© 2017 IJSRST | Volume 3 | Issue 7 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Web Page Noise Removal - A Survey

Dr. S. Vijayarani[1], K.Geethanjali[2]
[1]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India
[2]M.Phil Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT

Web mining is used to extract useful information from websites which includes web documents and hyperlinks of web sites. The World Wide Website contains a wide range of web pages which are very useful to many users. Web pages are composed of different kinds of data, such as text, audio, video and images. In addition to this, nowadays, web pages contain a large amount of unnecessary data, e.g., advertisement posters, navigation bars and disclaimer/copyright notices. These types of unnecessary data are called as noisy data. This has created the distractions to the user and also increases the time to perform searches and browsing tasks. To perform in-depth analysis of web data or web content mining, the first and essential step is to remove the noises which are existing in the web pages, and then we can extract useful information from the web pages. Removing noise from the web page is challenging task in web content mining. This main objective of this paper is to discuss the basics of web content mining, types of noises, techniques used for noise removal and different models used in the literature.

Keywords : Web Content, Web page, Global Noise, Local Noise, Filtering.

I. INTRODUCTION

Web mining is used to extract knowledge from web data. Web mining is classified into three main categories, i.e. Web content mining, Web structure mining and Web usage mining data. Web content mining is used to mine data from the content of web pages. Web pages consist of text, graphics, tables, data blocks and data records [1]. Web Content Mining uses the ideas and principles of data mining and knowledge discovery process. Web usage mining is also known as web log mining, which is used to analyze the behavior of website users. It can be used to predict the user behavior while the user interacts with the web. Web structure mining is based on the link structures. It can be used to categorize web pages and is useful to generate information such as similarity and relationship between different websites.

Extracting the useful information from web pages becomes essential task. The web page is a medium for accessing the information from different sources. Extracting the information from various resources has many problems like finding the useful information, extracting the knowledge from large data set and learning about individual users. To resolve these problems various methods and techniques are developed. The information technology field has a massive amount of data that needs to transform or extract into useful information. This extracted information can be used for several applications. To extract the useful information there are different kinds of algorithms and techniques are available for different types of data.

Web content mining includes various kinds of data such as: image, audio, video and text. In web mining web documents can be divided into three kinds namely core information, redundant information and hidden information [13]. Web documents also comprise “hidden information” like HTML tags, script language and programming comments, which is called ‘hidden information’. The repeated data in web documents are called as redundant information. The main content or information of the web page like, news article are known as the core information.

In a web mining system, the input data moves through the three different stages to reach its final result: namely preprocessing, data mining and post processing [2]. Pre-processing may include removing attributes that are irrelevant and cleaning the data from noisy information. Data mining is a generic term that includes the techniques and tools used to extract useful information

from big databases. Post processing is used for visualization technologies have recently been used in steering computation, in aiding directed the analysis, in query interfaces to complex multimedia databases, and in information presentation and navigation. Summarization is closely related to compression, machine learning, and data mining [3]. These noises are like banner commercials, navigational guides, garnishing images, etc.

A. Web Content Mining

Web mining has an important mission to discover useful knowledge or extract information from the web. Generally web mining is classified into three categories, they are, web content mining, web structure mining and web usage mining [4]. Web content mining is the process of extracting the knowledge from the web.

B. Unstructured Mining

Generally, web document has unstructured data. The unstructured data deals with data mining techniques which include knowledge discovery in text called text mining or text data mining. To provide effective available results, preprocessing steps for any unstructured data is processed by information extraction, text categorization, and/or applying NLP techniques [6]. The unstructured mining includes the following techniques.

1) Information Extraction
2) Topic Tracking
3) Summarization
4) Categorization
5) Clustering
6) Information Visualization

1) Information Extraction

To extract information from unstructured data i.e. web data using pattern matching algorithms and techniques. It focuses on the keywords and phrases and then finds out the connection between the text. Information extraction transforms unstructured text into a structured form [8]. Information extraction can be provided to the KDD module because the information extraction has to transform unstructured text into more structured data. First the information is mined from the extracted data and then using various kinds of rules [5].

2) Topic Tracking

Topic tracking checks the documents which are viewed by the user and completely analyze the user profile. Yahoo! Uses topic tracking. If the user enters the keyword to the web it shows the results which are related to the keyword. The advertisements that are displayed whenever the user enter into login. Based on the mail subject, users receive the advertisements. Topic tracking can be beneficial in the field of medicine and research also. Any advancement done anywhere in the world can be notified to the registered user. Individuals in the field of education could also use topic tracking which is to be useful, because, they have the latest references for research in their area of interest [7].

3) Summarization

Summarization is a technique which is used to reduce the length of the document by maintaining the important points of the document. It helps the user to make a decision whether to read the topic or not. The summarization technique uses two methods that are the extractive method and the abstractive method. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. This kind of summary might contain words, not explicitly present in the original document [5].

4) Categorization

This technique identifies the main theme by placing the documents in a predefined set of group. The technique counts the number of words in the document. According to the topic the rank is allocated to the document [7]. Categorization is the major technique for classifying the document. First it counts the number of words occurs in a document, then it to take the decision from the words count. It ranks the document according to the topic [9]. Documents having majority content on a particular topic are ranked first.

5) Clustering

This technique is used to group similar documents. Here, grouping of documents is not done on the basis of predefined topics. Some documents may appear in different group. As a result useful documents are not omitted from search results. This technique helps the users to select the topic of interest. It is very difficult to find out the relevant information from large unstructured document collection. To categorize the documents, categorization technique is used. In some situation, the same document can appear in different groups. The problem of finding the best group is done by clustering [10]. There are various clustering algorithms available which can help the user to easily select the topic of interest from the best relevant grouping.

6) Information Visualization

Visualization utilizes feature extraction and key term indexing. Documents having similarity are found out through visualization. Large textual materials are represented as visual maps or hierarchy where the browsing facility is allowed. It helps for visually analyzing the content. The user can interact by scaling, zooming and creating sub maps of the graphs [9].

C. Structured Data Mining

Data in the form of list, tables and tree is structured data. Structured data is easier to extract while compared to unstructured data. It is on the web are often very significant because it represents the host pages. The following techniques are used in the structured data.

1) Web Crawler
2) Page Content Mining
3) Wrapper Generation

1) Web Crawler

Computer programs are called as Crawlers. It travels in the hypertext structure in the web. Web crawlers can be used by anyone to collect information from the web. Search engines can use two types of crawlers. They are internal and external web crawler [10]. Internal web crawler crawls through internal pages of the website and the external crawler crawls through unknown websites. There are several uses of the program, perhaps the most popular being search engines using it to provide web surfers with relevant websites. Other users include linguists and market researchers, or anyone trying to search information from the Internet in an organized manner. Alternative names for a web crawler include web spider, web robot, web crawler, and automatic indexer [11].

2) Page Content Mining

The technique of page content mining is used to extract structured data which works on the pages that are ranked by the traditional search engines. The pages are classified by comparing the page content rank. [4]

3) Wrapper Generation

The information is provided by the wrapper generator with the capability of sources. Web pages are ranked by traditional search engines. By using the page rank value the web pages are retrieved according to the query. To facilitate effective search on the World Wide Web several Meta Search Engines have been formed which do not do the search themselves, but take help of the available search engines to find the required information. Meta Search Engines are connected to search engines by the means of Wrappers.

D. Semi-Structured Data Mining

1) Object Exchange Model

The relevant information is extracted from semi-structured and is collected in a group of useful information and then it is stored in the OEM (Object Exchange Model). This helps the user to accurately understand the structure of the information that is available on web [11].

2) Top down Extraction

This technique helps in extracting complex objects from a rich web sources and decompose them into less complex objects until atomic objects have been extracted.

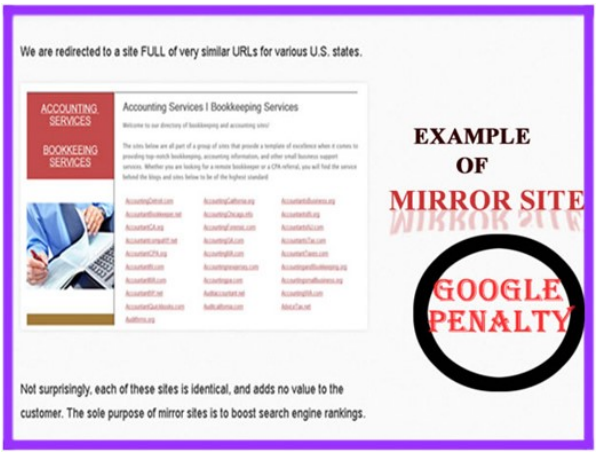
3) Web Data Extraction Language

This technique helps in converting web data to structured data and then delivers this data to end users.

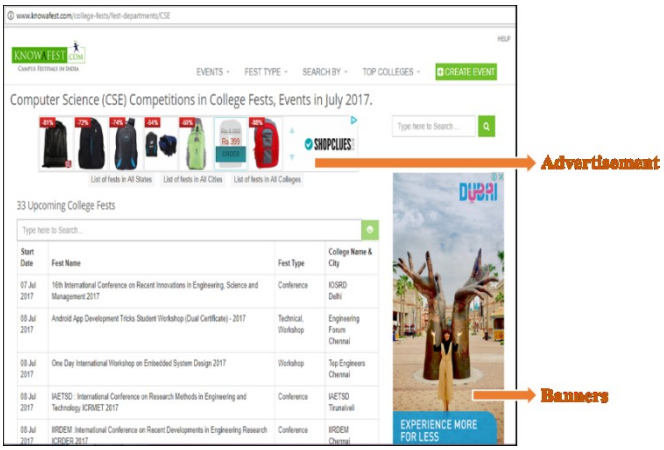
IV. TYPES OF NOISE

There are two types of noises in the web pages. They are global noise and local noise.

Global Noise: Global noises are replicates of the web pages over the internet i.e. mirror websites and legal or illegally duplicated web pages. These are considered as noises which are usually no smaller than individual pages [4].



Local Noise: The intra-page redundancies exist in the web pages are known as local noise. These are noisy regions/items/pattern/blocks within a web page of the website. Local noises are usually jumbled with web pages main contents. It includes banner ads, copyright notices navigational guides, decoration pictures, etc. and these noises should be removed using web mining. [5]



V. RELATED WORKS

Authors Name	Title	Description	Techniques	Inference
Hassan F.Eldirdiery, A.H.Ahmed - 2015	Detecting and Removing noisy data on web document using text density approach	This work presented BDBNE method that detect and remove the noisy data from the web document. This method is simple and works only on a single web page.	BDBNE (Block Density Based Noise Extractor)	The authors proposed the algorithm for validated against the consistency and accuracy measures. To evaluate consistency which means to test BDBNE if it has the ability to segment a web document into blocks and extract the noisy blocks from all blocks. The test is done by using a dataset of an electronic newspaper called Sudan Vision Daily.

Nikolaos pappes, Georgios Katsimpras, Efstathios Stamatatos - 2009	Extracting informative textual parts from web pages containing user generated content	This method has been applied two basic pre-processing steps, namely web page segmentation and noise removal. The proposed SD Algorithm combines visual and non-visual characteristics of web pages. Able to identify the type of web pages according to the properties of the detected regions.	SD Algorithm	The proposed algorithm evaluates its effectiveness on the extraction of informative textual parts from web pages.
S.S.Bhamare, Dr.B.V.Pawer - 2013	Survey on web page noise cleaning for web mining	The author discussed about techniques for web page cleaning and its methods and approaches.	Classification Based Cleaning method; a segmentation based cleaning method, a template based cleaning method, SST based cleaning technique, Feature weighting based cleaning method	The survey of web page noise cleaning for web mining is a newly proposed topic. Web page cleaning is comes under data preprocessing which is used to improve search results by filtering irrelevant or useless information.
Sandeep kaur, Abhishek Tyagi - 2014	Noise Reduction and Content Extraction from web pages using DOM based page segmentation	A DOM based segmentation method is proposed for noise reduction and extraction of web content from web pages. The navigational bar, Home page short description noise is removed using DOM based page segmentation which converts the web pages into blocks and regions removes the noise and extract the information based on regions and blocks.	DOM based page segmentation	A DOM Based Page Segmentation method is proposed for noise reduction and extraction of Web content from Web Pages. The navigational bar, Home page and short description noise are removed using Dom based page segmentation which converts the Web Pages into blocks and regions removes the noise and extract the information based on regions and blocks.
Ypgita K patel, MR. Narendrasinh Limbad - 2016	Noise removal from web pages for web content mining	The author reviewed the DOM tree approach that can be an effective solution to remove global noise.	DOM tree approach	Removing noise from the web pages can be improved the search results as well as explore the results. So this flow of information implies patterns inside, which makes noise removal approach convenient and effective for analysis.

Surabhi Lingwal - 2013	Noise reduction and content retrieval from web pages	The proposed content Extractor method applies a robust noise measure to discover templates. The experiments conducted shows that the algorithms are able to detect outlier with high accuracy in websites. The experiments also show that removing noisy information, i.e. templates can improve the accuracy of web mining tasks and content retrieval.	The content extraction and outliers detection technique	The proposed Content Extractor applies a robust noise measure to discover templates. The experiments conducted shows that the algorithms are able to detect outlier with high accuracy in websites. The experiments also show that removing noisy information, i.e., templates can improve the accuracy of Web mining tasks and content retrieval.
HUI Xiong, Gaurav pandey, michale Steinbach, Vipin kumar. - 2010	Enhancing Data Analysis with Noise Removal	The goal of the work presented in this paper is to boost the quality of data analysis, and capture the underlying patterns in the data reducing the effect of noise at the data analysis stage.	HCLEANER: A HYPERCLIQUE-BASED DATA CLEANER	Experimental study to compared HCleaner, CCleaner, and the two previously discussed noise removal techniques derived from the LOF and distance based outlier detection algorithms. Specifically, authors used these four techniques to remove the increasing amount of noise from the data and then applied clustering and association analysis to the cleaned data. These results were evaluated by measuring the entropy and the F-measure of the resulting clusters.
Michal Marek, Pavel Pecinal, Miroslav Spoustal – 2007	Web page cleaning with conditional Random Fields	The proposed method for web page cleaning based on sequence labelling with conditional Random fields and presented a few initiated experiments evaluated on the development data for Cleaneval 2007.	Sequence labeling with Conditional Random Fields	The proposed method for web page cleaning is based on sequence labeling with Conditional Random Fields and presented a few initial experiments evaluated on the development data for Cleaneval2007.
Anchal Garg, Bikrampal Kaur – 2014	Webpage Performance Enhancement by Removing Noise	This proposed algorithm aims to improve performance by using new technique; Least Recently Used and its variants. By using this algorithm, they found the least used links affects the performance of web pages.	Least recent Page Replacement Algorithm	The purpose of this proposed method is to eliminate noise and to increase the efficiency of web pages. Authors checked the complexity of both algorithms, i.e. DOM tree and LRU algorithm. A comparison is made between both algorithms; worked to remove the non- required advertisement. LRU algorithm will work faster than the DOM tree. LRU will give us the least used advertisement. This works by maintaining a linked list in the memory, with the most recently used page at the front and the least recently used page at the rear. Complexity of LRU algorithm is less than the DOM tree algorithm, which makes the LRU algorithm better.

VI. TECHNIQUES USED FOR NOISE REMOVAL

There are many techniques developed for identifying and removing the noise from the web pages. Mostly noise removal techniques are based on the DOM tree approach. The commonly used noise removal techniques are natural language processing (NLP), filters, artificial neural network (ANN), etc. along with a DOM tree representation [12]. Other major techniques are based on structural analysis and regular expression, layout based detachment approach (LBDA), pattern trees and heuristic based method, text density approach, image features, web page segmentation, featured DOM tree, etc.

A. Structural Analysis and Tag based Filtering on Regular Expression Technique

This technique used for removing the noise from web pages. In HTML, based on the content the tags are categorized into positive tags and negative tags [7]. The contents of the positive tag are retrieved from the useful part of web pages while negative tags contain includes <a>, <style>, <link>, <script>, <hr>,
 etc. The regular expression is used to remove the negative tags. After removing the negative tags, the web page is devoid of banner advertisements, images obtained from other websites like Amazon, mirror sites etc. In order to remove noise like navigation panel, menu bar etc.

B. Layout Based Detachment Approach (LBDA)

Layout Based Detachment Approach (LDBA) technique is used for extracting the content from web pages. Structural analysis, tag tree parsing, block acquiring page segmentation and content extraction also involved with LDBA. Structural analysis is applied to the web page for finding out the tags accessible in the web page. In this technique web page, i.e HTML format is converted into XML format. The DOM tree is created based on the XML file generated [14]. The HTML parser generates independent tag tree corresponding to each web page linked to a website. The tag trees are then incorporated into a single tree. The unnecessary tags include tags that are not closed and tags that will lack child node. The unwanted tags are removed using BAPS technique. After the data extraction, boundary of blocks is eliminated to get the required information.

C. Based on Visual Block Tree

The visual block tree is built for each web page using a page segmentation algorithm. Based on the visual block tree used to generate pattern tree which consists of a pattern node and information node. Each pattern node stores the information about the layout and order of nodes at a given level. Noise elements are differentiated from the main content by applying the following rules:
i) Important nodes have a different presentation styles.
ii) Distinctive pattern nodes at specified level through web pages are deemed to be important.
iii) Noisy content of a website shows consistency across web pages of that website.

To identify the significant content and noise element of a web page have two similarity measures, namely style importance metric and similarity count metric. Style importance metric measures includes the number of styles applied to a given node. Similarity count metric specifies the number of times the nodes are repeated across web pages.

D. Based on nX1 Table and XSL Display

The nX1 table and XSL display method are used to reduce the noises in the web pages. In this method the web pages are developed in the table format and the table having n rows and one column. The data is inserted into the following row in the form of the internal table. Each internal table is assigned as an attribute. The main content of the web page has an attribute value content and other internal tables may have attribute values like link, header and footer. The web page i.e. HTML format is converted into XML format. The XML document is displayed using XSL and the filter feature of XSL which is used to extract the content of web page [12].

E. Text Density Approach

The method of text density approach is used for discovering and eliminating the noisy information which is presented in the web pages. The HTML page is used as an input for the text density approach. The given web pages will divide into multiple blocks. Here, this approach is considering two types of blocks. One is valid blocks and another one is invalid blocks. After

dividing the multiple blocks only the valid blocks are carried out for further process. Blocks which contain a large number of blank characters, symbols, etc. are considered as invalid blocks. Noisy blocks are distinguished from non-noisy blocks using a threshold value which is calculated automatically. Text density of each block is calculated. Text density denotes to the number of words in a block. The neighboring blocks are compared by using the value which is computed based on text density. The value is compared with the threshold value. If the value is greater than or equal to the threshold value, then block having less number of lines is considered as noisy block. In this method a single web page is required to find noisy content of a web page and there is no need for generating a DOM representation of the web page. [7]

F. Noise reduction by removing advertisements

Advertisement is an example of noisy content of a web page. In the HTML tag the differentiator technique is used to remove image advertisement from web pages. Popular advertisement types of web pages are banner advertisement, video, and pop up, etc. In this technique DOM tree structure of the web page is first generated and image tag <img> inside anchor tags <a> are analyzed i.e. name of the image file, alternate text, aspect ratio etc. are gathered. The relevant features are collected to a given input of predefined rule based image classifier. The rules used by a classifier which include 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. These rules are differ from other web page contents. [7]

G. Case Base Reasoning (CBR) using neural network

The Case Base Reasoning (CBR) and neural networks are used to eliminate the noisy information in the web pages. There are most of the noise patterns are structured by using tags like <TABLE>, <DIV>, <FRAMESET>, <SELECT>, <INPUT> etc. In this method, the CBR is used to identify the noise patterns. The CBR is a machine learning approach which makes use of past experiences to solve upcoming problems. The past experiences are stored as cases and form the basis for taking decisions. Each case consists of past experience and also solutions. The case base is a collection of different cases. The different noise patterns in the websites are identified and they are stored as case in the form of DOM tree structure. Before the process, the web pages are converted into well-formed document form. According to the threshold value the DOM tree structure of the web page is divided into number of sub trees. Then the case base, searched for similar existing noise pattern. Artificial Neural Network (ANN) is used for matching the given pattern. The sub trees are converted into standardized numeric representation by using the equation (1)

$$X_i=S_n/T_n \quad (1)$$

Where X_i represents the input nodes at input layer, S_n is the number of occurrences of leaf nodes in sub-tree and T_n represents the total number of leaf nodes in sub tree. There are three classes of information, namely data class, noise class and mixture of data and noise class. Artificial Neural Network is trained using back propagation algorithm. Here the occurrence of noise pattern is modeled using standard sigmoid activation functions [5].

H. Identifying Informative Web Content Using Web Page Segmentation

The web page segmentation is a technique to identify the important content of web page and eliminate noise. A set of web pages is the input for the web page segmentation method. The web page is preprocessed by removing the noisy tags such as <a>, <b>, <script>, , comments, etc. Then the web page is represented as a DOM tree. Using only one depth child nodes, a sequence is made from the DOM tree. The key pattern in the sequence is found out. Key pattern denotes to the longest and most frequent repetitive pattern. The key patterns are matched with the sequence and then subsequences are found. Corresponding to the matching subsequence, virtual node is created as root node. The child nodes of the virtual node are the components of the subsequence. The significance of each block is computed by counting the number of significant tags in each block. If the block is importance of a given block is less than the threshold value, then the block is considered to be noisy and is removed. [5]

I. Removing Noise And Duplicate Contents

Removal of primary noises, duplicate content and noisy information is done according to the block importance of the web pages. The primary noise have copyright information, privacy notice, navigation bar, advertisements and etc. Block splitting technique is used to remove the primary noises. In block splitting is an important content of a web page which enclosed in div tag is considered. The main content is divided into a number of blocks. Simhash method is employed to remove the duplicate contents. Smash is a fingerprinting methodology where matching to each block a fingerprint is created. The keywords in each block are identified and the frequency of each keyword in a block is found out. The collection of fingerprints of blocks are analyzed. A block is considered as duplicate block if its fingerprint is different from other fingerprints by at most bit position where represents a small integer. Block importance is calculated based on keyword redundancy, link word percentage and title word relevancy. Keyword redundancy refers to the percentage of redundant words in a block. The percentage of link words in a block are also computed. Percentage of keywords present in a block is referred to as title word relevancy. A block is considered to be an important only if its block importance is greater than the threshold value. [4]

J. Using DOM and NLP

To extract the information like text and images from the web pages using DOM tree and NLP. In this method DOM tree structure is generated for each web page. HTML tag element is the information denoted in each node of the DOM tree. the two classes of HTML tag elements are block element and style element. Examples of block tags are , <p>, <bar>, <Li> etc. while the rest of the tags are considered as style tags. , <p>,
 tags are used to create paragraphs in web pages. [7]

K. Featured DOM Tree Method

Shine N.Das et.al. [14] has proposed a variation of DOM tree known as featured DOM tree to eliminate noisy contents from the web pages. Featuring, modeling and pruning are three main steps for noise removal. DOM tree shows the layout or presentation style of a web page and it is not sufficient to learning the content or meaning of a web page. Hence featured DOM tree was constructed which represents the feature set of individual blocks of a web page apart from the presentation style. After the preprocessing the feature set is generated. Tokenization, stop word removal, stemming, etc. are the preprocessing steps. The featured DOM tree is generated on the basis of optimal feature selection and feature weighing. In the modeling stage the DOM tree is generated. Tags form the internal nodes of the DOM tree and text, images, hyperlinks, etc. are represented in leaf nodes [13]. The web pages i.e. HTML pages are analyzed from the <BODY> tag. In the pruning phase, the noise checking is carried out. Minimum weight and overlapping (MWO) techniques are used to find out the feature set similarity. If the MWO value of a leaf node is less than the threshold value, then the leaf node is marked as noisy. Bottom up traversal of the DOM tree is carried out to eliminate the noisy nodes. A parent node is marked as noisy if all of its child nodes are noisy in nature. The marking process is propagated up the tree and the marked portion of the tree is pruned.

L. Other Noise Removal Methods

Oza and Mishra proposed [9] a method to remove the web page noise is based on the DOM tree. They observed that the web pages are not well formed. So the web pages are passed through HTML parser, which corrects the markup and generates the DOM tree structure. The maximum depth of the DOM tree is found and a suitable threshold value is determined. The technique of linear regression analysis is used to find the relationship between the maximum depth and threshold value. According to the threshold value the DOM tree is divide into multiple sub trees. Nodes of the DOM tree is less than the threshold level are considered as noise and they are removed.

Mehta and Narvekar suggested [10] technique based on the DOM tree and data filters to get rid of web page noise. The DOM tree is the graphical representation of a web page and the nodes in the DOM tree can be handled by using the appropriate methods. The DOM tree includes both noisy and non-noisy content. The unwanted contents are removed using the filters of the DOM tree. The filters contains a user defined function

which determines whether a node should be filtered or not.

VII. CONCLUSION

The noisy contents of the web pages are hidden then only meaningful content is available in the web page. Here, the advertisement banner, navigation bars, copyright forms, duplicate web pages and repeated words are considered as noise. There are many techniques are used to remove noise from the web pages. This paper discussed about related works, types of noise and the different noise removal techniques.

VIII. REFERENCES

[1]. Anchal Garg, Bikrampal Kaur "Web Page Performance Enhancement by Removing Noise" International Journal of Computer Applications (0975 - 8887) Volume 103 - No.6, October 2014

[2]. S. S. Bhamare, Dr. B. V. Pawar "Survey on Web Page Noise Cleaning for Web Mining" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 766-770

[3]. Jaideep Srivastava, Prasanna Desikan, Vipin Kumar "Web Mining Concepts, Applications, and ResearchDirections"

[4]. Surabhi Lingwal "Noise Reduction and Content Retrieval from Web Pages" International Journal of Computer Applications (0975 - 8887) Volume 73-No.4, July 2013

[5]. Mrs.Madhushree B, Yogita K Patel "A Review on Noise Removal from Web pages for Web Content Mining" International Institution for Technological Research and Development Volume 1, Issue 1, 2015

[6]. Sekhar Babu Boddu Assistant Professor, Department of MCA, KL University, Guntur, Andhrapradesh, India -522501 sekhar99@gmail.com "ELIMINATE THE NOISY DATA FROM WEB PAGES USING DATA MINING TECHNIQUES " GESJ: Computer Science and Telecommunications 2013|No.2(38) ISSN 1512-1232

[7]. Hassan F. Eldirdiery, A. H. Ahmed "Detecting and Removing Noisy Data on Web Document using Text Density Approach" International Journal of Computer Applications (0975 - 8887) Volume 112 - No. 5, February 2015

[8]. Mrs Madhushree B, Yogita K Patel "A Review on Noise Removal from Web pages for Web Content Mining" International Institution for Technological Research and Development Volume 1, Issue 1, 2015

[9]. Rajni Sharma, Max Bhatia Department of Computer Science and Engineering Lovely Professional University, Phagwara "Eliminating the Noise from Web Pages using Page Replacement Algorithm " Rajni Sharma et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3066-3068

[10]. Sandeep Kaur1, Abhishek Tyagi2 Department of Computer Science Engineering, Lovely Professional University Jalandhar, Punjab, India1 Email: Sandeepkaur.1489@yahoo.com Assistant Professor, Department of Computer Science Engineering, Lovely Professional University Jalandhar, Punjab, India2 Email: Abhishek.16857@lpu.co.in "Noise Reduction and Content Extraction from Web Pages Using DOM Based Page Segmentation" Sandeep Kaur et al, Int.J.Computer Technology & Applications, Vol 5 (6),2022-2027

[11]. Yogita K patel 1, Mr.Narendrasinh Limbad2 "Noise Removal from Web pages for Web Content Mining" IJARIIE-ISSN(O)-2395-4396 Vol-2 Issue-3 2016

[12]. Anchal Garg M.Tech CSE, Bikrampal Kaur Ph.D "Web Page Performance Enhancement by Removing Noise" International Journal of Computer Applications (0975 - 8887) Volume 103 - No.6, October 2014

[13]. Vıt Baisa "Web Content Cleaning" Masaryk University Faculty of Informatics Hui Xiong, Member, IEEE, Gaurav Pandey, Michael Steinbach, Member, IEEE, and Vipin Kumar, Fellow, IEEE "Enhancing Data Analysis with Noise Removal" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING , VOL. X, NO. X, XXX 200X

[14]. Shine N. Das, Pramod K. Vijayaraghavan, Midhun Mathew "Eliminating Noisy Information in Web Pages using featured DOM tree" International Journal of Applied Information Systems (IJAIS) - ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 2- No.2, May 2012 - www.ijais.org