Noise Removal from Web pages for Web Content Mining

Yogita K patel[1] ,Mr.Narendrasinh Limbad[2]

[1]Student,Department of Computer Engineering,L. J. I. E. T, Ahmedabad, Gujarat, India

[2]Assistant Professor,Department of Computer Engineering,L. J. I. E. T, Ahmedabad, Gujarat, India

Abstract

Nowadays web technology is getting an emergence importance in day to day life. Everyone is surfing on the web uploading important data on the web. A commercial websites typically contains noisy information blocks with main content, it usually such blocks as navigation panels, copyright, privacy notices and advertisements. There are various researches which are focusing on the extract relevant information from the web pages. This paper we are proposed noise elimination method that uses first outlier detection technique which remove the oulier content and second tag based filtering method implemented by the regular expression which remove the remaining tag from the web pages.This paper presents algoritham for remove the global noise from the web pages.

Keywords-*Web Mining, Outlier detection, Web Pages, DOM Tree, Noise.*

I.Introduction

The Rapid growth of the internet has made the www a popular place for collecting information. A web mining has important task to discover useful knowledge or information from the web. Web mining can be divided in to three categories:web structure mining,web usage mining and web content mining. Web structure mining is the process of discovering hyperlink and document structure information from the web. Web usage mining is the application of data mining techniques for finding interesting and useful usage patterns from web data which makes it more demanding for web based applications. Web content mining is the process of extracting useful information from the contents of web documents[1].

In the World Wide Web Noise on the web pages are not the part of main content and irrelevant information in web pages can affect web mining task.Noises present in web pages can be grouped into two categories according to their granularities[1]:

Local or Intra-page noises: These are noisy information blocks are jumbled with the main contents within a single web page. Local noises include banner and advertisements, navigational guides, some pictures for decoration, etc.

Global or Inter-page noises: These are noises on the web, which are usually no smaller than single web pages. Global noises include mirror web sites, legal or illegal duplicated web pages, some old versioned web pages, etc.[1].

There are several methods are available to segment web pages into blocks. In the DOM based segmentation approach an HTML document is represented as a DOM tree.DOM tree is generally provides a useful structure and better representation for a web page[6]. The DOM tree is hierarchically arranged and can be analyzed in the form of sections or as a whole, providing a wide range of flexibility. By parsing a webpage into a DOM tree, more control can be achieved while eliminating noise data [6].

The remaining of this work in organized as follows:first describe related studies in section 2.Then section 3 describe the proposed architecture for extracting the main content from the web pages.The result of our approach describes in section 4 and finally, we describes conclusion and future work in section 5.

II.RELEATED WORK

Although Web page cleaning is an important task,relatively list of the work has been done in this field.In Structural analysis and Regular Expressions based Noise Elimination from Web pages for Web content Mining [1] **Amit Dutta et al.** proposed the noise elimination method that uses tag based filtering followed by structural analysis of the web page. The system are uses two phases: first Filtering based on Regular Expression and second Structural analysis of the crawled web pages after filtering.This paper focus on detecting and eliminating local or intrapage noises from web pages.

In Noise Elimination from Web page Based on Regular Expressions for Web content mining[2] **Amit Dutta,Dipak Kole,Tanmoy Golui et al**. proposed approach to detect the global noises or inter-pages from web pages. The proposed technique consists of two phases. In the first phase, filtering method based on regular expression is used on web pages to remove noisy HTML tags The filtered document then undergoes to second phase where an entropy based measured is used for removing further noise.

In Mining Contents in web page using Cosine Similarity [3] **Swe Swe Nyein etal.** propose an approach to extract the main content from the web documents. The algorithm based on content structure tree (CST).firstly, proposed system use HTML parser construct DOM tree from construct construct DOM tree from content structure tree which can easily separate the main content blocks from the other blocks. The proposed system introduce cosine similarity measure to which part of tree represent less important and which part of tree represent the more important of the page.

In Elimination of Noisy Information from web page using DOM and Ant Colony Optimization [6] **Shaikh Sakina Banu et al.** Proposed method to eliminate noisy information from web page using DOM tree approach and Ant Colony Optimization to improve the efficiency of mining and also apply neural network algorithm to detect noisy data.

In Automatic News Extraction System for Indian Online News Papers[10] **Dipali B,Sachin Deshmukh et al.**proposed approach for the Indian online newspaper which extract contents from news web databases. The system first **browse** Web pages as per the input URL given by user andNext generate a DOM tree of the news Web page data. And at last, we not only identify and extract valuable news from the Indian news web pages but also remove noisy data. This paper proposed the novel approach for extract data from online Indian newspapers written in the many popular languages such as Marathi, Hindi, Tamil, Gujarati, Kannada, Oriya, Telugu, Punjabi, etc.

III. Proposed Approach

The proposed system is based on analysis of layout as well as actual contents of the web pages given websites for eliminating noisy information. Our method First of all crawled web pages of the same website are considered as input. Than create DOM tree for all crawled web pages and Remove the local noise from web pages. Apply outlier detection Technique to generate sentence wise keyword matrix find. If words are more frequent then remove sentence and detect global noise from web pages. To remove remaining noisy information need to take tag based filtering method based on regular expression. This phase to remove enclosed by predefined negative tag and finally extracts meaningful content from web pages.

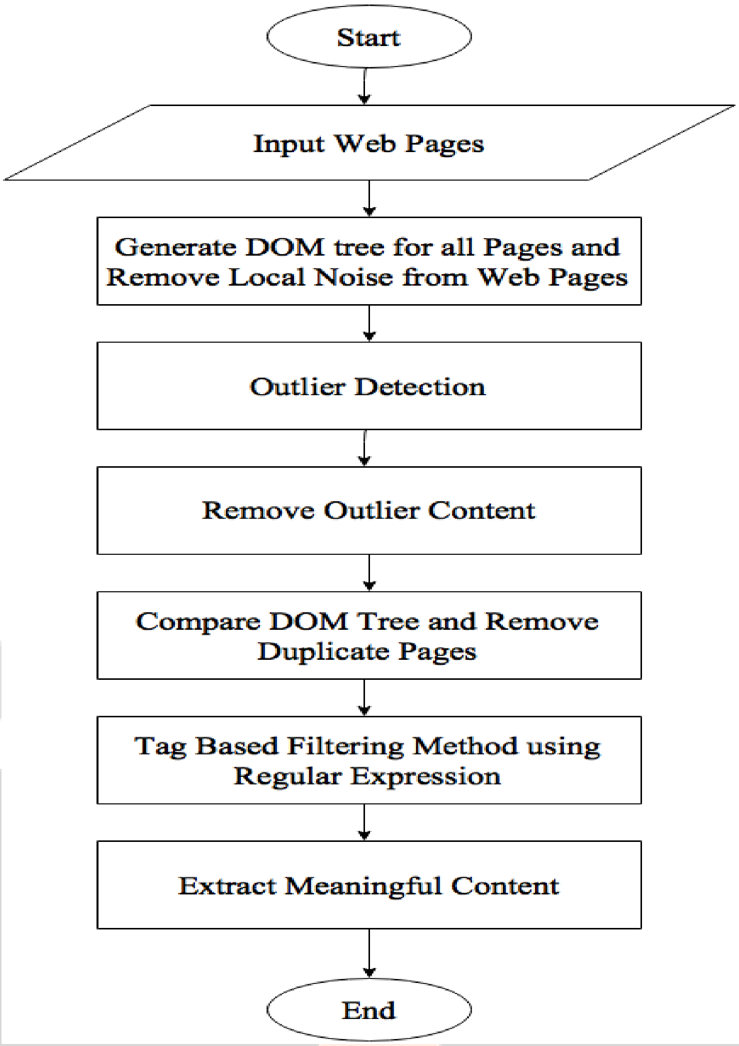
Start
Input Web Pages
Generate DOM tree for all Pages and Remove Local Noise from Web Pages
Outlier Detection
Remove Outlier Content
Compare DOM Tree and Remove Duplicate Pages
Tag Based Filtering Method using Regular Expression
Extract Meaningful Content
End

Fig 1: Proposed architecture

A.DOM Tree

DOM tree is a data structure which is used to represent the layout of a HTML web page. In the DOM tree tags are internal nodes and text, images or hyperlink are leaf nodes. In the DOM tree, each solid rectangle represents an internal node. The shaded boxrepresents the actual content of the node e.g., for the tag IMG, the actual content is src=myimage.jpg. Fig. 1 shows the DOM tree corresponding to the segment of HTML code.

The HTML tag looks like given below:

```
<BODY bgcolor=GREEN>
<TABLE width=650 height=500 >
.
.
</TABLE>
<IMG src="grep.jpg" width=650>
<TABLE bgcolor=BLACK>
.
.
</TABLE>
</BODY>
```

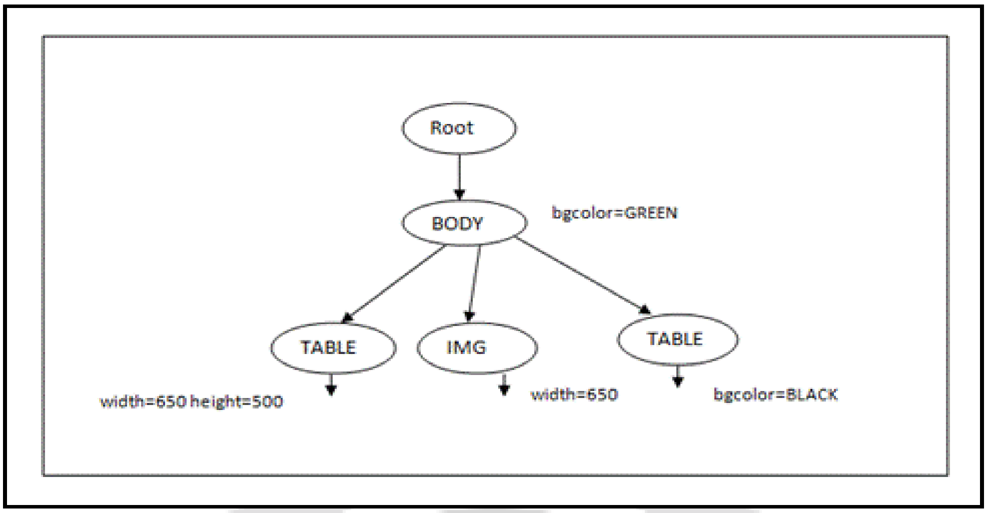

Fig 2: Construction DOM Tree

B. outlier content

Outlier detection is a process in which the noise that are irrelevant to the main content are detected. The Outlier Detection technique is based on document frequency of keywords. These documents are generating a list of words which is stored in a repository. In our proposed approach outlier detection technique apply and generate sentence wise keyword frequency matrix. If words are more frequent then remove sentence.

c. Tag based filtering

Tag Based filtering method implemented by regular expression.. A regular expression is a sequence of characters that forms a search pattern for find and replace operation..most of patterns are created using regular expression to remove contents enclosed by negative tags.Depending on the

content of HTML tags in a web page, the tags can be classified into two types: a) positive tag and b)negative tag ([1]). Positive tag contains useful information in a web page. All the tags except positive tags are referred to as Negative tags. Negative tags usually contain information that are not useful and degrades the performance of web content mining. In this work we have defined the following tags as negative tags to remove noisy information from a web page: Anchor tag (<a>), Style tag (<style>), Link tag (<link>), Script tag (<script>), Comment tag (<!-- … -->) and Noscript tag (<noscript>), Horizontal Ruler (<hr>) and Line Break (<br>).

IV. Experimented Results

We have implemented the above algorithm using Java, in 32-bit Windows 7 Professional with Service Pack 2. The web browser used for the purpose is Google Chrome Version 32.The processor used is Core 2 Duo, 2.00GB
RAM.
For the purpose of experiment we have taken ten popular commercial website. As these websites are dynamic in nature so its contents are varied from time to time. In the Table 1 a comparative analysis is done on how much percentage of noisy element. .In the First Graph how much percentage of noisy element between Base noise and proposed method noise.In the second graph how much percentage of main content between base content and proposed method content.

Page name	Base Noise	Base Content	Proposed Noise	Proposed content	Space complexity(B-P)
http://ijiere.com/Why us.aspx	75%	25%	78.96%	21.04%	228bytes
http://ijiere.com/Call for paper.aspx	57.88%	42.12%	75%	25%	183bytes
http://ijiere.com/aspxAutodetect cookie	57.81%	42.19%	72.93%	27.07%	455bytes
http://ijiere.com/AuthorsGuidelines.aspx	57.84%	42.16%	78.96%	21.04%	945bytes
http://ijiere.com/AuthorsRegistration.aspx	71.71%	28.29%	86.13%	13.87%	831bytes
http://ijiere.com/Check paperstatus.aspx	50.87%	49.13%	76.01%	23.99%	1448bytes

Table 1: Comparative analysis between noise and content

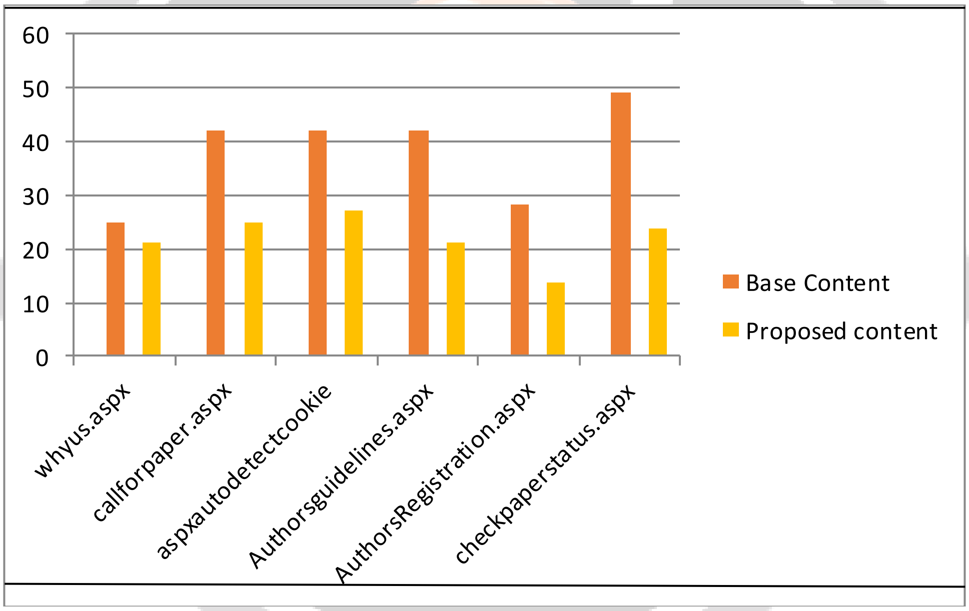

Fig 3: Compare between base content and proposed content

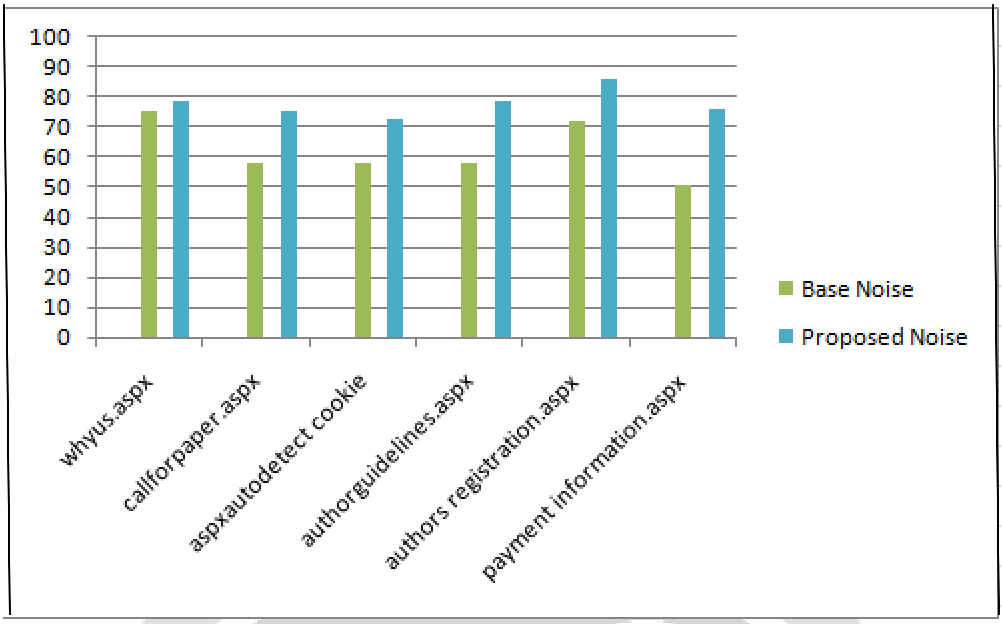

Fig 4: Compare between base Noise and proposed Noise

Conclusion

In recent area there are many research work has been done for noise removal technique. Data mining technique are easily applicable on web content mining. The information present in the local and global noise. The purpose of noise elimination is to improve web content mining. Accurate and effective method to find more relevant document from the web pages. Organizing and removing noise from web pages will get better on correctness of search results as well as explore results. The motivation is to explore new possibilities in improving this area and identifying new ways and methods. So this flow information implies patterns inside, which makes noise removal approach convenient and effective for analysis. DOM tree approach can be an effective solution to remove global noise as it can generate detection method from the flow information.

References

[1]Amit Dutta, Sudipta Paria, Tanmoy Golui and Dipak k. Kole “Structural Analysis and Regular Expressions based Noise Elimination from Web Pages for Web content mining” IEEE 978-1-4799-3080-7/14,PP.1445-1451,2014.

[2]Dutta Amit, Paria Sudipta, Golui Tanmoy and Kole Dipak, “Noise Elimination from Web Page based on Regular Expressions for Web Content Mining” Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014), Volume 27, pp 545-554, June 2014

[3]Swe Swe Nyein,”Mining Contents Web Page Using Cosine Similarity”, IEEE 978-1-61284-840-2/11,pp 472-475,2011.

[4]A. K. Tripathy, A. K. Singh “An Efficient Method Of Eliminating Noisy Information In Web Pages for Data mining” in Proceedings of the Fourth International Conference on Computer and Information Technology (CIT’04) 0-7695-2216-5/04 © 2004 IEEE

[5] Thanda Htwe, Khin Haymar Saw Hla “Noise Removing from Web Pages Using Neural Network”2010 ,978-1-4244-5586-7/10/$26.00 C 2010 IEEE Volume 1.

[6] Shaikh Sakina Banu, Hitesh Kumar Bhatia "Elimination of Noisy Information from Web Page using DOM and Ant Colony Optimization" International Journal of Engineering Research And Technology,Vol.3 Issue 2,PP 1227-1231,feb-2014.

[7] Debina Laishram, Merin Sebastian "Extraction of web news from web pages using a ternary tree approach"Second International Conference on Advances in Computing and Communication EngineeringICACCE, 978-1-4799-1734-1/15 2015 IEEE,PP 628-633, 2013.

[8]C Deepa**,** **"**LBDA: A Novel Framework for extracting content from web pages**"** International Conference on Advanced Computing and Communication Systems (*ICACCS* -2013),IEEE **978-1-4799-3506-2/13 2013 IEEE**

[9] Mingqiu Song, Xintao Wu "Content Extraction from Web Pages Based on
Chinese Punctuation Number" 1-4244-1312-5/07/ 2007 IEEE,PP 5573-5575

[10] Vivek D. Mohod, Dipali B. Gaikwad, Sachin N. Deshmukh ," Automatic News Extraction System for Indian Online News Papers"978-1-4799-6896-1/14/ 2014 IEEE

[11] Satish J. Pusdekar, "Using Visual Clues Concept for Extracting Main Data from Deep Web Pages", International Conference on Electronic Systems, Signal Processing and Computing Technologies 978-1-4799-2102-7/14 2014 IEEE DOI 10.1109/ICESC.2014.39, PP 190-193

[12] Shipra Saini, Hari Mohan Pandey,"**Review on Web Content Mining Techniques**",*International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 18, May 2015*