Journal of Theoretical and Applied Information Technology
30th November 2017. Vol.95. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TAG BASED NOISE REMOVAL FROM WEB PAGES

[1]MAYAJOHN, [2]JAYASUDHA J S

[1]Research Scholar, Department of Computer Science and Engineering, Noorul Islam University, Tamil Nadu,India

[2]Professor, Department of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, Kerala, India

E-mail: [1]maya.j.mail@gmail.com, [2]jayasudhajs@gmail.com

ABSTRACT

Over years the web has evolved as the largest repository of information available to mankind. Web pages have pieces of information which degrade the performance of mining data. The aforesaid type of information is termed as web noise which may be local or global in nature. A tag analysis based technique to eliminate local noises from a web page is proposed in this paper. Irrelevant images and links in a web page can be removed by analyzing the attributes and content of tags. Noisy information is eliminated either by filtering the tags representing noise or by modifying the attributes of tags. The contents in web page which are considered as noise by the proposed work include image advertisements, background images, unimportant link, search panel, copyright information etc. The efficiency in removing image advertisements was analyzed in terms of precision, recall and F-Score. The web pages after noise removal were found to have good compression ratio and showed a significant decrease in load time. As a result of removal of noise tags from web pages the size of source code of the web pages were also decreased considerably.

Keywords: *Advertisements, Block, Noise, Tag, Web page*

1. INTRODUCTION

It is observed that apart from the main content, web pages may contain information which may distract the attention of the user from the main content. The contents which cause distraction include advertisements, search and filtering panel, unwanted images or links etc. These contents may divert the user's attention from the main content of the web page. Wastage of bandwidth occurs due to the presence of advertisements in web pages [1]. In case the bandwidth available is less it is better if the user gets only the main content of the web page that is web page devoid of noise. The two categories of web noise are local noise and global noise [2,3]. Noises within a web page are called as local noise or intra-web page noise. Examples of local noise are unnecessary images, advertisements, links for navigation, copyright notice, privacy notice etc. Global noise or inter-web page noise refers to noise within a website. Prominent global noises are mirrored website, old versions of a web page, duplicate web pages etc. [4]. Web page noises can also be classified into fixed noise, web service noise and navigational assistance noise. Fixed noises are of three types namely decoration noise, statement noise and page description noise. Logos, decorative text or graphics are examples of decoration noise. Different types of statement noises are copyright details, terms and conditions, privacy statements, details regarding sponsors and partners etc. Date, time, visitor count etc. are examples of page description noises. Web pages contain service blocks which help users in performing certain tasks with ease, communicating with the server and receiving reports. Examples of contents of web service blocks include printing current page or email, navigating to another part of web page, weather reports, search panel, sign in forms, rating forms, quiz etc. Navigational assistance noise is classified into two types namely navigation guidance and recommended guidance. Navigation guidance deals with links which aid in exploring various pages in the web site. Recommended guidance guides the users to other websites which deal with either commercial aspects like advertisements, offers, promotions or to web sites which pertain to areas similar to web sites currently being browsed by the user. Document Object Model (DOM) based noise elimination technique is adopted by a majority of the existing techniques. In this paper, a novel method is proposed to identify

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

local noises by analyzing the attribute values and content of specific tags.

The various applications of removing noise from web pages include web page classification, clustering, enhancing the quality of search results, summarization, web mining, cellphone and personal digital assistant (PDA) rendering etc. Web mining is classified into three types namely web structure mining, web content mining and web usage mining. Web usage mining deals with analysis of web usage logs to gather information regarding the usage pattern of different users. Web structure mining deals with identifying the relationship between web pages which may be linked via information or direct link connection. The major steps which may be adopted to retrieve noise free content during web content retrieval are selection of features, assigning feature weights, splitting of blocks, eliminating of duplicate blocks and computation of importance of block. Many methods use DOM trees to identify important blocks in a web page. DOM tree is a good representation of the structure of a web page and is generated corresponding to each web page. The main challenge in noise removal is that if block based noise removal is employed, there is high chance that some useful part of the web page may be removed.

2. RELATED WORKS

Sivakumar and Parvathi [4] proposed a technique to remove primary noises by identifying noise patterns. This method considered only content within tags. The web pages was considered as a collection of blocks or rather collection of tags. Simhash method was employed to identify the duplicate blocks which are to be removed. In Simhash method the keywords in each block are initially identified and the number of times they appear in a block are found both. Corresponding to each block a figure print which is a collection of bit values is generated based on the keywords in the block and their frequency count. Two blocks are deemed to be similar if their fingerprint differ by less than a specific number of bits. The importance of each block was found by using the parameters keyword redundancy, linkword percentage and title word relevancy. Blocks having importance value greater than the threshold value were considered as main content blocks and rest of the blocks were eliminated as they were noise blocks. Although some local web page noises are useful for human viewers and necessary for the website owners, they hinder the process of automated information gathering and web data mining. A major disadvantage of the technique is that only contents in the tag are considered as main content.

Structural analysis and regular expression based method can be employed to remove noise from web pages [5]. This method consists of mainly two steps: tag based filtering based on regular expressions and structural analysis of web page. Based on the information present, HTML tags are classified into positive tags and negative tags [6].The positive tags contain information useful to a web page. Examples of negative tags are <a>, <style>, <link>, <script>, <hr>,
 etc. Patterns created using regular expressions are used to remove contents pertaining to negative tags. This filtering operation removes banner advertisements, images obtained from other websites, mirror sites etc. from web pages. Structural analysis of the web pages crawled from the website are carried out to eliminate navigation panel, menu bar etc. Obviously, the aforesaid noise is usually present in a majority of the pages of a website and it has same content and presentation style. Noise blocks have same content and presentation style in all web pages of a website. Here only some of the global noises are removed from the web page.

Deepa et al. [7] proposed a layout based detachment approach (LBDA) for extracting the main content from web pages. The various steps involved in LBDA are structure analysis, tag tree parsing and block acquiring page segmentation (BAPS). The purpose of structural analysis of the web page is to find the tags accessible in the web page such as child tags and tags over the inner blocks. DOM tree is created based on the XML file which is obtained by converting the web page from HTML format to XML format. The independent tag trees generated corresponding to each web page linked to a website are integrated into a single tree. The content of the web page is extracted by using BAPS technique. The tags that are not closed and the tags that lack child nodes are removed using BAPS technique. LDBA technique is more efficient than VIPS in removing unwanted information from web pages.

Raheja and Katiyar devised a method to reduce noise in web pages based on nx1 table and XSL display [8]. The web pages used in this technique should be in the form of a table having n rows and one column. The data is inserted in the corresponding row in the form of internal table, each of which is assigned an attribute. The internal

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

table representing the main content of a web page has attribute value content, other internal tables may have attribute values like link, header, footer. The web page is converted into XML format. The XML document is displayed using XSL and the filter feature of XSL is used to extract the content of web page. Major drawback of the aforesaid method is that it can be applied only to web pages which are designed in the specific table format.

A combination of case base reasoning (CBR) and neural network [9] can be used to remove noise from a web page. CBR is a machine learning approach used to identify noise patterns. In CBR technique, the past experiences regarding noise patterns are stored in the case base and form the basis for detecting noise. Case base is a collection of cases, where each individual case contains a past experience and its solution. An artificial neural network (ANN) trained using back propagation algorithm is used to classify patterns into three classes namely data class, noise class and a mixture of data and noise. The advantage of using neural network is that it is capable of learning from examples. Apart from that ANN is a highly effective in recognizing patterns and pattern matching. The major fault associated with the above mentioned method is that it is not much effective in removing noise from regions which contain both noise and data.

One of the most prominent types of noise in web pages is an advertisement. Image advertisements can be removed from web pages using HTML tag differentiator technique [1]. DOM structure of web pages are generated and attribute values like filename, alternate text, height, width of the <img> tags present within <a> tags are analyzed. These attribute values are given as input to the image classifier which separates advertisements from other images using rules like domain name difference rule, dimension rule, well-known ad-provider rule and related keyword rule, advertising by scripting, dynamic advertisement rule and flashy plug-in removal rule. The methods involved to find out advertisements should be changed as per changing web page creation methods.

Eldirdiery and Ahmed[10] proposed a text density approach for detecting and eliminating noises present in web pages. The method employed was named as Block Density Based Noise Extractor (BDBNE). In this technique, a web page is divided into blocks which are categorized into valid and invalid blocks. The information present is each block is compared with certain character patterns which help in differentiating between valid and invalid blocks in the web page. Invalid blocks contain a large number of blank characters, symbols etc. and are not further processed. Block text density or the number of words in a block is computed for all valid blocks. ε value is computed by comparing the text density of two neighboring valid blocks. The threshold chosen plays a vital role in detecting noise. Improper selection of threshold value will increase the chance of data block being considered as noise. A block is considered to be noisy if the value of ε is greater than or equal to the threshold and it has less number of lines compared to its neighboring block.

Narwal developed a web page noise elimination system based on the concept of visual block tree [11]. Corresponding to each web page visual block tree is generated using web page segmentation algorithm. Pattern tree which consists of pattern nodes and information nodes is generated from the visual block tree. The pattern trees of different pages of a website are analyzed to compute the values of metrics namely style importance and similarity count. Style importance refers to the number of styles associated with a node in the pattern tree. Similarity count of a node is the number of web pages in which a particular node is present. An important node is considered to have many style patterns. The peculiarity of noisy nodes are that they are present in many web pages of the website. Node importance is calculated based on style importance and similarity count. The value of node importance is normalized using zero mean normalization (z- score normalization). Heuristics rules are applied to remove noisy contents from the web page. The node importance of each node is compared with a threshold value. A node is deemed to be noise if its node importance is less than threshold value. A major pitfall associated with the above mentioned technique is that the heuristic rule that certain noises are positioned in same place in different pages of a website may turn out to be wrong at times.

Featured DOM tree [12] is a variation of DOM tree developed to eliminate noise from web pages. Here apart from presentation style the DOM represents the feature set of individual blocks of web pages. The procedure of noise elimination consists of featuring, modeling and pruning. The main steps involved in featuring phase are discarding the html tags, dividing data into tokens,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

removal of stop words, stemming of tokens and extraction of features. Here different term weighing approach is used to select the optimal subset of features. Modeling is the process of generation of featured weighed DOM tree for each web page. In the pruning phase each featured DOM is examined to check for noisy information. Feature set similarity measure is used to find out noise. The authors have proposed a new technique known as minimum weight overlapping for verifying the similarity. The concept of minimum weight overlapping is used to mark the noisy nodes and the noisy subtree is pruned off. The process of finding noisy nodes starts from the leaf nodes and propagates up the tree. A parent node is considered to be noisy in nature if all its children nodes are noisy.

Image feature based technique can be used to remove noise from web pages [13]. This was a variety method considering the fact that there was wide spread use of DOM trees to identify noise in web pages. In image feature based method the original set of web pages are known as web page template. Web page spider is employed to create a set of pages which are of same layer. These set of pages are known as objective pages. The objective pages are given as input to presentation component which converts the web pages into images. Each image is divided into sub images which are mutually exclusive in nature that is there is no overlapping between content of two sub images. Corresponding to each sub image, features such as color and texture are extracted. The color space employed here is YCbCr. The degree of color similarity between each sub image and its template sub image is computed. The three major steps involved in noise reduction are computation of global similarity degree, determination of sub-image type and identification of noisy information. Bhattacharya coefficients are used to calculate the global similarity measure. Sub images are classified into tentative noise sub images and tentative information sub images. The color similarity degree of a sub image is compared with that of sub images in the same position in other objective pages. A sub image is considered as tentative noise sub image if the similarity degree is large. A template sub image is considered as information sub image if probability of noise sub image is less than preset ratio. A drawback of this method is that the textual contents of web pages are not analysed.

Mehta and Narvekar proposed filters that can be applied to DOM tree to eliminate noise [14]. The system designed by them was capable of extracting content from news websites, blogs, forums and articles. The format followed by news website is neither same as that of blogs nor forums. The contents can be extracted using stored URL list or run time generated URL list. URLs which are to be used frequently are stored in list so that the contents can be retrieved with ease. At times user may not know the URL of information to be fetched. In this case user makes use of search engines like Google, Bing, Yahoo etc. which generates a list of URLs from which the user can fetch data. Here structure classifier is employed to find out DOM tree structure of the whole web page. Filters are used to remove unwanted/ noisy nodes from the DOM tree. The concept of threshold is also employed to extract the necessary contents.

Main text and images can be extracted from web pages using DOM and natural language processing (NLP) [15]. Here firstly DOM tree is generated corresponding to the web page. The HTML tags were classified as style and block tags Tags except div, p, br, li, ul, ol, td, tr, table, h1-6,hr are considered as style tags. It is to be noted that all paragraphs in the main content will not be present in the same level in the DOM tree. Paragraphs are created in web pages using tags like , <p> and
. and <p> are used to embed paragraphs within them but
 tag divides text in paragraphs. All sub trees containing atleast 500 textual characters are considered as potential article text. It is assumed that the main article will be present it the top most part of the web page. Hence the first potential article text is considered as main article. Here the authors assume that article images are embedded in article blocks and article images always have caption. The textual data in the first parent block element of image is considered as image caption. NLP and cosine similarity is used to find the semantic similarity between image caption and article text. The NLP technique used here is Named Entity Recognition (NER). A disadvantage of the proposed method is that statistical methods are not used to optimize the extraction of main content and measurement of similarity.

Dias and Gadge proposed a web page segmentation based method based on key patterns to identify informative blocks in a web page [16]. The HTML tags can be categorized into meaningful tags and less meaningful tags. Here less meaningful tags such as <a>, <b>, <script>, etc. are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

removed from the source code of web page. Sequences are generated from the DOM tree by considering only one level deep child nodes. The key patterns in the sequences are to be identified. Longest and most frequent pattern is known as key pattern. The key patterns are matched with sequences which results in subsequences. Corresponding to each subsequence obtained, a virtual node is added as root node which has subsequence as its children nodes. Since less meaningful tags were removed during the initial stages, the block importance is taken as the count of meaningful tags in a block. Blocks which have block importance less than threshold value are considered as noise blocks and are removed. This technique has good efficiency in terms of space and time.

Oza and Mishra observed that most HTML pages are not well formed [17]. Hence web pages are passed through HTML parser which corrects the markup and generates DOM tree. Linear regression analysis is carried out to find out the relationship between maximum depth of the DOM tree and threshold level. Here on the basis of threshold level DOM tree is partitioned into different sub trees. Nodes of the DOM less than threshold level are considered as noise and are eliminated.

Li and Ezeife proposed a system known as Web Page Cleaner to remove noise from web pages [18]. It is highly essential to clean web pages to improve the accuracy and efficiency of web page content mining. The three major processes involved in the system are extraction of blocks, computing block importance and generation of cleaned files. VIPS page division algorithm is used to extract blocks from a web page. The web page is represented as a combination of DOM tree and visual cues like block location, font size etc. The features associated with each block include PageID, LinkPer, ImLevel, PosLevel, BlockID BlockText, Fingerprint and SimilarLevel. The importance of each block depends upon features such as position of the block, fraction of links in a block and degree of similarity between the block and other blocks. Blocks which have highest block importance values are treated as significant blocks and can be further used for mining data.

The paper discusses how different types of noise present in web pages can be removed by analyzing tags. The paper is organized as follows Section 3 deals with how noises can be identified and removed by analyzing the tags in a web page. The results obtained are presented in Section 4. Section 5 concludes the work carried out.

3. PROPOSED SYSTEM

The proposed system analyses the various tags to identify different types of noise. The different types of noise which can be removed are given below.

1. Background images
2. Advertisement
3. Plug-ins, audio, video
4. Unimportant links
5. Search panel
6. Copyright information

3.1 Removal of background images

Other than providing decoration background images serve no other purpose. They, in turn, increase the time to load web pages. The URL of the background images are specified in tags like td, table, tr, div and style. Background images can be removed by clearing the value of following attributes:

1. background attribute of td, tr, table, body, style tag.
2. background-image attribute of style tag.
3. background-image feature of style attribute of td, tr, table, div.

3.2 Removal of advertisements

Advertisements are usually provided in web pages to divert the attention of viewers towards advertiser's site. They may also be provided to promote certain events or brands. The different types of advertisements include text advertisements, banner advertisements, video advertisements, pop-up advertisements, content sponsoring advertisements, interstitial advertisements etc. Images embedded in anchor tags are examined by inspecting the values of different attributes to determine whether they are advertisements or not. The rules adopted to check for image advertisements are discussed below.

3.2.1 Different Domain Name Rule

If the domain of the src attribute of an image is different from that of the web page then it is more likely that the image may be an advertisement.

3.2.2 Dimension Rule

Image dimensions are a strong cue to differentiate image advertisements from other images. Web banner sizes are "standardized" as outlined by the Interactive Advertising Bureau (IAB). The list of standard web ad banners in exact dimensions of width and height in pixels along with their special names are given in Table 1.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2.3 Popular Advertisement Rule

The contents of a web page which come from well-known ad providers are to be blocked. This is done by checking if the URL belongs to the ad providing website.

3.2.4 Advertisement Related Keywords Rule

Presence of certain words in the name attribute or URL attribute of an image is a strong indicator that the image may be an advertisement. Examples of such words are Ad, free, buy, join, click here, advertisement, Ads, ad, now, click, hits, counter, free, shop, soon etc. This rule is highly effective in identifying advertisements.

3.2.5 Using script tag for advertising

<script> tags may be used to incorporate advertisements in a web page. Javascript codes are clubbed with script tag for the purpose of advertising in web pages. The Presence of advertisements can be found out by checking the SRC attribute of script tag for details pertaining to advertisement providers. Some common terms indicating advertisements present in the SRC are adsbygoogle.js, dcmads.js, googlead, banner, ad, AdSense, AdChoice, pagead2.googlesyndication etc.

Pop-up advertisements can be created using script tag. This can be identified by checking whether the content part of script tag contains terms like window.open, window.showModelessDialog. Advertisements by google, AdSense information, tracking information can be identified from the content part of script tag by the presence of terms like window.adsbygoogle, AddAdSenseService, GA_googleAddSlot, GoogleAnalytics Object.

3.2.6 Dynamic Advertisement Rule

Dynamic advertisements can be inserted into a web page by using the <ins> tag. INS is semantic tag describing something that is inserted to the text after the text was already published. The presence of terms like dcmads, adsbygoogle etc. in the class attribute of ins tag indicate the tag pertains to an advertisement.

3.3 Removal of plug-ins, audio, video

Plug-in is a software component that adds a specific feature to an existing computer program. They are mostly used for audio and video files. The <embed> tag defines a container for an external application or interactive content (a plug-in). The <object> tag defines an embedded object within an HTML document. It is used to embed audio, video, Java applets, ActiveX, PDF and Flash into web pages.

<iframe> tag can be used to embed videos into web pages. This can be identified by checking SRC attribute of The <iframe> for the content /embed. <audio> tag is used to define sounds or music of the file format MP3, Wav, and Ogg. <video> tag is used to specify video clips or video streams with file extension MP4, WebM and Ogg.

3.4 Unimportant Links

Most web pages contain unimportant links which may be avoided. Nowadays majority of web pages contains links to social networking websites like Pinterest, Linkedin, facebook, twitter, google plus which may distract the attention of the viewer from the current web page.

It has been observed that links related to terms, policy, sitemap, disclaimer, the template of website etc. are of no much use and may be neglected. Hyperlinks having href value as # should be removed. href="#" doesn't specify an id name, but does have a corresponding location - the top of the page. Clicking an anchor with href="#" will move the scroll position to the top.

In websites like Wikipedia in a single paragraph, there will be many links in the content part, some of these links may be related to simple information. If a paragraph consists of a large number of links then removal of links reduces navigation from that web page. But caution should be taken for not removing links with the text more, click here etc. as they are important in nature.

3.5 Search Panel

Option for searching information is provided in many web pages. Removing search option from web pages can reduce the number of navigations associated with a web page. It is observed that in a majority of the cases the default content of the text box where search text has to be typed is search and the text associated with the search button is go or search. Hence by checking for the aforesaid values, one can remove search panel.

3.6 Copyright Information

Most web pages contain information related to copyright, rights reserved etc. this information can be categorized as noise as they not part of the information required while mining data. The aforesaid noise may be present in div or td tag and are to be removed.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Standard Web Advertisement Banner Size.

Special name of web banner	Dimensions of web banner
Half banner	234X60
Full banner	468X60
Button 1	120X90
Button 2	120X60
Vertical banner	120X240
Micro button	88X31
Button	80X15
Square	250X250
Square button	125X125
Skyscraper	120X600
Wide skyscraper	160X600
Full banner with navigation bar	392X72
Rectangle	400X260, 180X150
Medium rectangle	300X250
Vertical rectangle	240X400
Large rectangle	336X280
Half square banner	150X150
Head banner	745X100
Leader board	728X90

4. RESULT AND DISCUSSION

Figure 1(a) and 1 (b) represents a web page before and after removal of the background image. Background images are not required to the normal users and hence it can be considered as noise and removed.

Links pertaining to the disclaimer, sitemap etc. are not so important links and may be removed. Information related to copyright is rarely searched during content mining and hence may be eliminated. Figure 2(a) and 2(b) refers to sample web page before and after removal of copyright information and unwanted links.

In the case of sites like Wikipedia, it has been observed that in a paragraph there are many links and these links are corresponding to the words in the paragraph. Many of these words may be familiar to the user but still, users have a tendency to click those links. This diverts the user's attention from the currently viewed web page. Figure 3(a) depicts a web page with a large number of links in a paragraph. In Figure 3(b) the links in a paragraph is removed.

In case the download rate is less such type of user distractions may be eliminated that is if a paragraph has a large number of links then those links should be removed. It is observed that many web pages have provision for searching information. In case the download rate is very slow it would be better if such a provision is curtailed to restrict the user from viewing more web pages. Figure 4 (a) and 4(b) represents a web page with and without search panel.

Removal of noise from the web page reduces the size of the web page and hence the response time becomes less. The efficiency in removing image advertisements can be evaluated using measures like precision, recall and F-score. Precision refers to the ratio of the number of advertisements correctly removed and the total number of advertisements removed. Recall is the ratio of the number of advertisements correctly removed and the number of advertisements actually represent in a web page. F-Score is computed from precision and recall. The efficiency in removing advertisements is shown in Table 2.

Let a be the number of advertisements identified as advertisements,

b be the number of advertisement identified as non-advertisement and

c be the number of non-advertisements identified as advertisements.

$$\text{Precision} = a/(a+c) \quad (1)$$

$$\text{Recall} = a/(a+b) \quad (2)$$

$$\text{F-Score} = (2*\text{Precision}*\text{Recall})/(\text{Precision}+\text{recall}) \quad (3)$$

As a result of filtering of noisy tag information, the size of the source code of the web page gets reduced. The compression ratio of the source code is the ratio of the size of the source code before noise removal to the size of the source code after noise removal. The compression ratio of the source file of various web pages are given in Table 3.

Journal of Theoretical and Applied Information Technology
30th November 2017. Vol.95. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

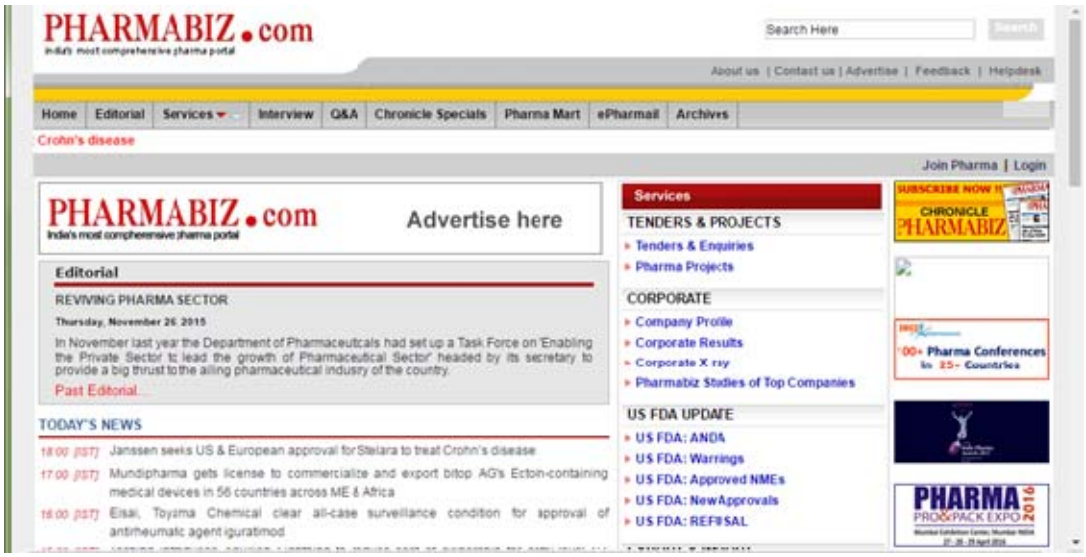

Figure 1(a): Web page with background image

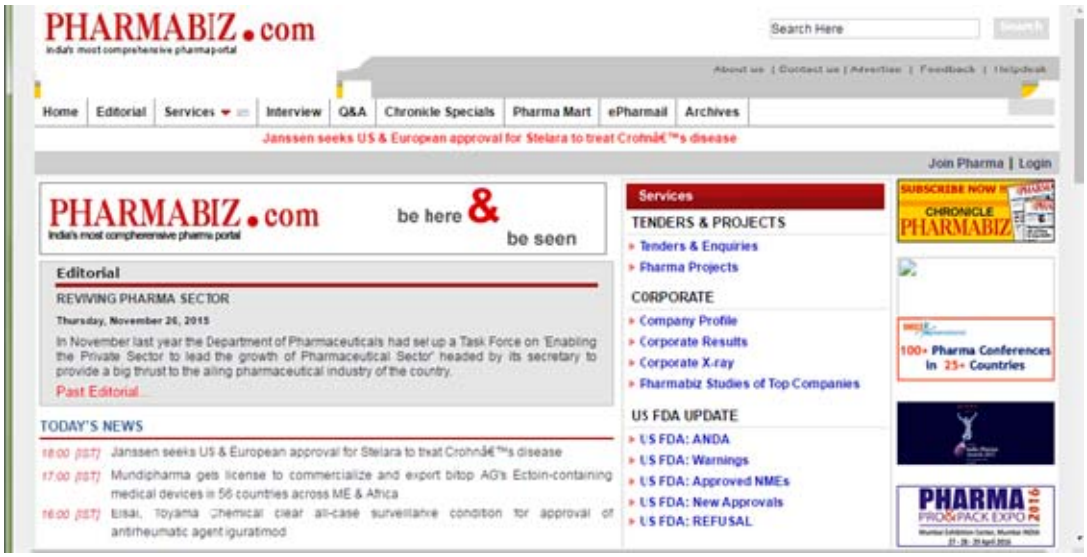

Figure 1(b): Web page with background image removed

Figure 2(a): Web page with unwanted links and copyright information

Journal of Theoretical and Applied Information Technology
30th November 2017. Vol.95. No 22

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 2(B): Web Page With Unwanted Links And Copyright Information Removed

Figure 3(A): Web Page With Large Paragraph Having Many Links

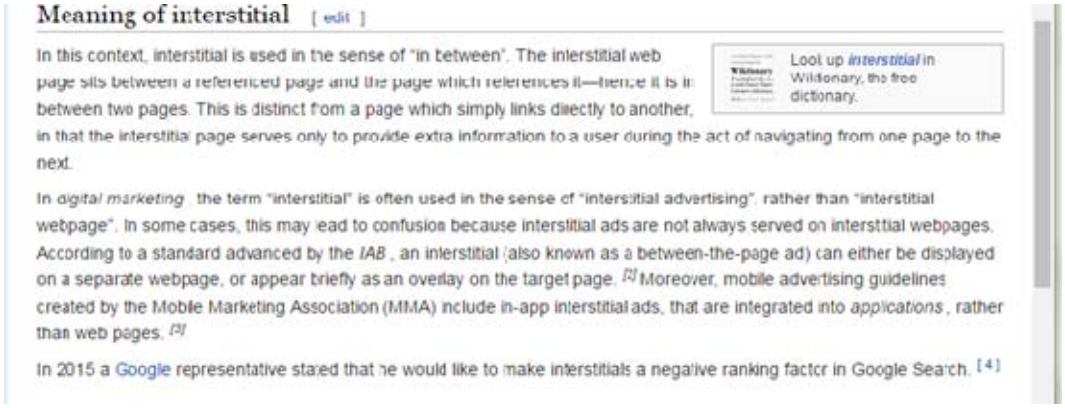

Figure 3(B): Web Page With Large Paragraph Having Links Removed

Figure 4(A): Web Page With Search Panel

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

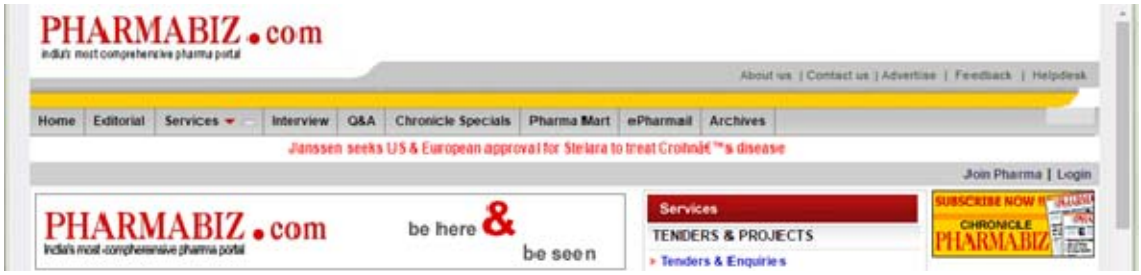

Figure 4(B): Web Page With Search Panel Removed

Table 2: Efficiency In Removing Image Advertisements

Sites	Precision	Recall	F- Score
pharmabiz	0.78	0.78	0.78
myiris	0.67	1	0.80
zdnet	1	1	1
pcmag	1	1	1
yahoo	1	1	1

Table 3: Compression Ratio Of Source File Of Web Page

Sites	Compression ratio
pharmabiz	1.02
myiris	1.08
zdnet	2.07
pcmag	1.53
yahoo	1.06

The compression ratio of a web page is the ratio of the size of a web page before noise removal to the size of a web page after noise removal. The compression ratio of different web pages is given in Table 4.

Table 4: Compression Ratio Of Web Page

Sites	Compression ratio
pharmabiz	1.50
myiris	2.16
zdnet	4.28
pcmag	2.38
yahoo	1.46

Load time of the web page is reduced as a result of noise removal. The decreased percentage of load time of web pages on account of noise removal is given in Table 5.

Table 5: Percentage Decrease In Load Time Of Web Page

Sites	Percentage decrease in load time of web page
pharmabiz	15.6%
myiris	63.43%
zdnet	94.29%
pcmag	31.17%
yahoo	58.22%

5. CONCLUSION

In most of the techniques adopted to remove din from web pages, the web pages are divided into blocks and the noisy blocks are identified. In case the noisy blocks are wrongly chosen then the informative content of the web page will be lost. Majority of the web page noise removal techniques convert web page into DOM tree so that they can be analyzed to identify noises. In this paper, different methods are adopted to remove different types of noise and emphasis is given on analyzing the HTML tags to identify noise. Apart from removing the local web page noises we have considered background image as noise as they usually don't serve any purpose. In this work we have proposed different types of links which may be considered as noise and are to be eliminated. Since in this work we are identifying noise by analyzing the tags, there is no chance of blocks being wrongly removed. This improves the efficiency of web mining operations as no useful blocks are removed. A major benefit of removing noise from web pages is that it helps in efficient and speedy mining of web page contents. A potential application of generating noise free web pages are that they can be presented to the user in case of less bandwidth availability.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES:

[1] J. Gadge and H. R. Parmar, "Removal of Image Advertisement from Web Page," *International Journal of Computer Applications*, Vol. 27, No. 7, 2011, pp. 1–5.

[2] L. Yi, B. Liu and X. Li, "Eliminating Noisy Information in Web Pages for Data Mining," *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2003.

[3] 3.L. Yi and B. Liu, Web Mining Through Feature Weighting," *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence*, 2003.

[4]P. Sivakumar and R. M. S. Parvathi, "An Efficient Approach of Noise Removal from Web Page for Effectual Web Content Mining," European Journal of Scientific Research, Vol. 50, No. 3, 2011, pp. 345–356.

[5] D. K. Kole, A. Dutta, S. Paria and T. Golui, "Structural Analysis and Regular Expressions based Noise Elimination from Web Pages for Web Content Mining," *Proceedings of IEEE International Conference on Advances in Computing, Communications and Informatics*, New Delhi, India, 2014, pp. 1445–1451.

[6] B. H. Kang and Y. S. Kim, "Noise Elimination from the Web Documents by Using URL paths and Information Redundancy," *Proceedings of International Conference on Information & Knowledge Engineering*, Las Vegas, Nevada, US, 2006, pp. 26–29.

[7] C. Deepa and A.S. Vijendran, "LBDA : a Novel Framework for Extracting Content from web pages," *Proceedings of IEEE International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, 2013.

[8] N. Raheja and V. K. Katiyar, "A Noise Reduction Approach based on n x 1 Table and XSL Display Method for Efficient Web Data Extraction," *International Journal of Computer Applications*, Vol. 64, No. 11, 2013, pp. 1–6.

[9] T. Htwe and K. H. S. Hla, "Noise removing from Web pages using neural network," *Proceedings of 2nd International Conference on Computer and Automation Engineering*, 2010, pp. 281–285.

[10] H. F. Eldirdiery and A. H. Ahmed, "Detecting and Removing Noisy Data on Web Document using Text Density Approach," *International Journal of Computer Applications*, Vol. 112, No. 5, 2015, pp. 32–36.

[11] N. Narwal, "Improving Web Data Extraction by Noise Removal," *Proceedings of IET 5th International Conference on Advances in Recent Technologies in Communication and Computing*, 203, pp. 388-394.

[12] S. N. Das, M. Mathew and P.K. Vijayaraghavan, "Eliminating Noisy Information in Web Pages using featured DOM tree," *International Journal of Applied Information Systems*, Vol. 2, No. 2, 2012, pp. 27–34.

[13] H.Yao, Z. Yin, F. Zhu and C. Gong, "The Noise Reduction Method of Web Pages Based On Image Features," *Proceedings of IEEE International Conference on Computational Intelligence and Software Engineering*, 2010, pp. 1-5.

[14] B. Mehta and M. Narvekar, "DOM Tree based approach for Web Content Extraction," *Proceedings of IEEE International conference on Communication, Information and Computing Technology*, Mumbai, India, 2015.

[15] P. M. Joshi and S. Liu, "Web document text and images extraction using DOM analysis and natural language processing," *Proceedings of 9th ACM symposium on Document Engineering*, 2009, pp. 218-221.

[16] S. Dias and J. Gadge, "Identifying Informative Web Content Blocks using Web Page Segmentation," *International Journal of Applied Information Systems*, Vol. 7, No. 1, 2014, pp. 37–41.

[17] A.K. Oza and S. Mishra, "Elimination of noisy information from web pages," *International Journal of Recent Technology and Engineering*, Vol. 2, No. 1, 2013, pp. 115-117.

[18] J. Li and C. I. Ezeife, "Cleaning Web pages for Effective Web Page Content Mining", Database and Expert Systems Applications, DEXA 2006. *Lecture Notes in Computer Science*, Vol. 4080.