Kim, "Noise Elimination from The Web Documents By Using URL Paths and Information Redundancy", In Proceedings of the International Conference on Information & Knowledge Engineering, Las Vegas, Nevada, US, pp. 26-29, 2006.

[9] Ye Shiren, Chua Tat-Seng. Detecting and Partitioning Data Objects in Complex Web Pages, Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence[C]. Washington: IEEE Computer Society,pp. 669-672, 2004.

[10] Ye Shiren, Chua. Tat-Seng Learning object models from semistructured Web documents [J]. IEEE Transactions on Knowledge and Data Engineering, pp. 334-339, 2006.

[11] Lin Shian-Hua, Ho Jan-Ming. Discovering informative content blocks from Web documents, Proceedings of the eighth ACM SIGKDD[C]. New York: ACE, pp.588-593, 2002.

[12] Haitao YAO, Zhiyi YIN, Fuxi ZHU and Changsheng GONG "The Noise Reduction Method of Web Pages Based on Image Features International Conference on Computational Intelligence and Software Engineering, pp. 1-5, CiSE 2009.

[13] Yan Guo, Huifeng Tang, Linhai Song, Yu Wang and Guodong Ding, "ECON: An Approach to Extract Content from Web News Page", In Proceedings of the 12th International Asia-Pacific Web Conference (APWEB), pp. 314 – 320, April 06-08, Buscan, Korea, 2010.