Noise Detection and Web Content Extraction Using Page Segmentation and Text Density

A Dissertation Submitted

By

Sandeep Kaur

11304827

To

Department of Computer Science

In fulfilment of the Requirement for the

Award of the Degree of

Master in Technology in Computer Science and Engineering

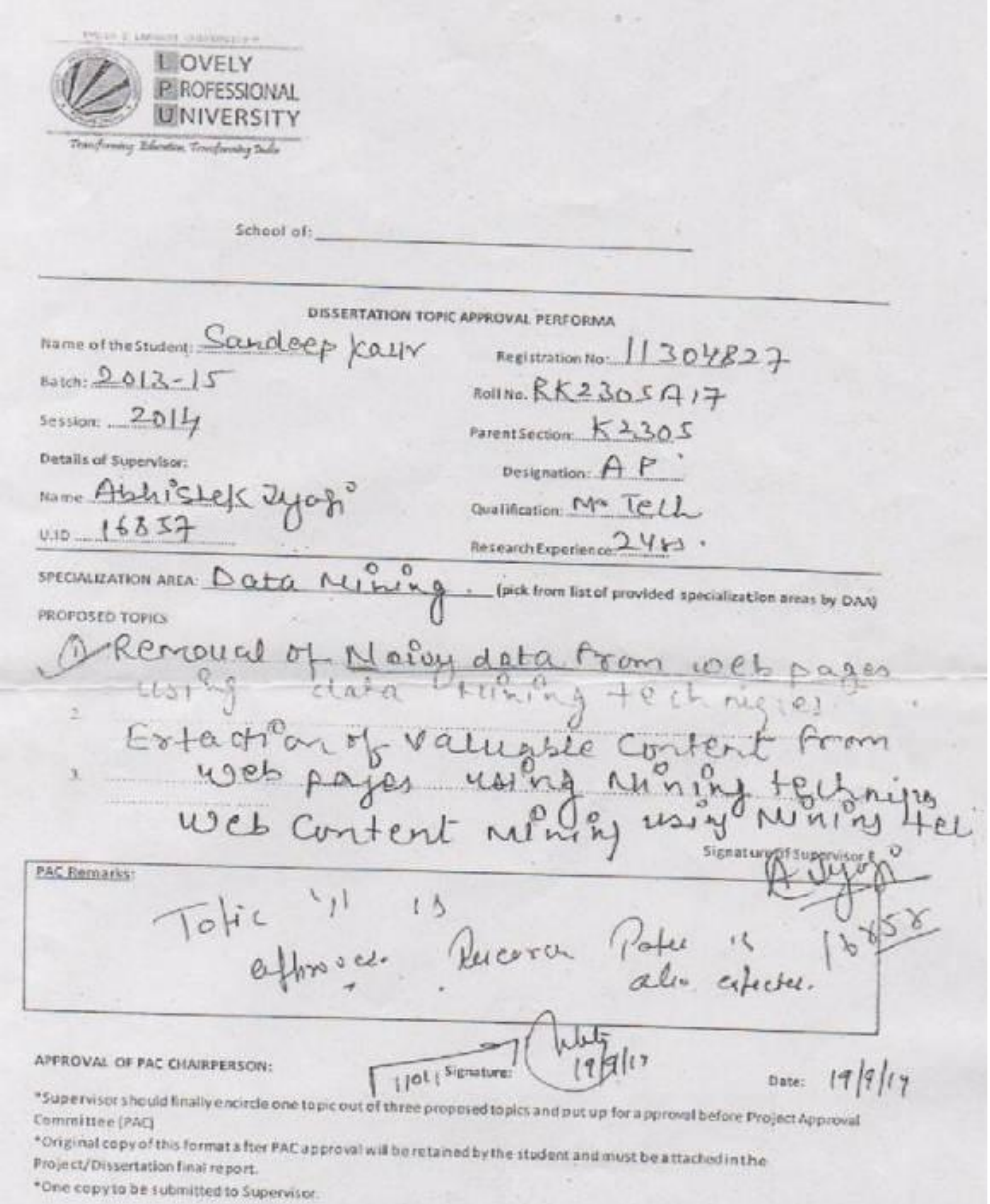
Under the guidance of

Mr. Abhishek Tyagi

(Assistant Professor, Lovely Professional University)

(MAY 2015)

PAC FORM



LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: ________________

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Sandeep kaur

Registration No: 11304827

Batch: 2013-15

Roll No. RK2305A17

Session: 2014

Parent Section: K2305

Details of Supervisor:

Designation: AP

Name Abhishek Tyagi

Qualification: M. Tech

U.ID 16857

Research Experience: 2Yrs.

SPECIALIZATION AREA: Data Mining. (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Removal of Noisy data from web pages using data mining techniques
2. Extraction of valuable content from web pages using mining techniques
3. Web Content mining using Mining tech

Signature of Supervisor

A Tyagi 16858

PAC Remarks:

Topic "1" is approved. Research Paper is also expected.

APPROVAL OF PAC CHAIRPERSON: Signature: 19/9/17 Date: 19/9/17

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

A web page contain large amount of information and information present in web page is also include additional content such as navigational bar, advertisements, decorated images which are not related to main textual content or may hamper the performance of PDA and small cellular devices. The user is not interested this additional content which is called the noisy information. Extracting the useful textual information from web page content extraction approach is provided. The text density approach is used with page segmentation to extract the text data and composite text density smoothing with pattern matching algorithm used to detect the noise from web page. It is fast and accurate method then previous approach for extracting the content and also maintains the original structure of web page. The Content Extraction Composite Text density and Density Sum (CECTD-DS), Content Extraction Text Density and Density Sum algorithms (CETD-DS), Content Extraction Composite Text density smoothing (CECTD-S) and Extended Content Extraction Composite Text Density Smoothing (ECECTD-S) with pattern matching are compared to provide the best effective result.

ACKNOWLEDGMENT

I take this opportunity to express a deep sense of gratitude towards my guide Asst. Prof. Abhishek Tyagi, for providing excellent guidance, encouragement and inspiration throughout the work. I would like to thank him for his constant assistance and overwhelming interest in the research that has made it successful. I express my deep sense of gratitude and appreciation towards him. Without his invaluable guidance, this work would never have been a successful one.

I would also like to thank my family and friends who have been a source of encouragement and inspiration throughout the duration of this research.

SANDEEP KAUR
(11304827)

DECLARATION

I hereby declare that the dissertation proposal entitled "Noise Detection and Web Content Extraction Using Page Segmentation and Text Density", submitted for the M-Tech (Computer Science and Engineering) degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: **Investigator**

Reg. No. 11304827

CERTIFICATE

This is to certify that she has completed M. Tech dissertation proposal titled “Noise Detection and Web Content Extraction Using Page Segmentation and Text Density” under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any degree or diploma.

The Dissertation is fit for the submission and fulfilment of the conditions for the award of M. Tech Computer Science & Engineering.

Date: _______________

Signature of Advisor

Name:

UID:

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Data Mining

Data Mining is any kind of data repository. Different databases like relational database, data warehouse and multimedia database is studied under Data Mining. Data Mining is defined as mining or extracting the useful information from the large amount of data. It can also define as data mining is mining the information from data. There is large amount of data available that require being bitter into useful information and various applications like production control, market basket analysis.

Data Mining used various Processes for extracting the useful information from large amount of data called as knowledge discovery process, knowledge mining, and knowledge extraction [7].

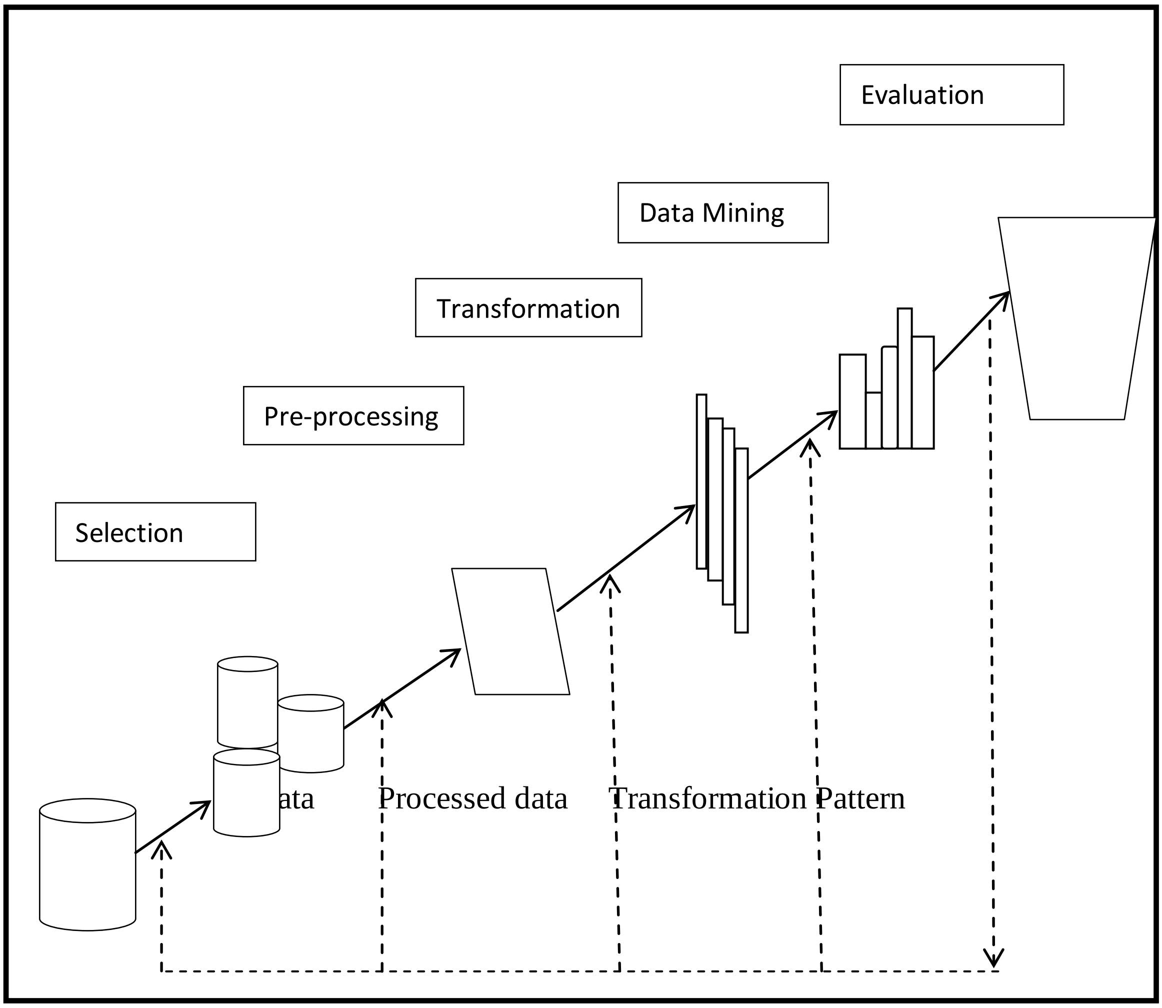


Fig.1.1 Knowledge Discovery Process

In other words Data Mining is a technique used discovery the various patterns and further based on these patterns decision are developed. It includes some steps which describe as:

- Exploration
- Pattern Identification
- Deployment

Exploration: Data Exploration is used to clean and transformed the data into another form. It is an informative search that is used by data consumers to form true analysis from the information gathered. Data is always gathered in large bulk which is in rigid form. After that there is a need to narrow down the bulk data. Data exploration is used to analyse data and information from the data for further data analysis. Data which is come from various data ware houses are in various types of data format. Data exploration is used to collect relevant information from various sources.

Pattern Identification: The pattern identification recognizes and choosing the best patterns and based on these patterns makes prediction.

Deployment: The desired outcome is deployed based on patterns. Pattern evaluation is that module which interacts with the data mining modules so as to focus the look for interesting patterns and graphical user interface which communicates between users and the data mining system allow the user interface with system.

Data Mining is a method used for discovering useful information from a large set of data which is stored in database, warehouse, or in repositories. Data mining architecture has the following components.

- Database, Data Warehouse, other data repositories: The various techniques are performed on database, warehouse, spread sheet and other information repositories are data cleaning and data integration.
- Data Warehouse Server, Database: Extracting the data base and user’s data mining request are handled by data warehouse server.
- Knowledge Base: It searches or evaluates the interesting results and patterns.
- Data Mining Engine: Data mining system perform various tasks such as categorization, association analysis, classifications, evaluations etc.
- Pattern Evaluation: It employs interesting measure with data mining modules.

- Graphical User Interface: GUI communicates between users and data mining system which allow the user to interact with the system by using data mining query and providing the information based on the query [3].

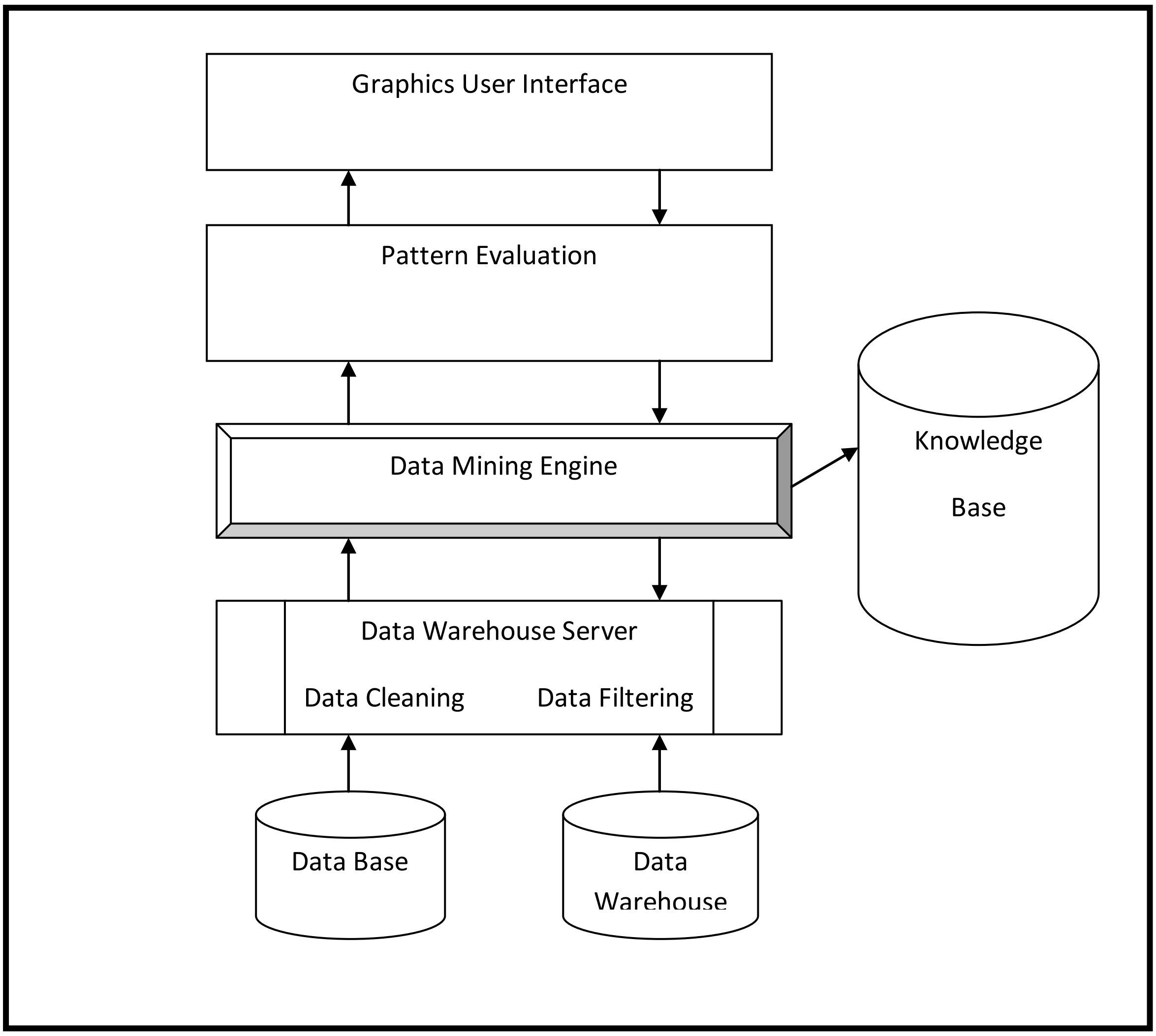


Fig.1.2 Data Mining Architecture

Data Mining includes the techniques clustering, classifications, decision tree analysis, and association rules. The most important data mining applications is a Web Mining. World Wide Web has a popular place for dissipating and accumulates the information. Extracting the useful information from web pages becomes essential task. Web is a medium for accessing the information store in 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 resolving these problems various methods and techniques are developed. Web Mining is a developing research area motivated on resolving these problems. The various techniques include Web Content Mining, Web Usage Mining, and Web Structure Mining WWW is a well-known standard by which people can spread and gather the information. Web Pages contain large amount of undesired information, which is called noisy or irrelevant content. The navigational panel, header, Footer, copy right, advertisements known as noisy content [8]. Noise is mainly of two types:

- Global Noise: Global Noise with large granularity. Global noise do not presents individual pages in Web. Global Noise over all in Web site like Mirror sites, duplicate page etc.
- Local Noise: Local Noise is not related to main content of Web Page .Local Noise includes Navigation Bar, Header Footer, Advertisements, Copy right, Decoration Pictures etc. A user is interested in main content of a web page. Identifying main content blocks from a web page is called content extraction or information extraction.

How does Data mining works

As the technology growing fast, data mining provides relation between the transaction and analytical system. Data mining software analyse patterns, relationship and store data based transactions. Many types of software available like statistical, neural network and machine learning. Therefore four type of relationship are described as following:

- Classes: Data is stored and used to locate the predefined group is known as classes.
- Clusters: Data items are grouped together according to their logical relationship on the preference of customers are known as Clusters.
- Associations: Associations are identify based on data mined.
- Sequential Patterns: Data is mined to expect the behaviour patterns and trends.

Data Mining system: Data mining system are classified into various classifications as following:

1. Classification according to type of data source mined: In the Classification according to type of data source mined system is categorized according to times series data, multimedia data, spatial data handled.
2. **Classification according to the data model:** In the Classification according to the data model system is categorized according to relational database, object oriented database, relational database handled.

3. **Classification according to the kind if knowledge discovered:** In the Classification according to the kind if knowledge discovered system is categorized according to the knowledge discovered like association, clustering, classification etc.
4. **Classification according to the mining techniques to be used:** In the Classification according to the mining techniques to be used system is categorized according analysis approach like machine learning, neural network, genetic algorithm and visualization.

1.2 Web Mining

The Web Mining is a technique of data mining which is used to automatically find and extract the meaningful data and information from web documents. The Web Mining presents the data in the World Wide Web and these databases in the form of web pages.

Web data can be as:

- Contents of Web Pages.
- HTML or XML tags are used intra page structure.
- Structure of links between the web pages are used inter page structure.
- Web Page accessed by visitors describe by usage data method.

Web Mining decomposed into following Subtasks:

- Resource Discovery: It is used to find the information and task of retrieving the web documents.
- Information Selection & Pre-processing: It is used for gathering the information pre- processed that information which is retrieved from web resources.
- Generalization: It determines the common patterns at different web site or multiple sites.
- Analysis: It Justification understanding of the various mining patterns.

Typical Sources of Data:

- Data is automatically stored in server which can access logs; refer logs, agent logs and client-side cookies.
- Product oriented and E-commerce events.
- User ratings and user profiles.
- Page attributes, page content, site structure and Meta-data.

Concept of Web Mining: The Main function of web usage mining is to find out meaningful data generated from client-server transactions on number of web servers.

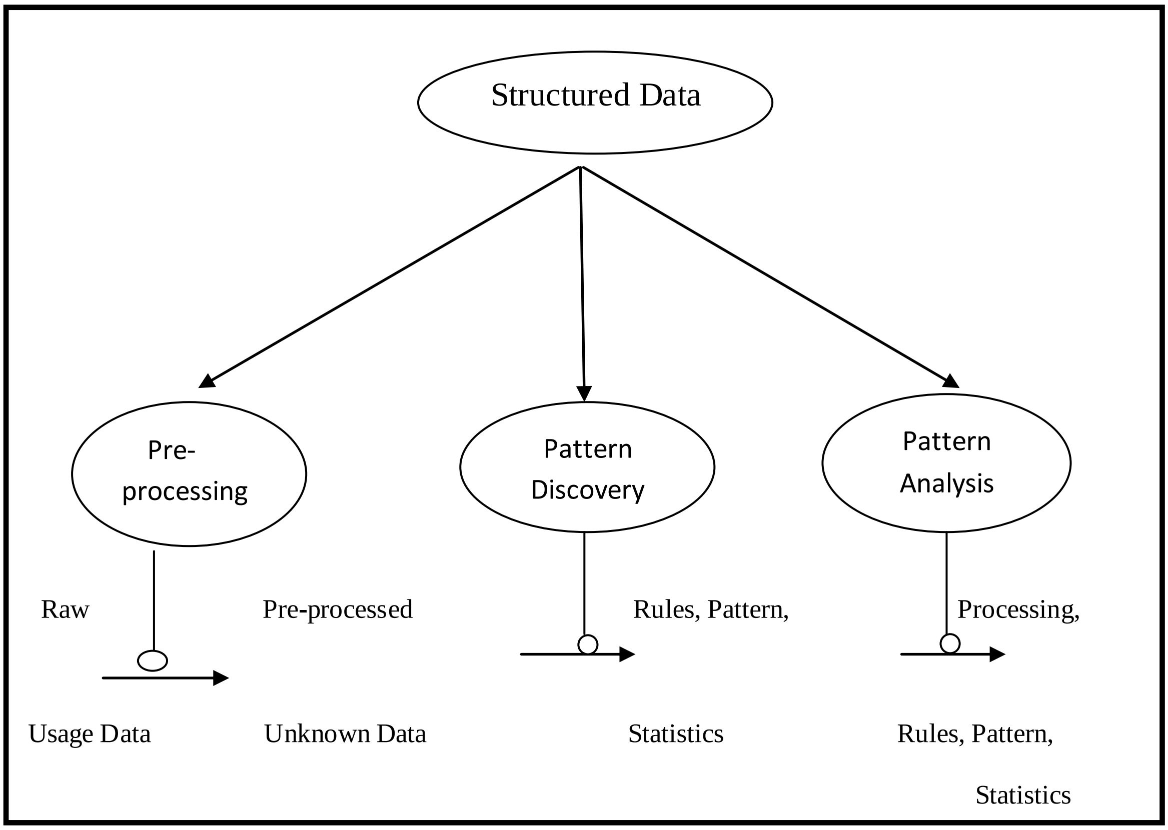


Fig.1.2.1 Concept of Web Mining

Web Mining is Categories: Web Mining categories as following:

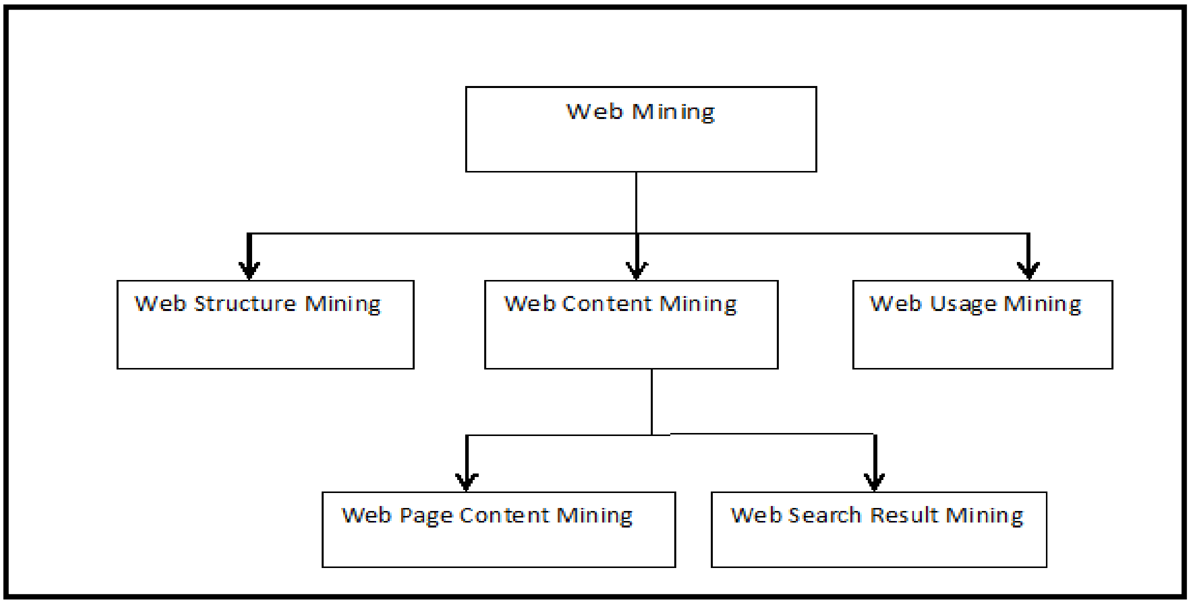


Fig.1.2.2 Web Mining Classifications

1.2.1 Web Structure Mining: Web Structure Mining includes of web pages as a nodes and links which Contains another linked pages. The Web Structure Mining two types:

1.2.1.1 Hyperlinks: In this Web Page located in diverse location any similar Web Page or a diverse Web Page.

1.2.1.2 Document Structure: In this content inside a Web page is also organised in a tree-structured form founded in many HTML or XML tags e.g. DOM (Document Object Mode). Mining efforts have focused upon automatically extracted webpages.

In typically web graphs which are consisting of web pages as nodes, hyperlinks as edges connecting to related pages. Web structure mining is the process of discovering structure information from the web.

1.2.2Web Content Mining: The Knowledge discovery process followed by Web Content Mining. It is suitable for extracting the information contents from Web Documents. The main objective of Web Content Mining is collection of text documents or multimedia documents e.g. image, video, audio. These are embedded in Web Pages as shown below.

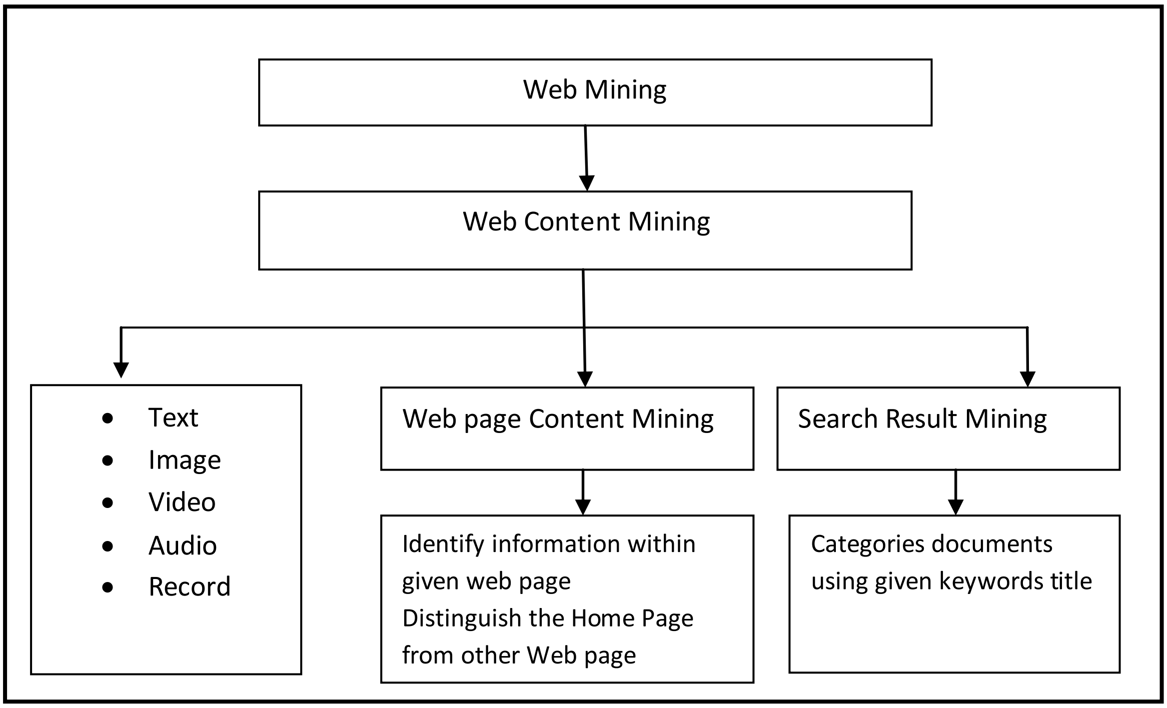

Web Mining
Web Content Mining
Text
Image
Video
Audio
Record
Web page Content Mining
Search Result Mining
Identify information within given web page
Distinguish the Home Page from other Web page
Categories documents using given keywords title

Fig.1.2.2.1 Web Content Mining

- Web Page Content Mining: Searching of Information from web pages with the help of Content.

- Search Result Mining: Search result mining based on further search of pages which is found in previous search.

Web Content Mining distinguished based on two points of views:

- Agent Based Approach: This technique used to improve the information finding and filtering. The informational retrieval techniques based on hypertext web documents retriever, filter, and categorize them.
- Database Approach: This approach modelling the data on the Web and apply database querying or data mining applications to analyse it [10].

1.2.3 Web Usage Mining: Web Usage Mining is used to extract the various usage patterns from web usage data. Usage data find and captured along with user browsing behaviour at a web site.

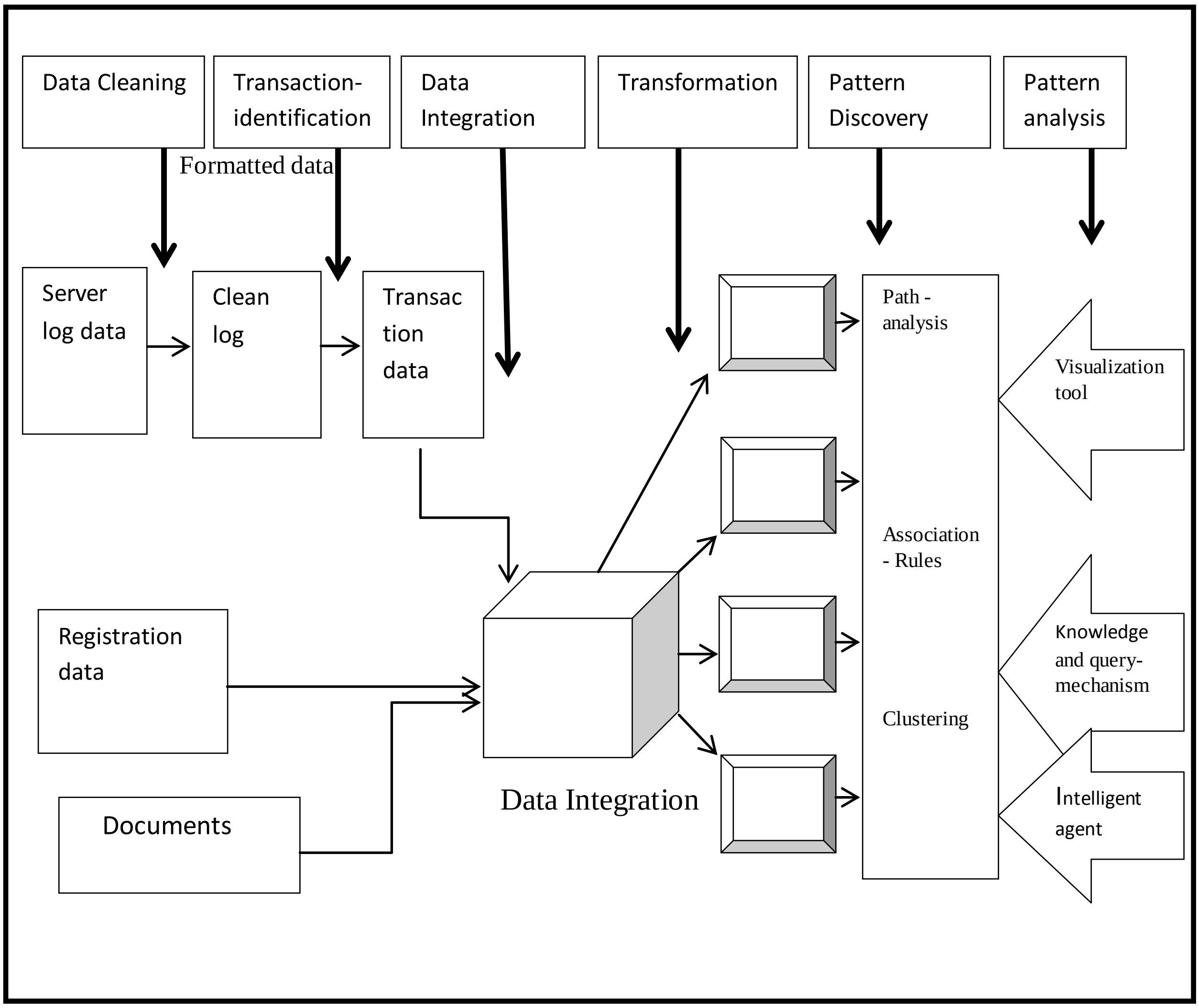


Fig.1.2.3 Web Mining Architecture

The main purpose is Web Usage Mining gather the information from log file on the Web Server. Data mining method used to identify the navigational behaviour of the users based on the web log data.

The Web Usage Mining process is process which involving of data preparation, pattern finding and pattern analysing. The data is pre-processed to identify user, session, page views etc. In the pattern discovery association rules, clustering applied in order to detect interesting patterns. Further patterns are stored that can be analysed for the Web Usage Mining process.

1.3 Techniques of Content Extraction and Noise Removing From Web Pages:-

1.3.1 Extraction of Main Content with Text Density and Visual Importance.

The large amount of information is present in web sites and web pages on internet. Web pages also contain the additional information such as banners, advertisements, duplicate pages, copyright etc. which is not related to the main contents. Such information is useful for only site owners and user browsing. That type of information is treated as noise. The Content extraction technique used find the main information and eliminate the noisy data which is useful to improve the performance for web mining and retrieve the useful data from web pages Dandan Song [2] proposed a DOM (document object model) based approach for content extraction which is used to define the text information and visual information of web pages. Content Extraction technique using DOM tree can be expressed as:

1.3.1.1 DOM Tree: Document object Model (DOM) interface used for retrieving and updating content and structure of documents. The tags are represented by internal nodes or detailed text and images are leaf nodes. DOM tree describe the complete structure of the web pages. Example of html Code given below:

```
<HTML>
<HEAD>
<TITLE> text</TITLE></HEAD>
<BODY>
<P> p text</P>
<IMG SRC= “1.jpg”></IMG>
</BODY>
</HTML>
```

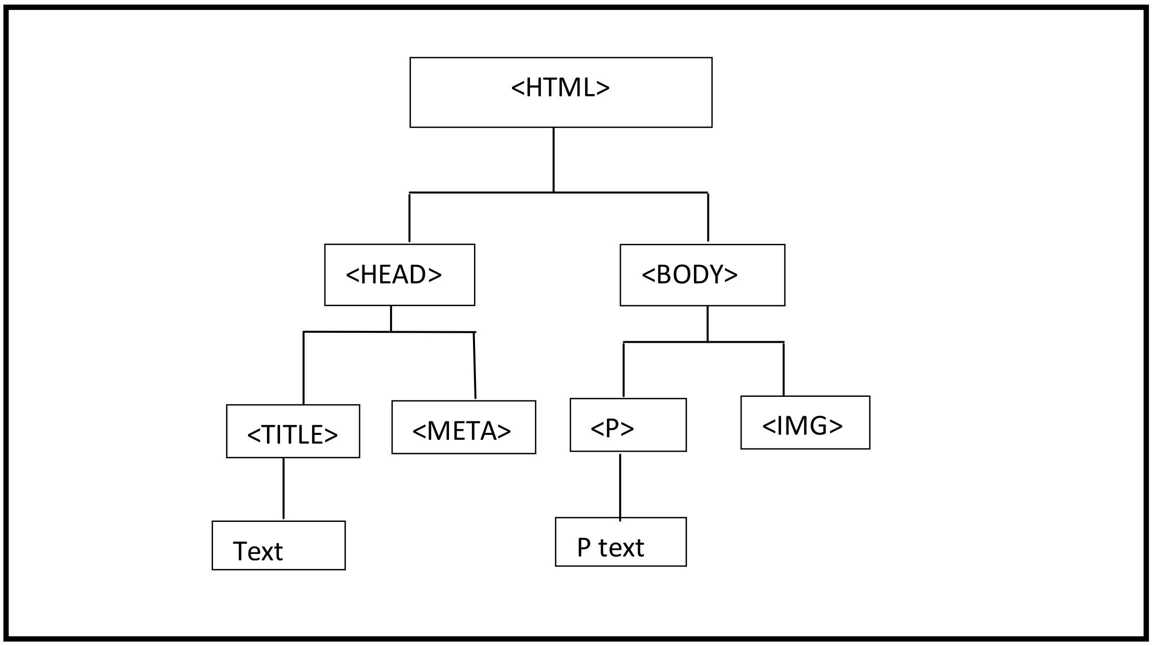


Fig.1.3.DOM Tree Example

1.3.1.2 Text Density: Text Density found the noise which is more formatted and contain small text and detailed sentences. The main information is usually lengthy and less formatted. When HTML document parsed and represented with the help of DOM number of characters and tags each node presented.

CharNumber: Number of character in its subtree.

TagNumber: Number of all tags in its subtree.

The Text Density technique can be defined as follows:

$TD_{i=}C_i/D_i$

In the statistics of the huge number of websites, the main page contains large number of advertisements and decorated images and hyperlinks for a new web page. The main content is organised into hierarchal structure and centralized in nature. The threshold value for node containing data is always different from other nodes at the same level.

1.3.1.2.1 Link Text Density (LTD): It is the ratio of all link text length to all content text length under a given node. The outcome value is a significant indicator judge advertising block. The greater the value, the greater the possibility of advertising block.

1.3.1.2.2 Link Amount (LA): It is the number of link nodes of all child nodes under a given node. It is a main judgment index for the accurate content. The peak value indicates that there are more child nodes, thus related work has to be carried out to check its child nodes.

1.3.1.2.3 Link Amount Density (LAD): It is the ratio of the number of all child links to the number of all child nodes under a node. The value gives the number of link nodes in a node. It judges whether the content contains a number of links.

1.3.1.2.4 Node Text Length (NTL): It is the length of all text pages with labeled nodes removed. LAD and NTL should be combined to perform the functions of accurate content judgment.

1.3.1.3 Composite Text Density: The large amount of noise in web pages contains hyperlinks and statistical information distinguished from Text Density is called the Composite Text density.

1.3.1.4 Visual importance and Hybrid Text Density: The useful content of web page mostly resides in the central part of screen. To combine the text density, textual or Visual Information for DOM nodes defines visual information and the measure of Visual Importance is combined into composite text density which is defined Hybrid Text Density. Before Text Density Compute, the algorithm eliminates unimportant parts from a HTML Documents like Scripts and Comments. Density Sum and threshold approach is used for content extraction.

1.3.2 Recognition of Navigation Page by using DOM and Text Block Identification.

The growth of web pages on internet continues and the Web Page organizations are very essential. The Web Pages can be categorised into Navigation page and content page.

- Navigation page guide the users to find the information and home page.
- Content pages provide the users information, such as page reporting.

Navigation page detection is important it contain significant content pages and used for topic tracking .The Navigation page contain content pages with some same topics. The Navigation page contains more links and some text, where the content pages include more text and some links. There are few pages that contain noise like navigational bars, advertisements, header footer etc. A DOM based block text identification method proposed which detects the Navigation Page. This approach used to extracting the text segment block from a Web Page. If the number of segments is very small that page is mostly navigation page or large segments are mainly content page and the content page is not divided into many small blocks [5].

1.3.2.1 Outlinks/pagesize: The ratio of outlinks in navigation page is more than content page.

These are calculated as following:

OL/PS=Numberoflinks/PageSize ………………………….. eq. (1)

1.3.2.2 Anchor Text Density (Anchor TD): In this whole page's Composite Text Density is calculated. The anchor text ratio of Navigation Page is more than in content page.

AnchorTD= AnchorText/NonAnchorText….…………………eq. (2)

1.3.2.3 Pre-process, Block-Text Identification: Delete noisy nodes when Web Pages parse into a DOM tree such as script, text area, style etc. Block Text Identification includes following steps:

- Average Text Density (ATD) is calculated as:

  ATD=SizeOfText/NumberOfLines……………………eq. (3)
- If the text density is larger than ATD then that block is text block.
- If the number of text block is more in web page then that is navigation page.

CHAPTER 2
LITERATURE REVIEW

Satish J.Pusdekar et al. (2014) Information extraction from web pages is a challenging task due to the complex structure of web pages. The data on web pages are always displayed in regularly for users to browse. The information in deep web pages presents in the form of data records when the user queries the result is returned by web databases. So it is necessary to extract the structured data from deep web pages for further processing. A visual block tree is implemented for data record extraction from web pages. VIPS algorithm is used to generate the visual tree and extract the data records and data items. Information extraction of web pages consists of both text and images through visual features. Using VIPS approach web page transpose into a visual block and then extract the information from these blocks. This visual tree segment the web pages in which there is a root block that describe the whole page and each blocks represent the corresponding rectangular region on the page. Extracting the data record from web pages discover the boundary of data record and locate the data region then extract data record from data region[9].

Suresh Subramanian et al. (2014) presented a Genetic Algorithm and Duplicate Web Documents Identification function which is used to improve relevance of retrieved documents by removing the duplicate records from the dataset. The Genetic algorithm, for html web content mining (GAHWM) matching the title, content and number of anchor tag used to detecting the duplicate documents. GAHWM is work on a sentence level and compared sentence by sentence [15]

Anna Saro Vijendran et al. (2013) present a very effective technique LBDA for extracting the information. Layout based detachment approach used to extract the main information from web page and remove the unnecessary information like header footer, advertisements navigational bar. In this web page is converts into DOM tree to examine the structure of web page and integrate the all tags into a single tree which have common features of web pages. Structure analysis is used to find the tags in web pages and require accessible in page block. The unwanted tags are removed using tag tree parsing and block segmentation is created using block acquiring technique, block filtering is applied for content extraction. It utilizes the less execution time then other approaches [1].

Dandan Song et al. (2013) proposed a technique to identify and remove the noisy part from web pages. Two techniques are used find the textual information and find the visual information from web pages. Text density is used to extract the content from web pages calculate the textual data and noisy data. The noisy data in web page are that which are highly formatted and contains more number of tags. The content in web page lass formatted and contains more text and less number of tags. The more number of hyperlinks in web pages are known as noise and identify using composite text density. Visual information is defined using find the location of main content called as hybrid text density. After that density sum and data smoothing function is used to find the main content and noisy content. To find the more than one content block in web page the maximum density sum tag algorithm is used. CETD provide the original structure of web page which can be best suited for applications such as small screen devices [2].

Shobhit Srivastava et al. (2013) a new approach is presented class based attributes with DOM segmentation. The class attributes is find present in the HTML page inside the body section. In this HTML page is converted into XHTML using standardize tag and segment the web page using DOM tree then apply the class attribute technique. Cleaning of web page in which unnecessary tags are removed from the XHTML documents. Using class attribute based approach different classes of element inside body section creates with its own properties. All class attribute with the DIV tag extracted the web page information. With the help of class attribute approach main content, table of items, section which contains hyperlinks and several structural elements based on html segments defined and produced the content extraction results [11].

Xuhong Zhang et al. (2013) a new low cost vision based web page segmentation and information content extraction technique proposed. Using VIPS algorithm various rules is generated to find the html elements and extract the information from these elements. Web page is segments using row column splitting technique and choose the best information part from various segments of web page with the help of Degree of coherence (DOC).This algorithm is used to segment the web page and find the information block from the segmented web page. Low Cost VIPS algorithm provide the efficient information extraction and achieves high degree of performance as compared traditional clustering process in detecting informative block [18].

Surabhi Lingwal (2013) proposed a method to filter the web pages eliminate the noise part and retrieve the useful informative content from web pages. A web page is segmented in various blocks than tokenize the web pages to remove the unwanted tags from HTML page. Redundancy removal algorithm is used to remove the duplicate pages from dataset. Further outliers detection algorithm is used to detect the noise that is not related to main content. Content extraction process used to extract the textual content from documents. A high length text block known as text of main content block and minimum text length block known as noisy block. This content extraction technique works also on local noise removing as well as global noise removing and provide efficient results [14].

Wu Qi et al. (2012) present the content extraction technique from html pages based on the web clustering process and retrieved the useful informative part. In this paper the unnecessary information is cluttered and only important content is extracted from web pages.HTML page is converted into DOM tree and various nodes is generated. Every node given id and adjust according to the relations with each other nodes. Valid information is obtained from these nodes according the longer text node from page. Smoothing score is applied to all nodes in HTML pages in case any content node is declared by mistaken as non-content node. With the help of smoothing process recognized all content nodes and adjust the various score of content nodes effectively. The ID mechanism technique provides the best result find the information content from HTML documents [17].

Shuang Lin et al. (2012) present the density based approaches for content extraction from web pages. Web pages contains noisy information and main content information, users are mainly interested in main content. To identify the noise and extract the useful information a content extraction approach is used to combined the segmentation approach and density based approach. Segmentation converts the web page into small blocks and block extractor tool used to extract the content from various blocks. The BLE&IE approach used to find the redundant blocks and noise blocks and removes these blocks from pages. Density based approach is used to find the length of text block. Main content always contain the large number of words and few number of tags.so with the help of density calculation main content is extracted from web pages. Segmentation and density algorithm combined to obtain the better result to extract the useful information and identify the duplicate documents [12].

Li Yue et al. (2012) introduced a DOM Based block text identification technique that is used for detect the navigation web pages. This Navigation pages detected based on URL feature as some attributes of navigation pages. The outcome was that the ratio of out-links to navigation page is greater as compared the content pages. Anchor tag density and Non anchor tag density is find for detect the page is navigation page or content page. After that Pre-Process the web page and block text identification method is applied and provide the effectiveness result detecting navigation pages [5].

Tim Weninger et al. (2010) present content extraction via tag ratios approach to extract the content information from many web pages based on HTML documents tag ratios. Various tags computed line by line, clustered based on similarity of tags and provide the result content area or non-content area. To extract the content from web pages first of all unnecessary tag is removed from HTML documents and then further tags clustered and grouped and histogram is generated based on tag ratios. Minimum tags and large text in clusters called text area and less text large number of tags called noise area or non-content area. With the help of smoothing find the important content lines might be lost. CERT provides the fast and accurate method extracting the content from HTML web pages [16].

Michal Marek et al. **(2007)** proposed a method in which HTML code is parsed and not useful text, scripts, are removed from HTML structure after that pre-cleaned HTML text is again for text blocks separated by one or more HTML tag. Labels assigned manually to content block and used to produce the cleaned output. The result is noisy blocks which has no interest should be eliminated [6].

Shumeet Baluja (2006) presents an approach web page segmentation that is re-examines the task through the decision tree learning. The number of sample classified at each node and the subset reach the parent node. The conditions are examined and feature is selected based on the sample. The Recursive Segmentation procedure are visualize through a vertical and horizontal cuts. Many rules are generated to ignore the DOM elements and extract the useful elements [13].

Kushmerick N (1999) introduced effective technique ADEATER that removes advertisement images from internet pages. The removing advertisement images from web pages results internet pages downloaded faster. An inductive learning approach is used where rules are generated and apply those rules for removing advertisement from internet pages. ADEATER achieves high level of accuracy result [4].

CHAPTER 3

PRESENT WORK

3.1 PROBLEM FORMULATION

With the fast development of internet large amount of information are available in web sites and web pages. Web pages also contain the additional information such as banners, advertisements, duplicate pages, copyright etc. these are not related to the main contents. Such content is useful for site owners and user browsing. That type of information is treated as noise. The information extraction technique used to extract the main content and remove noisy data is useful for improve the performance and retrieve the required information from Web Pages. Dandan Song proposed a DOM (document object model) based approach for content extraction which is used to define the text information and visual information of Web Pages. DOM is language independent model which is used foe extract and update the structure and content in a WEB Page Document. In the document there is various tags text and images, internal nodes include tags and detailed text, are leaf nodes includes images. DOM tree describe the complete structure of the web pages. Text Density found the noise which is more formatted and contain some text and sentence. The main content in Web Page is lengthy and simple formatted. DOM tree parsed and structures the HTML document and all tag, text in each node is configured in a tree structure. Mostly the noise in web pages contains hyperlinks. The statistical information is distinguish from Text Density is called Composite Text density. The main content of Web Pages is represented in the central part of the Web Page and Visual Importance is defined based on text, text density and visual information based on DOM node. The degree of Visual Important is combined into composite Text Density, which is known as Hybrid Text Density. The various types of noisy data present on the web pages and remove these type of noisy data from the web pages various techniques had been proposed previous times as discussed in the literature review section. In this work, we will remove the noisy data from the web pages by apply the pages segmentation with Text density and Pattern Matching techniques.

3.2 SIGNIFICANCE

World Wide Web has a popular place for dissipating and accumulates the information. Extracting the useful information from web pages becomes essential task. The web is used for extracting the information stored in different sources. Extracting information from large amount of data and different sources creates many problems like difficulty in finding useful and relevant data. The various category of Web Mining are Web Content Mining for content extraction, Web Usage Mining for user browsing behaviour and Web Structure Mining for analyse the structure of web page. The main objective of Web Content Mining is collection of traditional text documents, and multimedia documents. Web Structure Mining used to extracting the organization of the Web Page by scanning the tags and uses interconnections to give weight to the web page. Web Usage Mining extracts the usage pattern of web user by observing the web log files, user profiles, user session.

WWW is a medium which are used by people to spread and gather the information. But the web pages of contain undesired information, which is called noisy or irrelevant content. Mostly navigational panel, footer and header, advertisement are noisy content. Mostly user is interested accessing the main information from Web Pages. The procedure of extracting the main content block from a Web Page is known as information and content extraction. Content and information extraction technique was originated by Dandan Song. The effective information extraction technique used to access the main content from Web Page is based on two techniques. The first technique is, a Web Page, noise is more formatted and contains some text and sentence, where the useful information and main content is simply formatted or contains the more text, data and fewer links than in the noise. The other technique is the main information part is mostly presented in the central part of a Web Page.

3.3 OBJECTIVES

1. To study and analyse page segmentation approach for splitting the web pages into small blocks.
2. Enhancement in Text density algorithm to increase efficiency and accuracy of the algorithm.

3. A page segmentation pattern matching approach is proposed by using text density algorithm to overcome the problems like extract the main text content and detecting and removing the noise which the user not interested and make the performance of PDA devices slow.

4. Implementation of the enhanced technique and existing technique. Analyse the results graphically in terms of accuracy.

3.4 RESERCH METHODOLOGY

Web Pages contain many information parts and noisy parts. The noisy parts can harm the web content mining. In Dom based Content Extraction by integrating Textual and Visual Importance approach, the HTML document splits into tree structure. The Text Density, Composite Text Density is computed or visual importance considers the main content in the web page. But this approach has some limitations: One is that the HTML parser is quite slow. DOM tree structure effects performance and does not perform well with lots of menus and short descriptions or hyperlinks. Another problem is that various Web Pages are in HTML code and mostly traditional solutions based on analysing the HTML code.

3.4.1 ALGORITHM OF PROPOSED TECHNIQUE

- Web pages: For the purpose of detect noise information and extract textual content web page is selected and loaded with advertisements , Navigational and menu bars, header footer , unnecessary images and hyperlinks.
- Page Segmentation: A web page split into various rows and columns using horizontal and vertical cuts .These row columns splitting approach based on fixed length of segments and <div> tag.
- Content Extraction Text Density Approach: CETD is used to extract the textual information from web page Text density algorithm is used which calculate the number of words white spaces and character in the <div> tag under segment of web page.

Example.1 Find the text content and noise content using Text Density Approach

Number of div tag= 1, chars=52, tag=4;

Number of div tag=3, char=12, tag=10;

- Content Extraction Composite Text Density: CECTD Calculate the content and noise based on the time and date of the web page .It calculating the starting date of web page loaded and check the updated information in web page.
- Content Extraction Composite Text Density Smoothing: CECTDS calculate hyperlinks in the web page and calculated the hyperlink and advertisement noise from web pages.
- Composite Text Density with Pattern Matching: CTDPM calculated the updated and non-updated hyperlinks and advertisement noise under the <div> tag in the web page. The pattern matching algorithm is used with the CECTD-S algorithm to increase its efficiency of noise detection from the web page. Pattern-matching algorithm scans the text with the help of window in which the size of the window is equal to the length of the pattern. Pattern Matching is used to find a pattern which is relatively very small or assumed to be large.

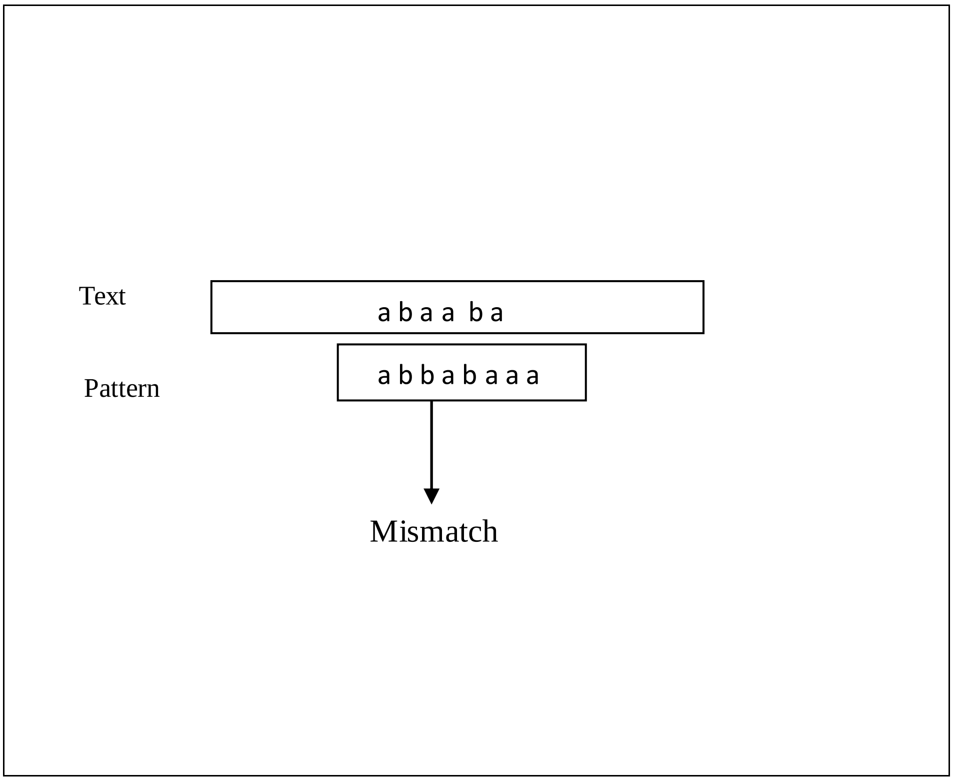


Fig.3.4.1 Text and Pattern Matching

In the first step pattern is set to left end of the text, and matching process start. In second step after a mismatch is found, pattern is shifted one place right and new matching process start, and so on. The pattern and text are in arrays pattern [1...m] and text [1...n] respectively.

3.4.2 FLOW CHART OF PROPOSED TECHNIQUE

To extract the useful information from web pages Content Extraction with Text density and page segmentation approach is proposed which split the web page into various small blocks.

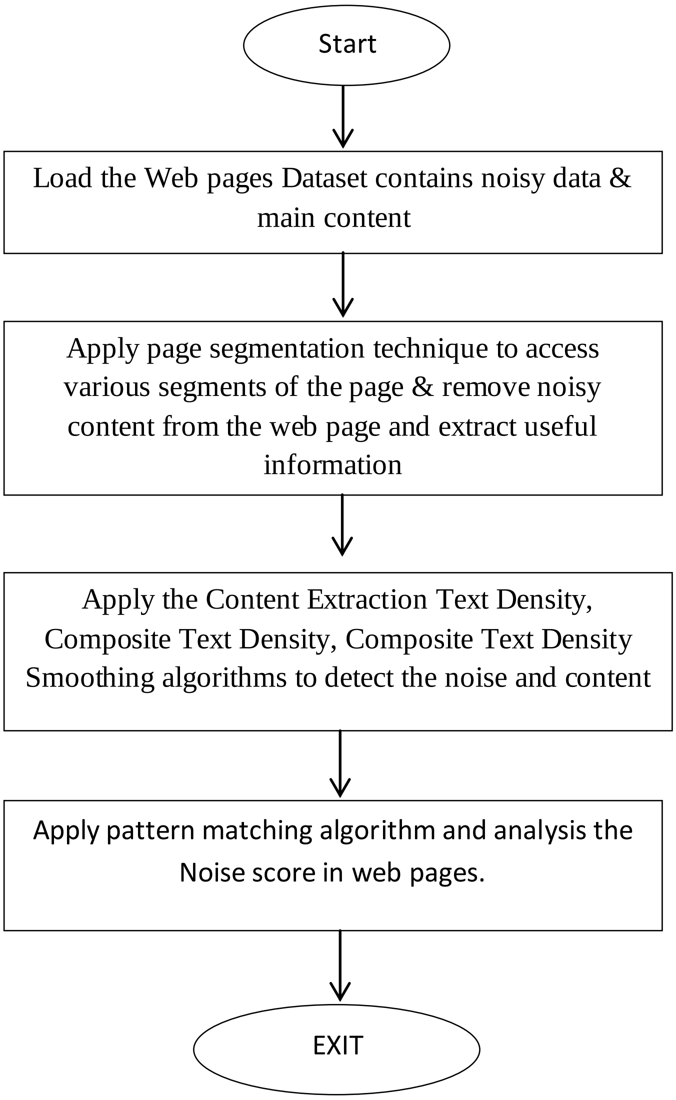

Start
Load the Web pages Dataset contains noisy data & main content
Apply page segmentation technique to access various segments of the page & remove noisy content from the web page and extract useful information
Apply the Content Extraction Text Density, Composite Text Density, Composite Text Density Smoothing algorithms to detect the noise and content
Apply pattern matching algorithm and analysis the Noise score in web pages.
EXIT

Fig.3.4.2 Proposed Methodology Flow Chart

The Content Extraction Text Density, Content Extraction Composite Text Density, Content Extraction Composite Text Density Smoothing and Composite Text Density Smoothing with Pattern Matching are computed Textual information and Noisy information from web pages.

In this work, a page segmentation approach is proposed with composite text density technique to extract the textual information of the web page. It can detect the noise information and retrieve the textual main information effectively and provide the essential content block.

CHAPTER 4
RESULTS AND DISCUSSIONS

4.1 MATLAB R2010A TOOL

MATLAB is “Matrix Laboratory”. We can perform image enhancement, noise reduction, image segmentation in Matlab. It is an interactive program which provides numerical computation and visualization of data. With the help of its programming capabilities it provides tool which is very useful for all areas of science and engineering.

a) **STANDERED WINDOWS IN MATLAB R2010A**

- Command window: This is main window where you can type and execute the commands.
- Workspace Window: In this window we can edit variables, load variables and clear variable.
- History window: In history window commands are re executed by double clicking and we can see the previously executed commands.

b) **GUIDE TOOL IN MATLAB R2010A**

Various tools are available in Matlab R2010A for creating the GUI interface. In our implementation we use the Guide tool and creating the GUI interface for Noise calculations. It is a Graphical tool where we can easy create the text box, check box, Buttons, Labels Changing properties of interface according our requirements.

4.2 RESULT AND DISCUSSION

In order to test our methodology, we organised various execution on the sequence of work. Our proposed Web Content Extraction and Noise detection using Page Segmentation with Composite Text Density to differentiates the noise and content from various web pages. Depending on different web pages like yahoo, BBC News, Wikipedia our approach produce wide variety of outputs. Most of Noise content like Navigational bars, decorated pictures, logos and irrelevant elements are detected reduced in size.

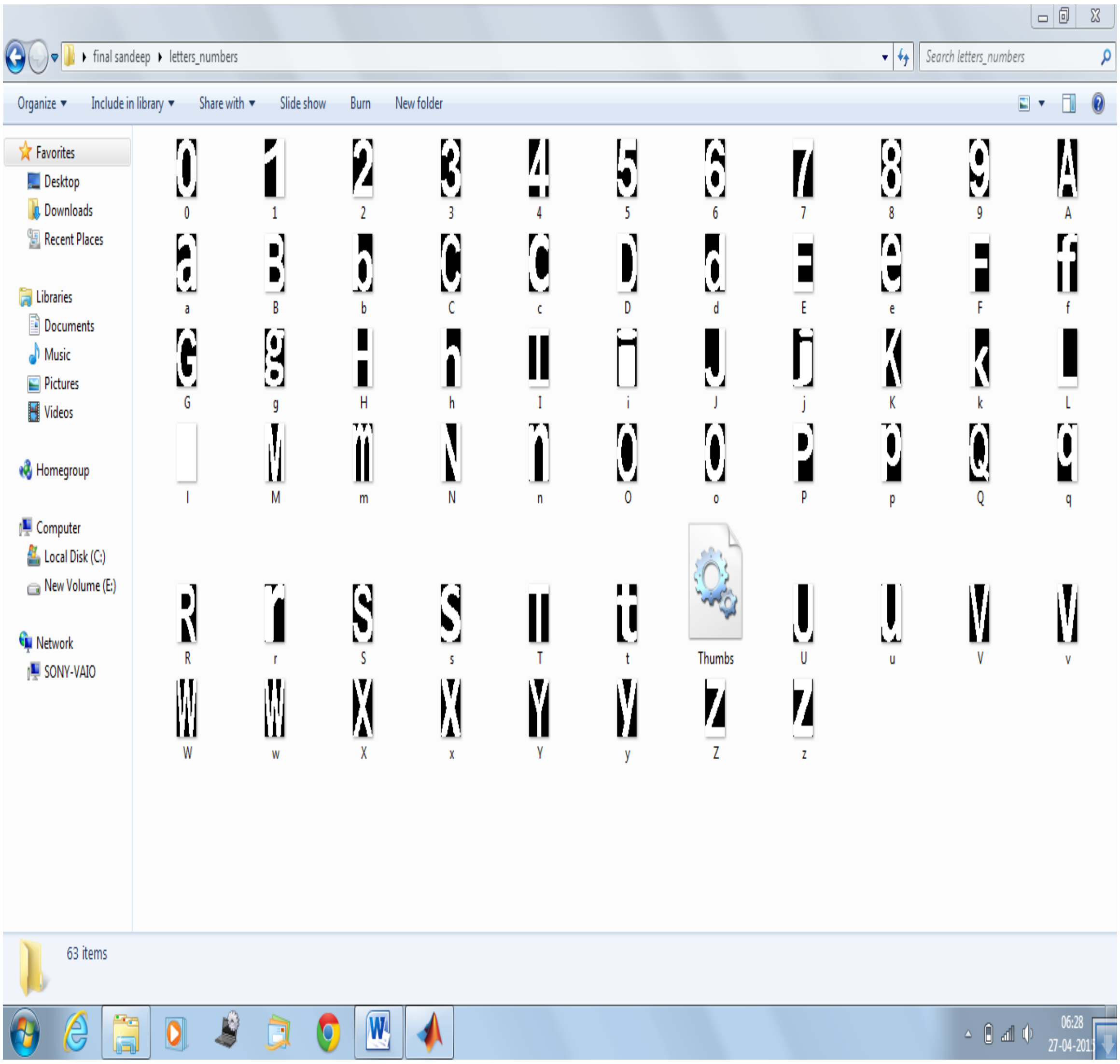


Fig.4.2.1 Template for Content Extraction and Noise Detection

In the Figure.4.2.1.Upper and Lower case letter template is load for content extraction using page segmentation. There are following steps performed for load and create the template.

Step1: Template Creation

In first step template different upper and lower letters with numeric zero to nine numbers create in Matlab.

Step2: Template images reads for Noise detections.

In second step function is create for call the different upper and lower numerical data images which match the text in web page for page segmentation. For reading the template images imread inbulit matlab function is used to load the different upper, lower and numeric data set and web page images for content extraction and noise detection process.

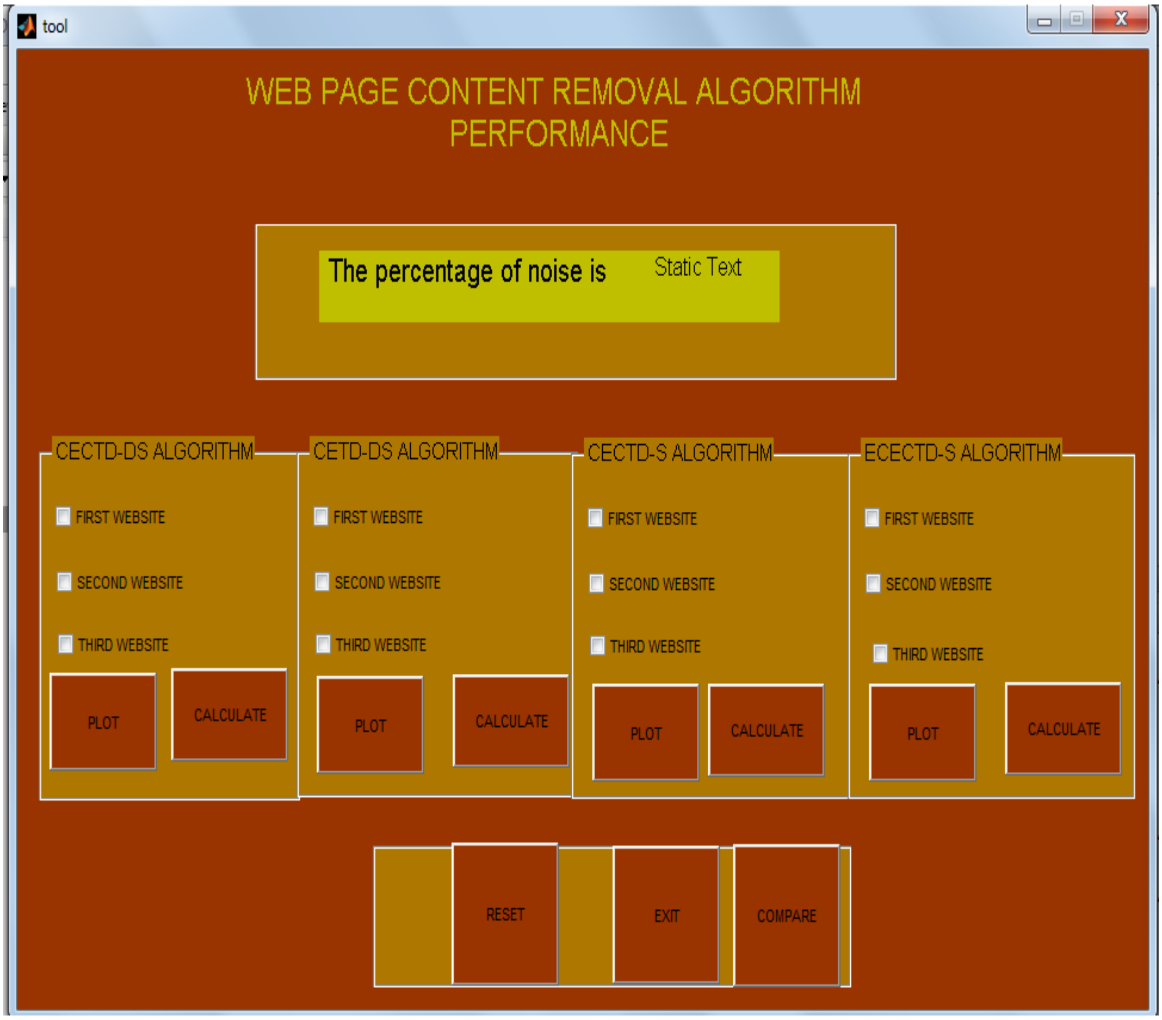
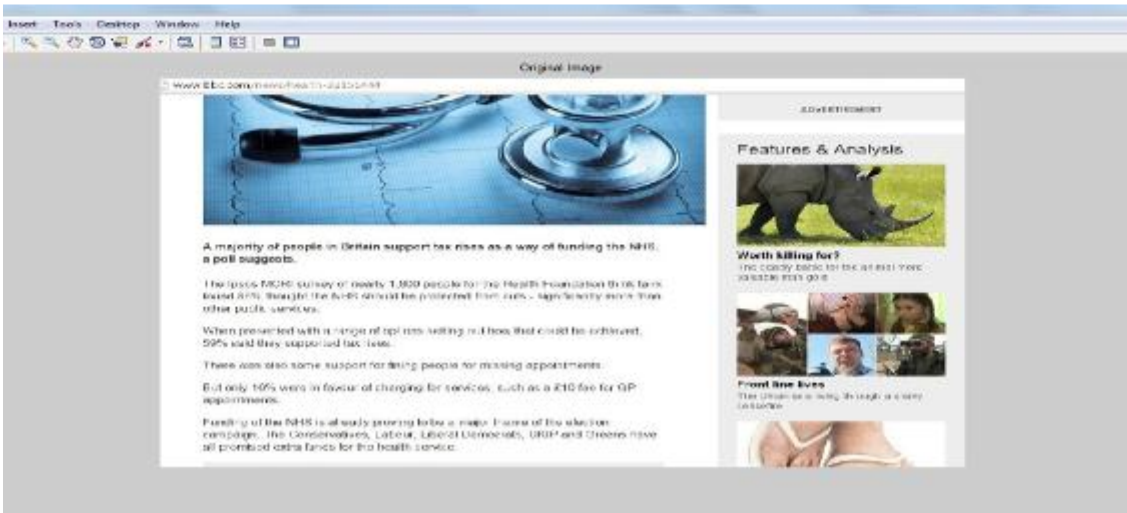


Fig.4.2.2 GUI interface to calculate the Noise and Content

The Figure 4.2 shows the Graphical User interface which is created in matlab using Guide tool for calculating the noise and content in Different Web pages. In the GUI interface checkbox, labels, Text Box, Buttons are used for selecting the options and submit, display the result and plots graph based on calculations.

As display in the GUI interface three WEBSITE are used for calculate the noise and content in Web Pages. Four algorithms Content Extraction Composite Text Density (CECTD-DS), Content Extraction Text Density –Density Sum (CETD-DS), Content Extraction Composite Text Density–Smoothing (CECTD-S), Extended Content Extraction Composite Text Density-Smoothing (ECECTD-S) are used for calculate the Noise and content in different Web Pages. Calculate tab used to analysis the result of selected WEBSITE and algorithm and PLOT tab used to generate the graph based on calculations. Compare tab used to provide the comparison result of selected WEBSITE and four algorithms.



A majority of people in Britain support tax rises as a way of funding the NHS, a poll suggests.

The Ipsos MORI survey of nearly 1,000 people for the Health Foundation think tank found [illegible]% thought the NHS should be protected from cuts - significantly more than other public services.

When presented with a range of options setting out how that could be achieved, 59% said they supported tax rises.

There was also some support for fining people for missing appointments.

But only 16% were in favour of charging for services, such as a £10 fee for GP appointments.

Funding of the NHS is already proving to be a major theme of the election campaign. The Conservatives, Labour, Liberal Democrats, UKIP and Greens have all promised extra funds for the health service.

Fig.4.2.3 BBC Web Page before Content Extraction

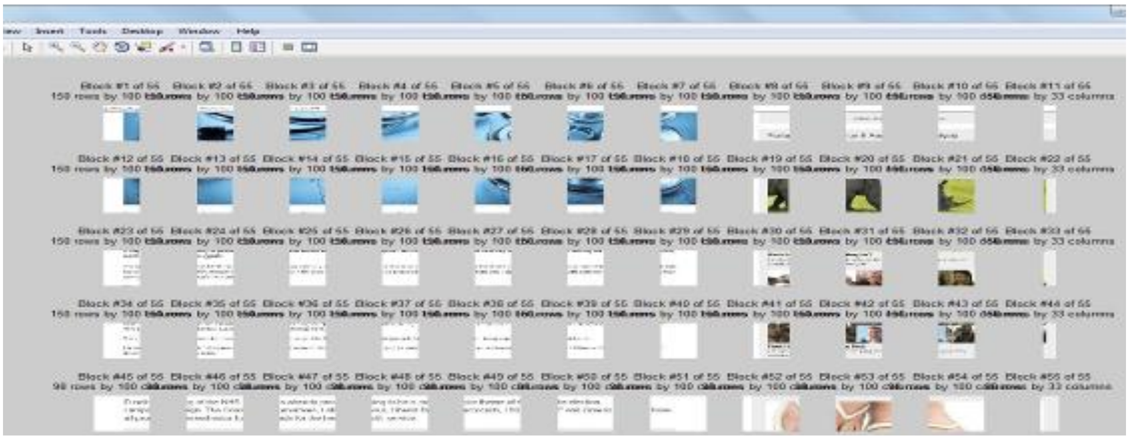
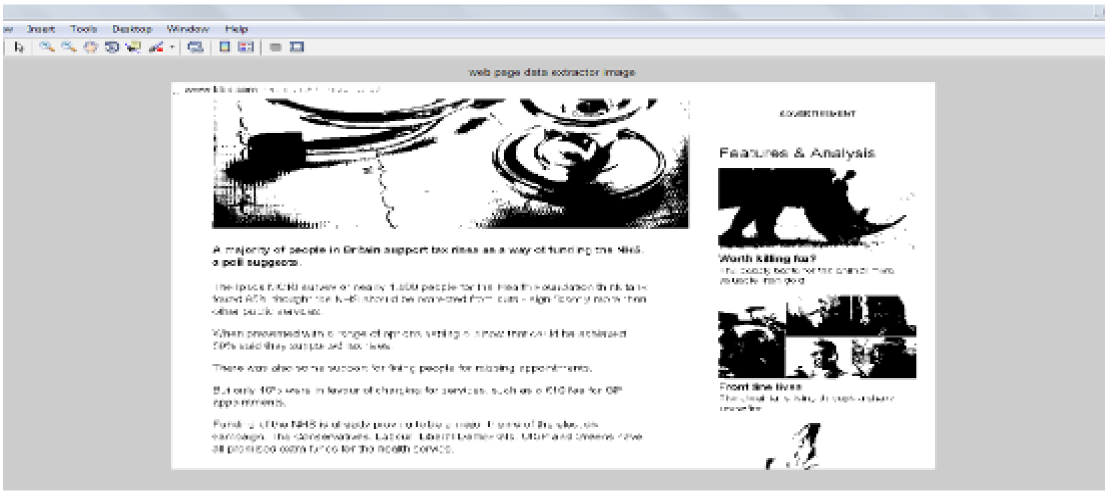


Fig.4.2.4. BBC Web Page Segmentation



A majority of people in Britain support tax rises as a way of funding the NHS, a poll suggests.

The Ipsos MORI survey of nearly 1,000 people for the Health Foundation think tank found [illegible]% thought the NHS should be protected from cuts - significantly more than other public services.

When presented with a range of options setting out how that could be achieved, 59% said they supported tax rises.

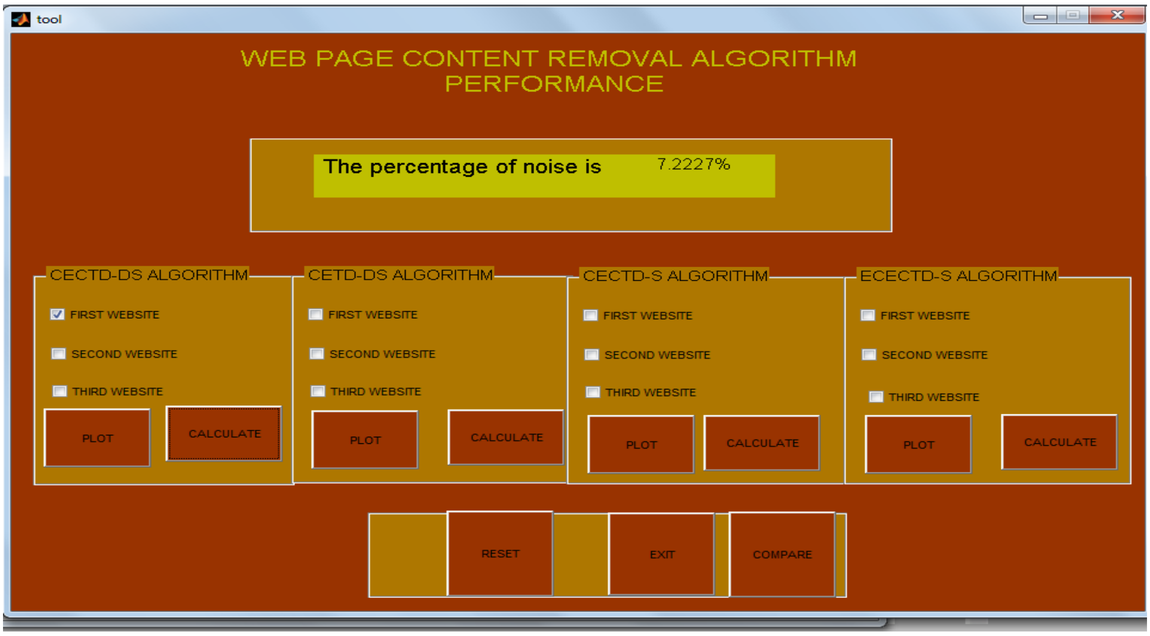
There was also some support for fining people for missing appointments.

But only 16% were in favour of charging for services, such as a £10 fee for GP appointments.

Funding of the NHS is already proving to be a major theme of the election campaign. The Conservatives, Labour, Liberal Democrats, UKIP and Greens have all promised extra funds for the health service.

Fig.4.2.5 BBC Page after Content Extraction

The Fig.4.2.3 shows the BBC Web Page with main content and noisy content. Page segmentation algorithm used to split the web page in Fig.4.2.4various blocks. In the next step content extraction and noise detection perform shows in fig.4.2.5.



The percentage of noise is 7.2227%

CECTD-DS ALGORITHM
☑ FIRST WEBSITE
□ SECOND WEBSITE
□ THIRD WEBSITE

CETD-DS ALGORITHM
□ FIRST WEBSITE
□ SECOND WEBSITE
□ THIRD WEBSITE

CECTD-S ALGORITHM
□ FIRST WEBSITE
□ SECOND WEBSITE
□ THIRD WEBSITE

ECECTD-S ALGORITHM
□ FIRST WEBSITE
□ SECOND WEBSITE
□ THIRD WEBSITE

Fig.4.2.6 BBC Page Noise and Content Analysis using CECTD-DS Algorithm

In Fig.4.2.6 We select the First web site and Content Extraction Composite Text Density-Density Sum algorithm used to calculate the Noise and Content in Web Page. With the help of Composite text density algorithm we calculate the starting date of web page loaded and check the updated information in web page

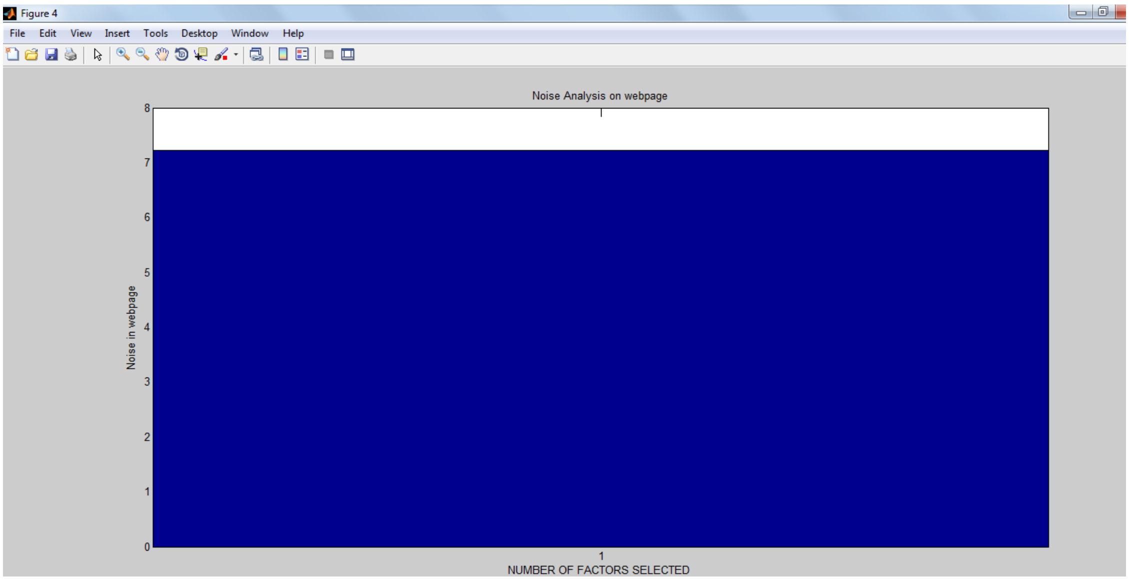
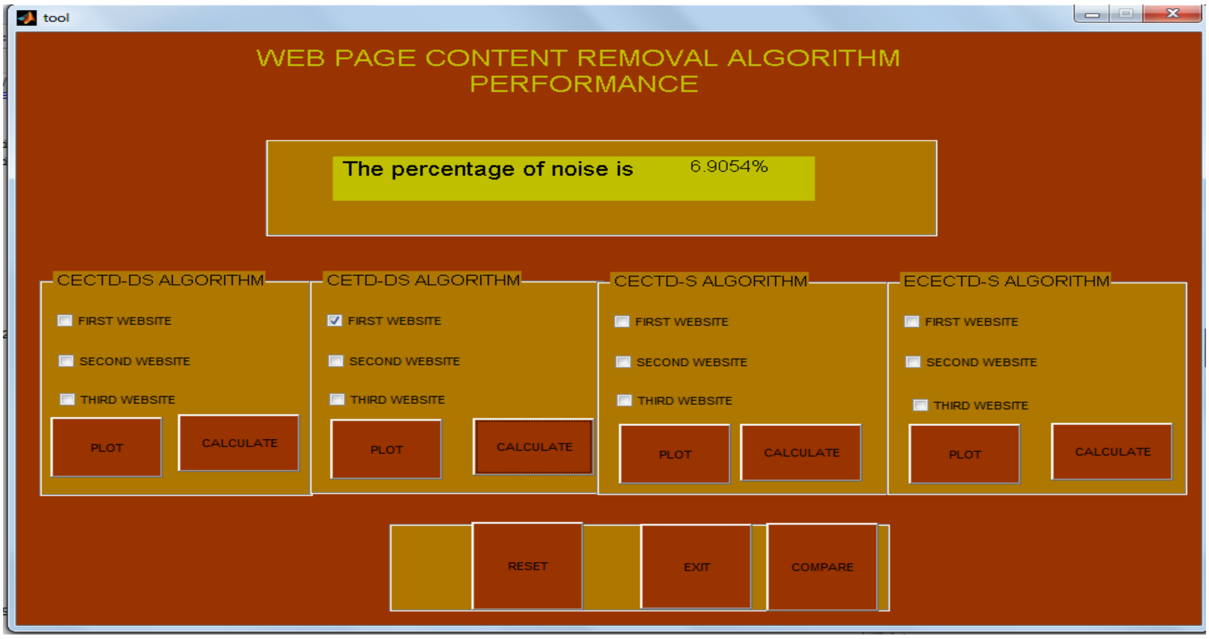


Fig.4.2.7 BBC Page Noise and Content Analysis using CECTD-DS Graph

In Fig.4.2.7 First BBC Web Page result analysis in graphical Order based on the Noise and Content analysis using CECTD-DS algorithm. After calculation show the 7.2227 analysis result based on the CECTD-DS algorithm.



WEB PAGE CONTENT REMOVAL ALGORITHM PERFORMANCE

The percentage of noise is 6.9054%

CECTD-DS ALGORITHM
- □ FIRST WEBSITE
- □ SECOND WEBSITE
- □ THIRD WEBSITE

CETD-DS ALGORITHM
- ☑ FIRST WEBSITE
- □ SECOND WEBSITE
- □ THIRD WEBSITE

CECTD-S ALGORITHM
- □ FIRST WEBSITE
- □ SECOND WEBSITE
- □ THIRD WEBSITE

ECECTD-S ALGORITHM
- □ FIRST WEBSITE
- □ SECOND WEBSITE
- □ THIRD WEBSITE

Fig.4.2.8 BBC Web Page Noise Analysis using CETD-DS Algorithm

In Fig.4.2.8 We select the First web site and Content Extraction Text Density-Density Sum algorithm used to calculate the Noise and Content in Web Page. The Text density calculation is done based on the white space and text length in Web Page.

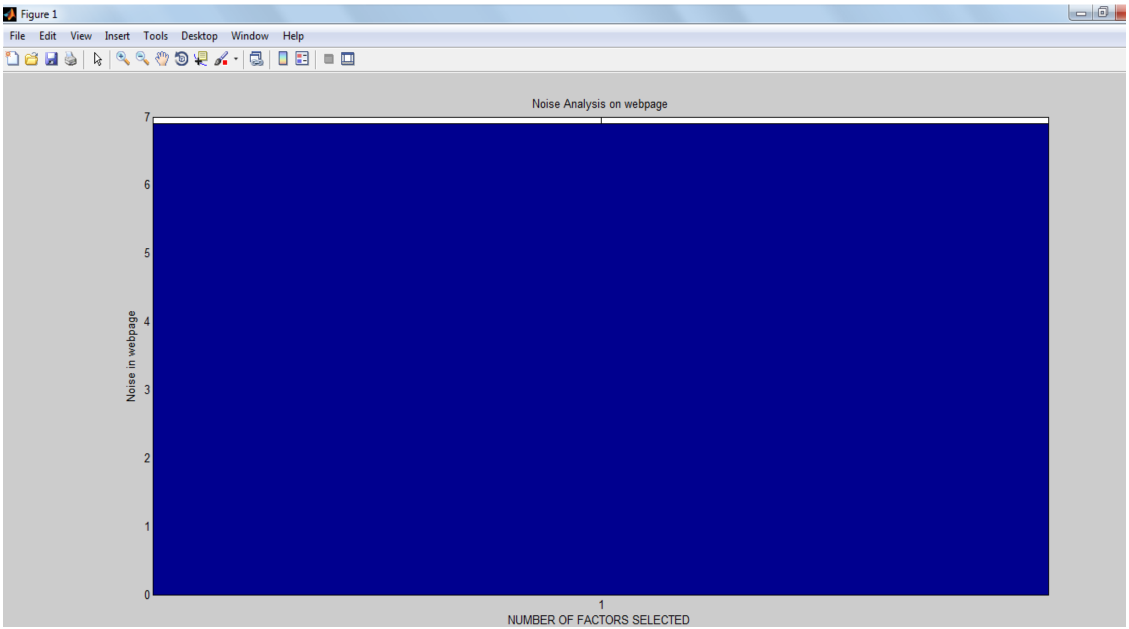
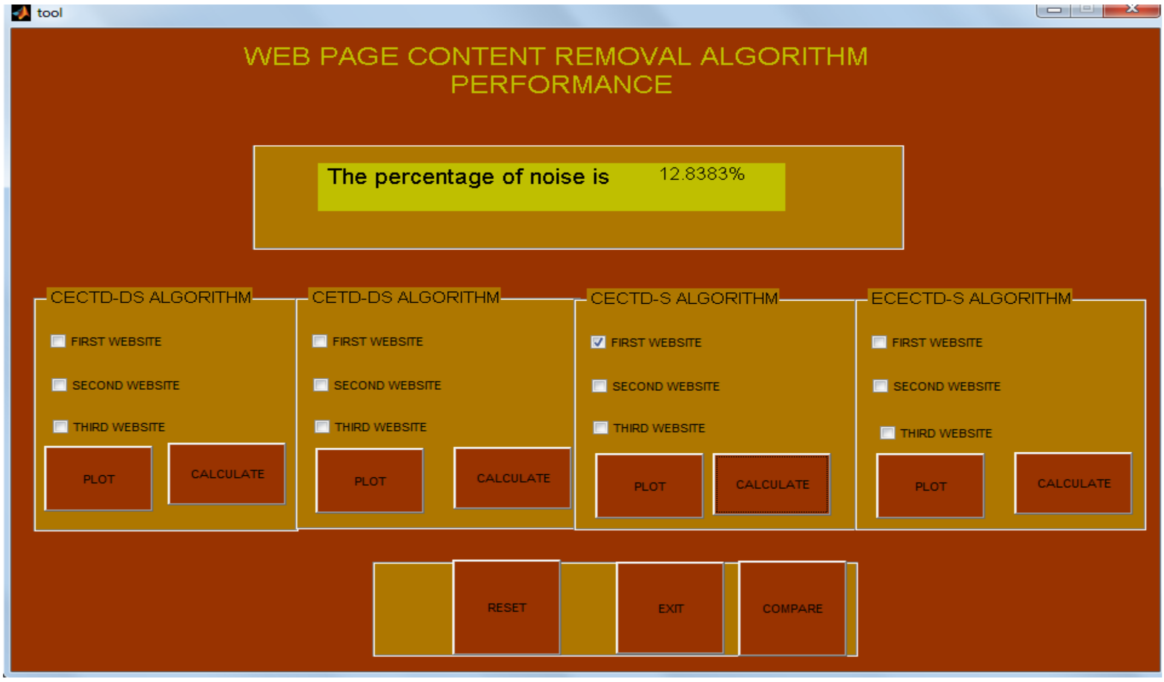


Fig.4.2.9 BBC Web Page Noise Analysis using CETD-DS Algorithm Graph

In Fig.4.2.9 First BBC Web Page result analysis in graphical Order based on the Noise and Content analysis using CETD-DS algorithm**.** After calculation show the 6.9054 analysis result based on the CETD-DS algorithm.



WEB PAGE CONTENT REMOVAL ALGORITHM PERFORMANCE

The percentage of noise is 12.8383%

CECTD-DS ALGORITHM
- ☐ FIRST WEBSITE
- ☐ SECOND WEBSITE
- ☐ THIRD WEBSITE
- PLOT | CALCULATE

CETD-DS ALGORITHM
- ☐ FIRST WEBSITE
- ☐ SECOND WEBSITE
- ☐ THIRD WEBSITE
- PLOT | CALCULATE

CECTD-S ALGORITHM
- ☑ FIRST WEBSITE
- ☐ SECOND WEBSITE
- ☐ THIRD WEBSITE
- PLOT | CALCULATE

ECECTD-S ALGORITHM
- ☐ FIRST WEBSITE
- ☐ SECOND WEBSITE
- ☐ THIRD WEBSITE
- PLOT | CALCULATE

Fig.4.2.10 BBC Web Page Noise Analysis using CECTD-S Algorithm

In Fig.4.2.10 We select the First web site and Content Extraction Composite Text Density-Smoothing algorithm used to calculate the Noise and Content in Web Page. Using CECTD-S Algorithm we calculate hyperlinks in the web page and calculated the hyperlink and advertisement noise from web pages.

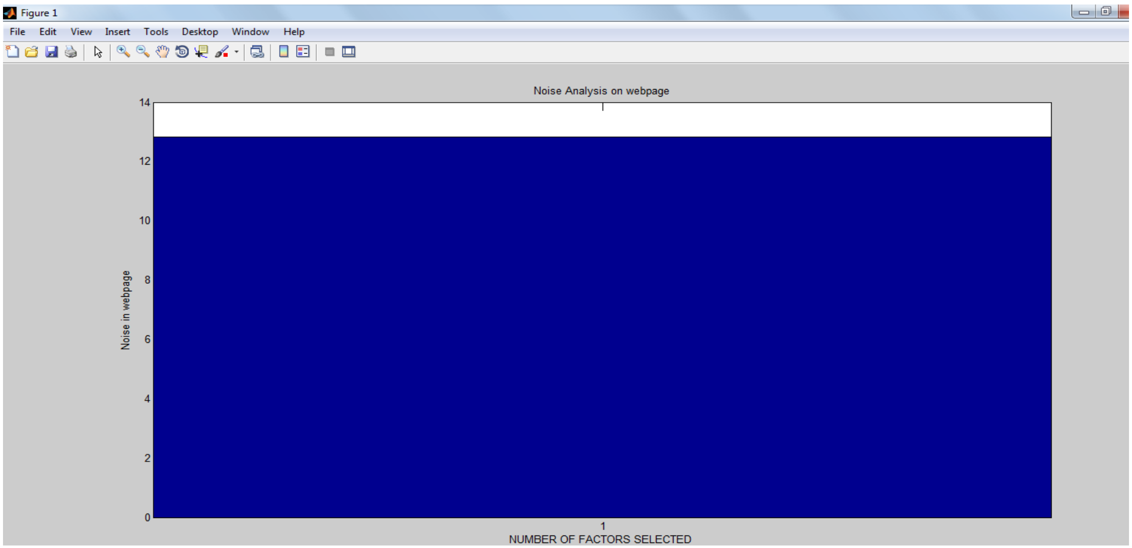
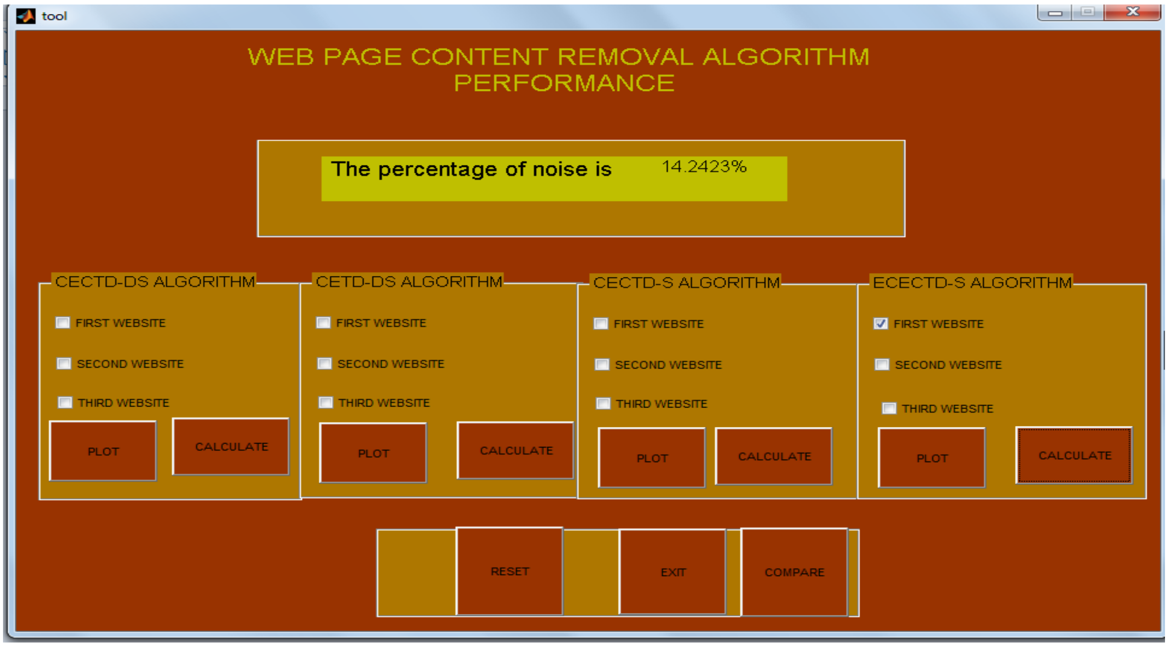


Fig.4.2.11 BBC Web Page Noise Analysis using CECTD-S Algorithm Graph

In Fig.4.2.11 First BBC Web Page result analysis in graphical Order based on the Noise and Content analysis using CECTD-DS algorithm. After calculation show the 12.8383 analysis result based on the CECTD-DS algorithm.



WEB PAGE CONTENT REMOVAL ALGORITHM PERFORMANCE

The percentage of noise is 14.2423%

CECTD-DS ALGORITHM

□ FIRST WEBSITE

□ SECOND WEBSITE

□ THIRD WEBSITE

CETD-DS ALGORITHM

□ FIRST WEBSITE

□ SECOND WEBSITE

□ THIRD WEBSITE

CECTD-S ALGORITHM

□ FIRST WEBSITE

□ SECOND WEBSITE

□ THIRD WEBSITE

ECECTD-S ALGORITHM

☑ FIRST WEBSITE

□ SECOND WEBSITE

□ THIRD WEBSITE

Fig.4.2.12 BBC Web Page Noise Analysis using ECECTD-S Algorithm

In Fig.4.2.10 We select the First web site and Extended Content Extraction Composite Text Density- Smoothing algorithm used to calculate the Noise and Content in Web Page. ECECTD-S calculated the updated and non-updated hyperlinks and advertisement noise under the <div> tag in the web page.

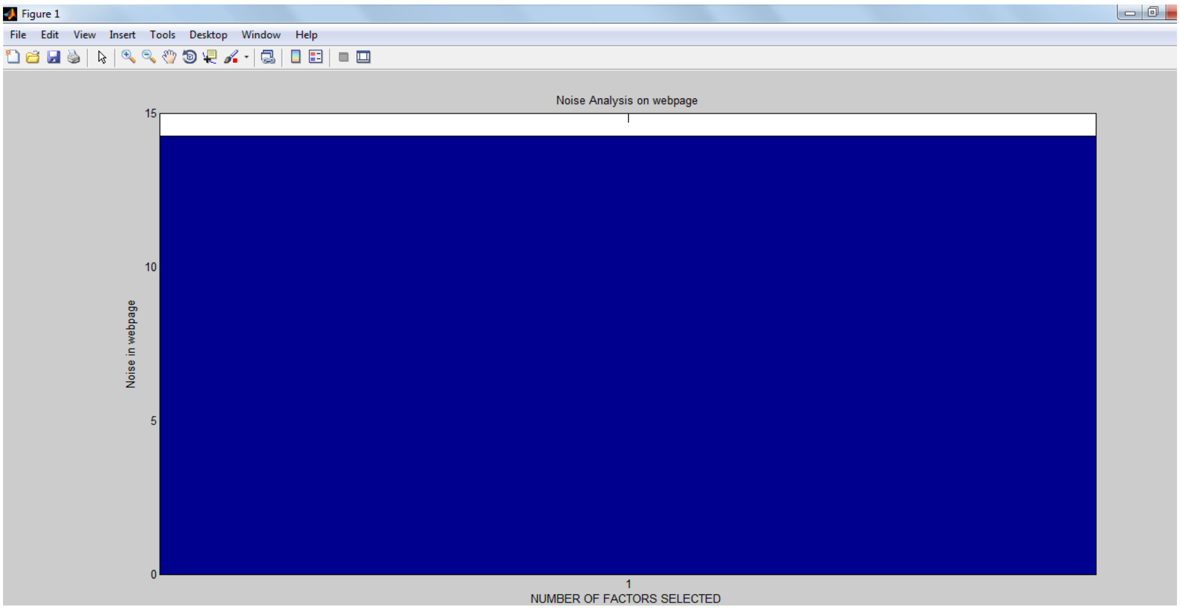


Fig.4.2.13 BBC Web Page Noise Analysis using ECECTD-S Algorithm Graph

In Fig.4.2.11 Fourth BBC Web Page result analysis in graphical Order based on the Noise and Content analysis using ECECTD-S algorithm**.** After calculation show the 14.2423 analysis result based on the ECECTD-S algorithm.

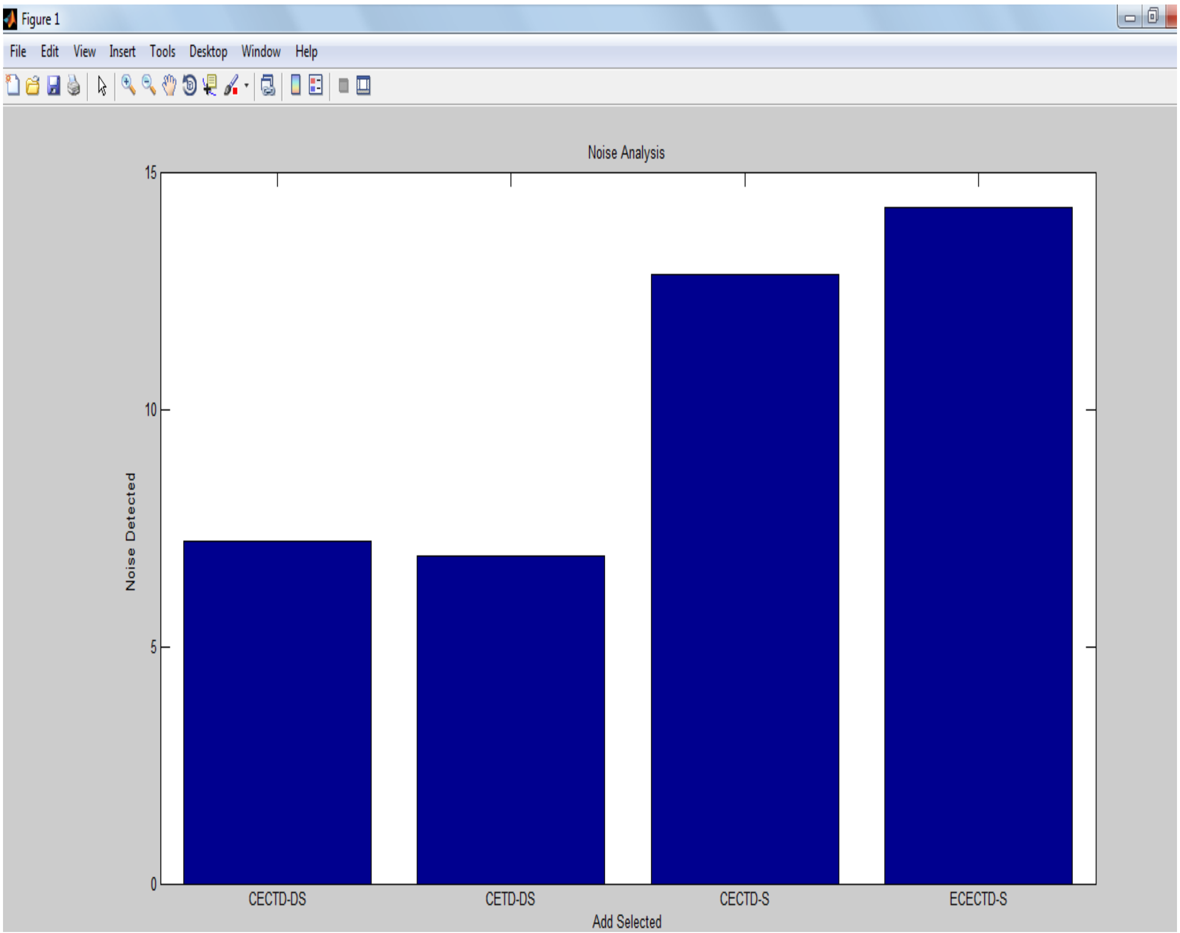


Fig.4.2.14 BBC Page Noise and Content Analysis Comparison of four Algorithms

In Fig.4.2.14 Four Algorithms in Content Extraction Composite Text Density and Density Sum (CECTD-DS), Content Extraction Text Density and Density Sum (CETD-DS), Content Extraction Composite Text Density and Smoothing (CECTD-S), Extended Content Extraction Composite Text Density and smoothing (ECECTD-S) are applied on BBC Web Page to analysis the noise and content in web page and represent the graphical results and compared result of four algorithms.

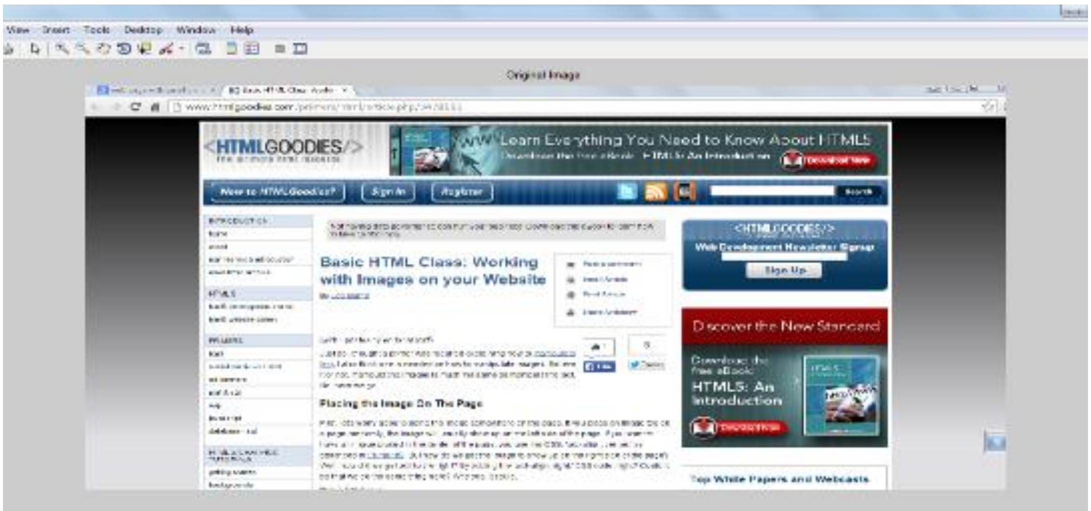

Original Image
HTMLGOODIES
Learn Everything You Need to Know About HTML5
New to HTMLGoodies?
Sign In
Register
Basic HTML Class: Working with Images on your Website
Placing the Image On The Page
Discover the New Standard
HTML5: An Introduction
Top White Papers and Webcasts

Fig.4.2.15 HTML Web Page before Content Extraction

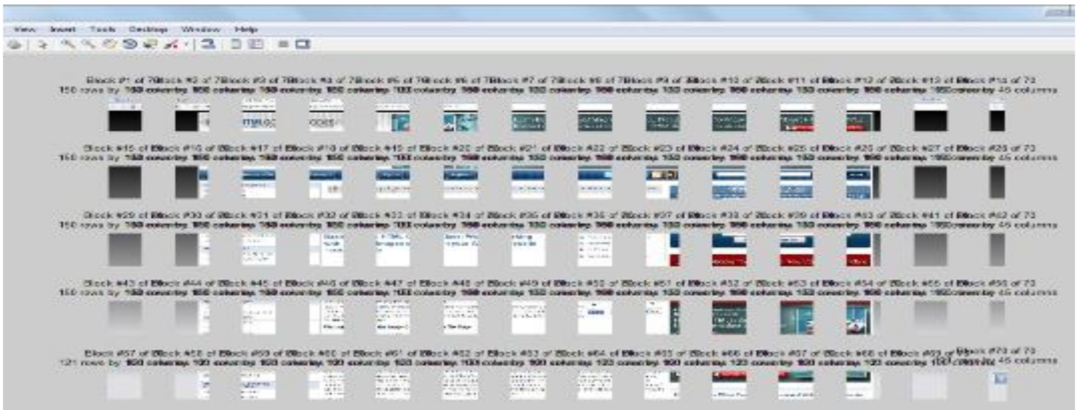


Fig.4.2.16 HTML Web Page segmentation

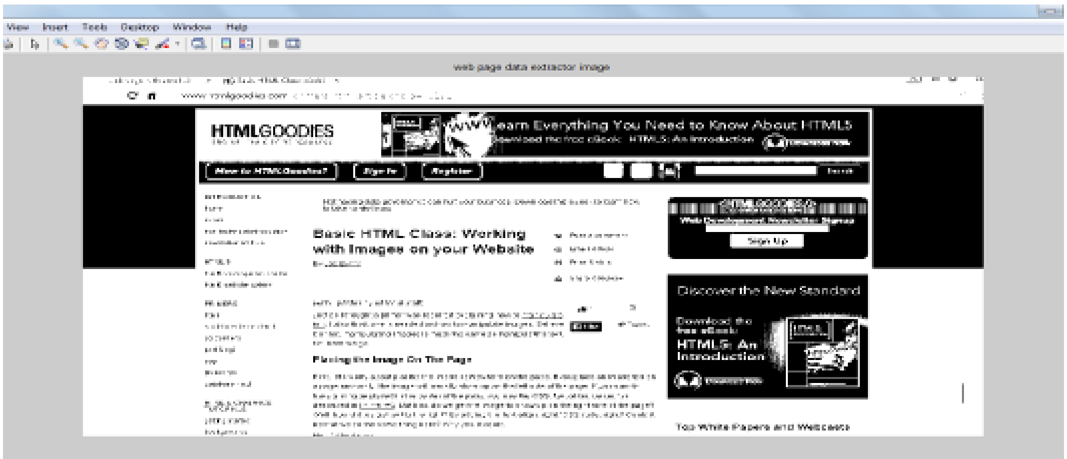
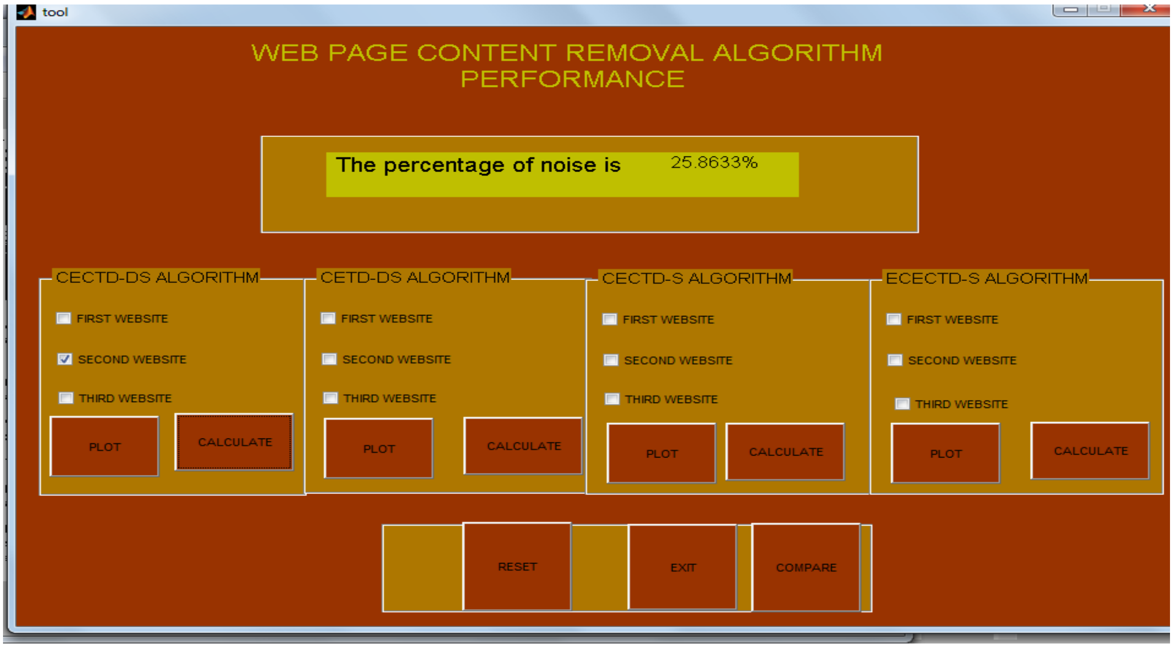

web page data extractor image
HTMLGOODIES
Learn Everything You Need to Know About HTML5
New to HTMLGoodies?
Sign In
Register
Basic HTML Class: Working with Images on your Website
Placing the Image On The Page
Discover the New Standard
HTML5: An Introduction
Top White Papers and Webcasts

Fig.4.2.17 HTML Web Page after Content Extraction

The Fig.4.2.15 shows the other html Web Page with main content and advertisements. Page segmentation algorithm used to split the web page in blocks Fig.4.2.16. In the next step content extraction and noise and content detection perform shows in fig.4.2.17.



The percentage of noise is 25.8633%

Fig.4.2.18 HTML Page Noise and Content Analysis using CECTD-DS Algorithm

In Fig.4.2.18 We select the second web site and Content Extraction Composite Text Density-Density Sum algorithm used to calculate the Noise and Content in Web Page. With the help of Composite text density algorithm we calculate the starting date of web page loaded and check the updated information in web page.

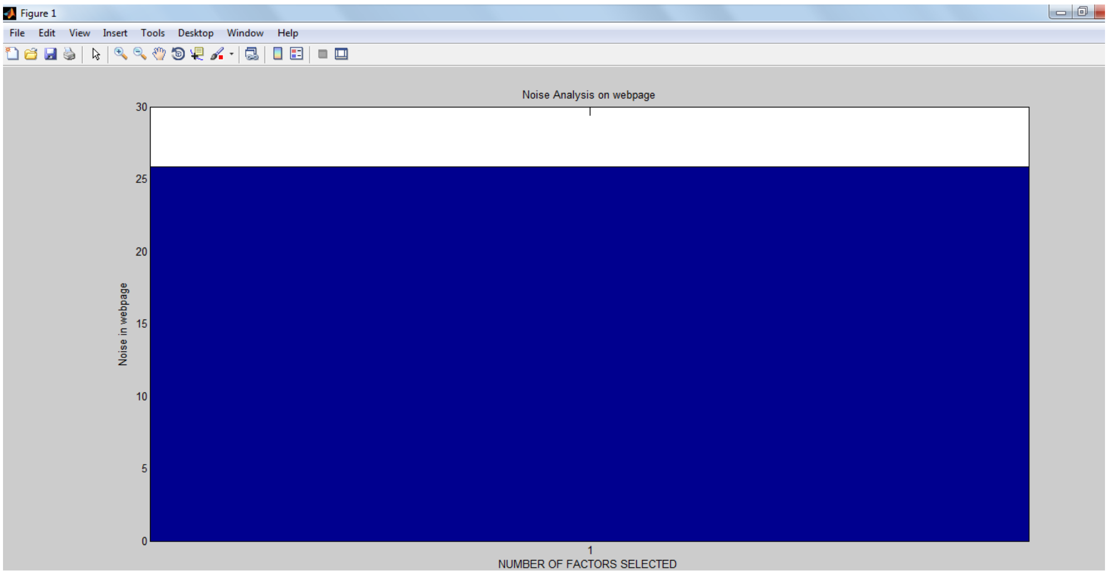
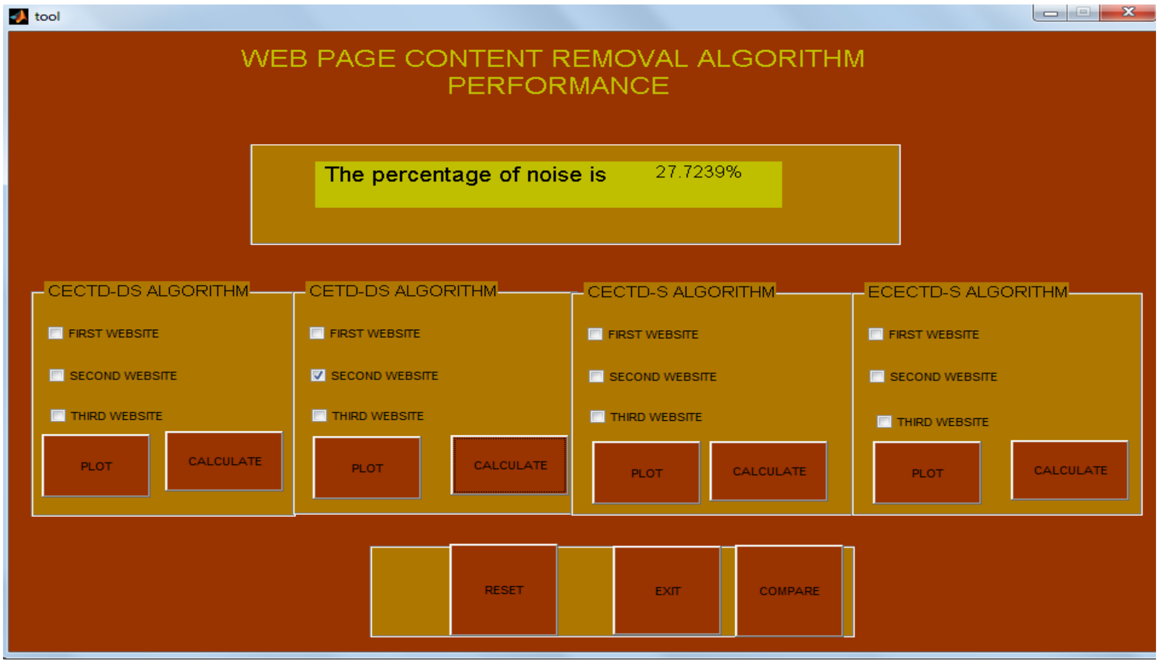


Fig.4.2.19 HTML Page Noise and Content Analysis using CECTD-DS Graph

In Fig.4.2.19 Second HTML Web Page result analysis in graphical Order based on the Noise and Content analysis using CECTD-DS algorithm**.** After calculation show the 25.8633 analysis result based on the CECTD-DS algorithm.



WEB PAGE CONTENT REMOVAL ALGORITHM PERFORMANCE

The percentage of noise is 27.7239%

CETD-DS ALGORITHM: ☑ SECOND WEBSITE

Fig.4.2.20 HTML Page Noise and Content Analysis using CETD-DS Algorithm

In Fig.4.2.20 We select second web site and Content Extraction Text Density-Density Sum algorithm used to calculate the Noise and Content in Web Page. The Text density calculation is done based on the white space and text length in Web Page.

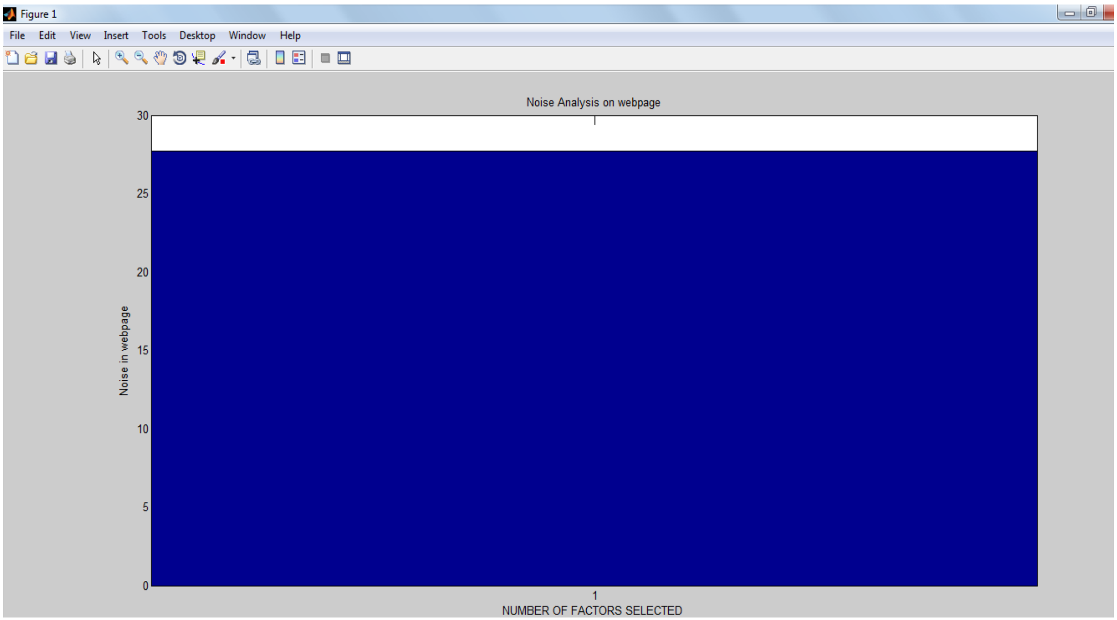
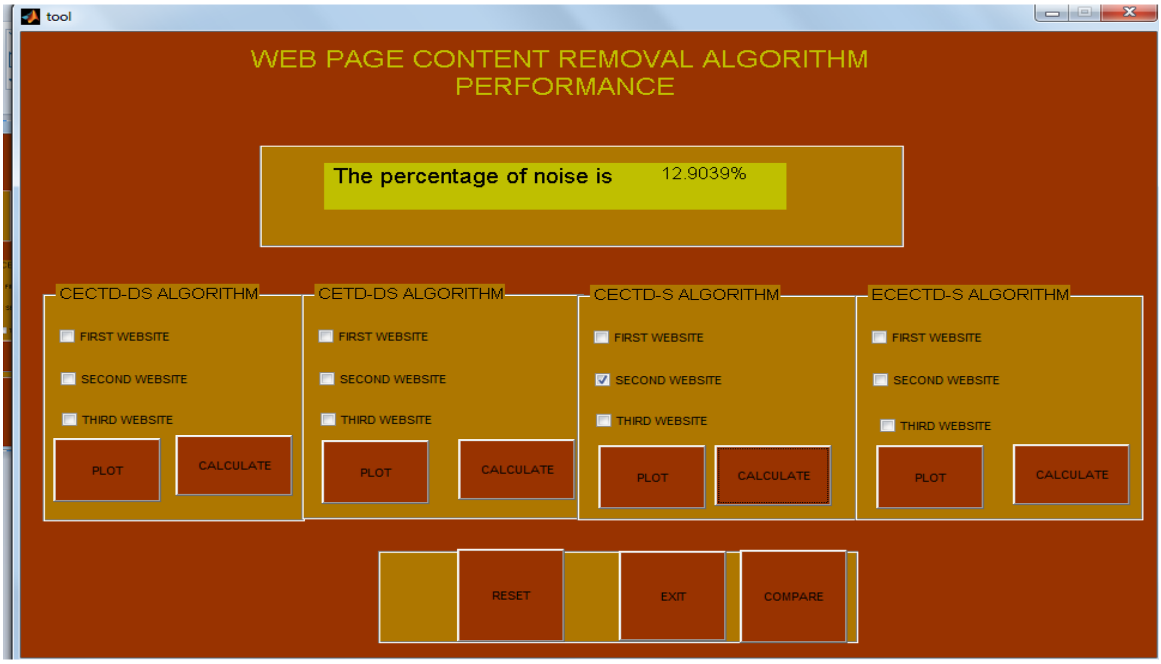


Fig.4.2.21 HTML Page Noise and Content Analysis using CETD-DS Graph

In Fig.4.2.21 Second HTML Web Page result analysis in graphical Order based on the Noise and Content analysis using CETD-DS algorithm**.** After calculation show the 27.7239 analysis result based on the CETD-DS algorithm.



WEB PAGE CONTENT REMOVAL ALGORITHM PERFORMANCE

The percentage of noise is 12.9039%

CECTD-DS ALGORITHM

□ FIRST WEBSITE

□ SECOND WEBSITE

□ THIRD WEBSITE

CETD-DS ALGORITHM

□ FIRST WEBSITE

□ SECOND WEBSITE

□ THIRD WEBSITE

CECTD-S ALGORITHM

□ FIRST WEBSITE

☑ SECOND WEBSITE

□ THIRD WEBSITE

ECECTD-S ALGORITHM

□ FIRST WEBSITE

□ SECOND WEBSITE

□ THIRD WEBSITE

Fig.4.2.22 HTML Page Noise and Content Analysis using CECTD-S Algorithm

In Fig.4.2.22 We select the Second web site and Content Extraction Composite Text Density-Smoothing algorithm used to calculate the Noise and Content in Web Page. Using CECTD-S Algorithm we calculate hyperlinks in the web page and calculated the hyperlink and advertisement noise from web pages.

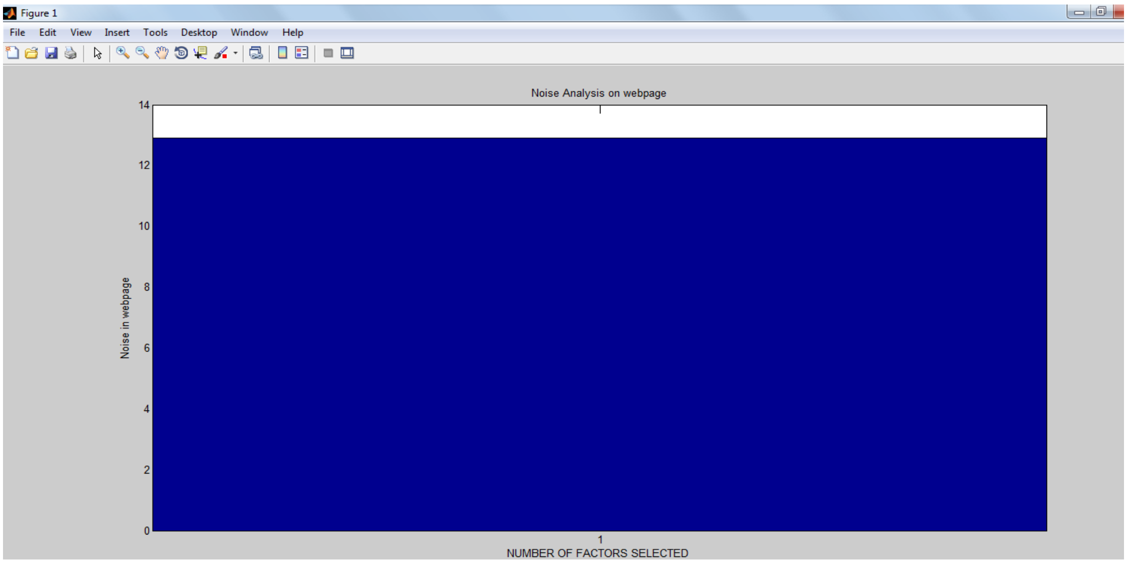
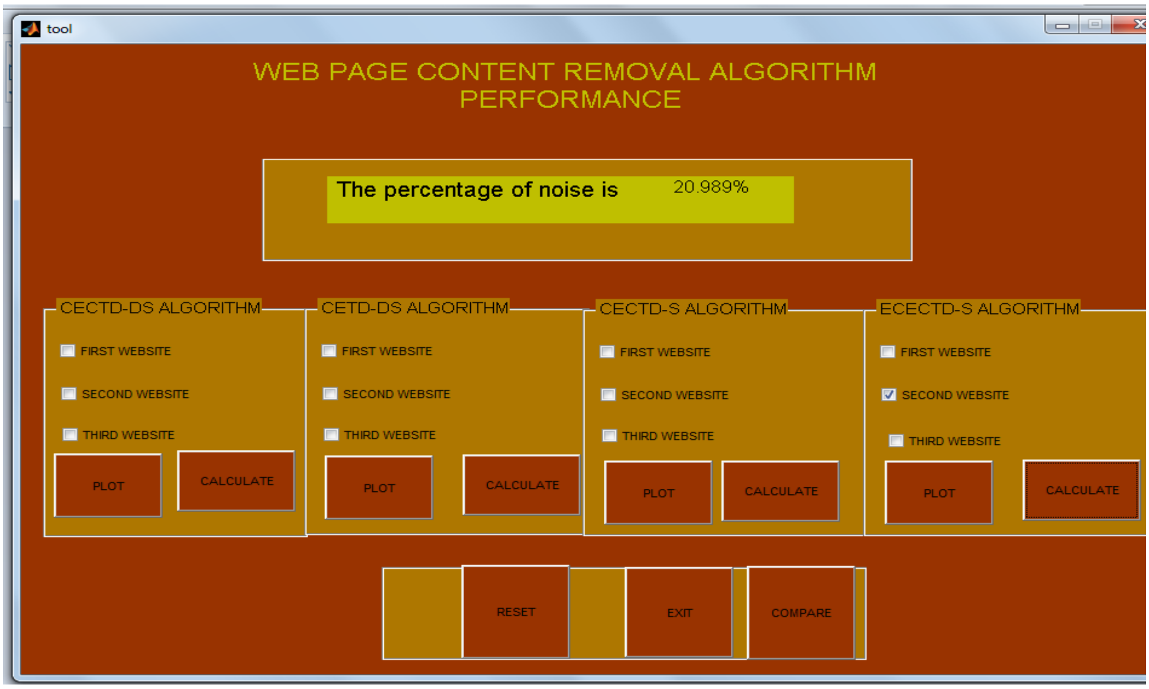


Figure.4.2.23 HTML Page Noise and Content Analysis using CECTD-S Graph

In Fig.4.2.23 Second HTML Web Page result analysis in graphical Order based on the Noise and Content analysis using CECTD-S algorithm. After calculation show the 12.9039 analysis result based on the CECTD-S algorithm



The percentage of noise is 20.989%

CECTD-DS ALGORITHM
□ FIRST WEBSITE
□ SECOND WEBSITE
□ THIRD WEBSITE

CETD-DS ALGORITHM
□ FIRST WEBSITE
□ SECOND WEBSITE
□ THIRD WEBSITE

CECTD-S ALGORITHM
□ FIRST WEBSITE
□ SECOND WEBSITE
□ THIRD WEBSITE

ECECTD-S ALGORITHM
□ FIRST WEBSITE
☑ SECOND WEBSITE
□ THIRD WEBSITE

Fig.4.2.24 HTML Page Noise and Content Analysis using ECECTD-S Algorithm

In Fig.4.2.24 We select the second web site and Extended Content Extraction Composite Text Density-Smoothing algorithm used to calculate the Noise and Content in Web Page. CTDPM calculated the updated and non-updated hyperlinks and advertisement noise under the <div> tag in the web page.

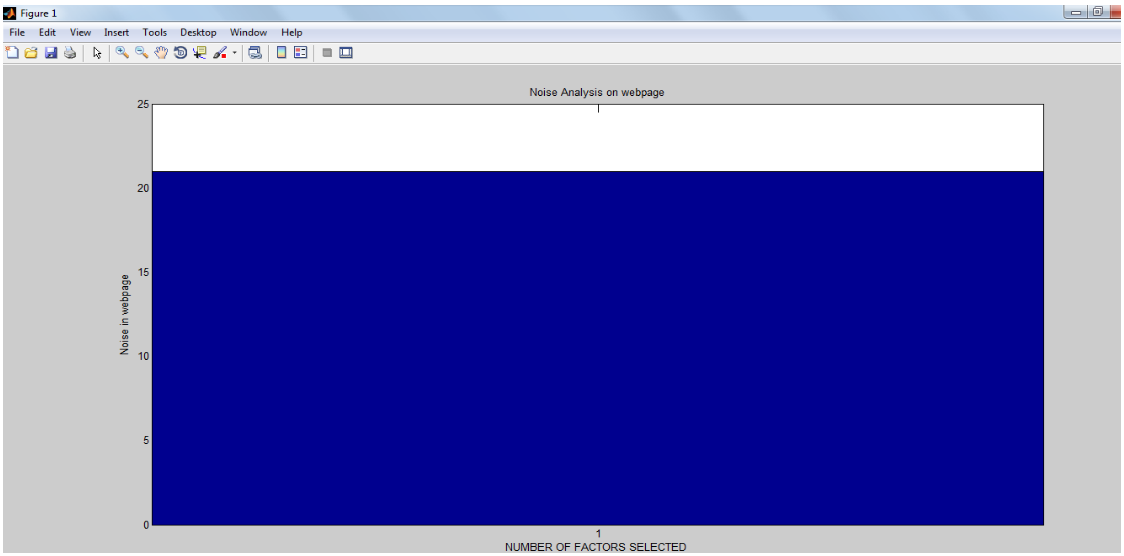


Fig.4.2.25. HTML Page Noise and Content Analysis using ECECTD-S Graph

In Fig.4.2.25 Second HTML Web Page result analysis in graphical Order based on the Noise and Content analysis using CECTD-S algorithm**.** After calculation show the 20.989 analysis result based on the CECTD-S algorithm.

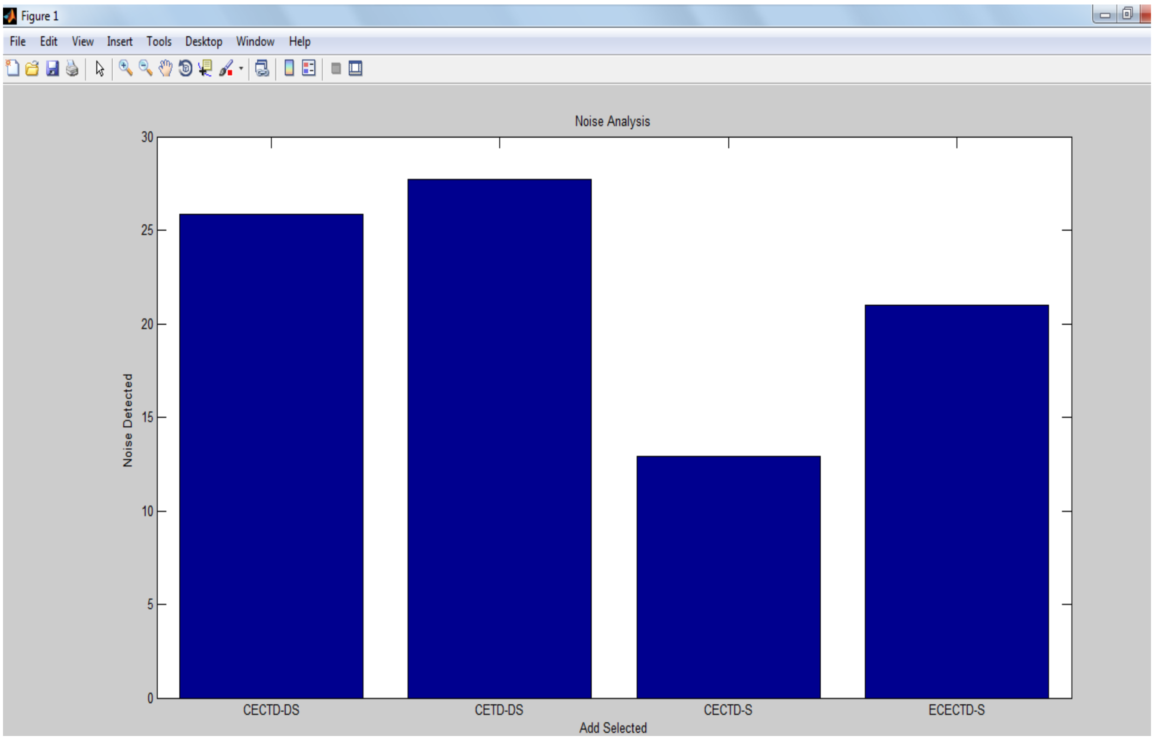


Fig.4.2.26 HTML Page Noise and Content Analysis Comparison of four Algorithms

In Fig.4.2.26 Four Algorithms in Content Extraction Composite Text Density and Density Sum (CECTD-DS), Content Extraction Text Density and Density Sum (CETD-DS), Content Extraction Composite Text Density and Smoothing (CECTD-S), Extended Content Extraction Composite Text Density and smoothing (ECECTD-S) are applied on HTML Web Page to analysis the noise in web page and represent the graphical results and compared result of four algorithms.

Web Pages	CECTD-DS	CETD-DS	CECTD-S	ECECTD-S
BBC	7.2227	6.9054	12.8383	14.2423
HTML	25.8633	27.7239	12.9039	20.989

Table.4.2.1. CECTD-DS, CETD-DS, CECTD-S, ECECTD-S algorithms Comparison

In the above Table.4.2.1 shows noise and content analysis on different web page like BBC, HTML, Wikipedia using four different CECTD-DS, CETD-DS, CETD-S, ECECTD-S algorithms. We implement our enhanced algorithm ECECTD which used pattern matching algorithm to detect the noise and content in web page and provide the better noise and content detection result as compared to other techniques.

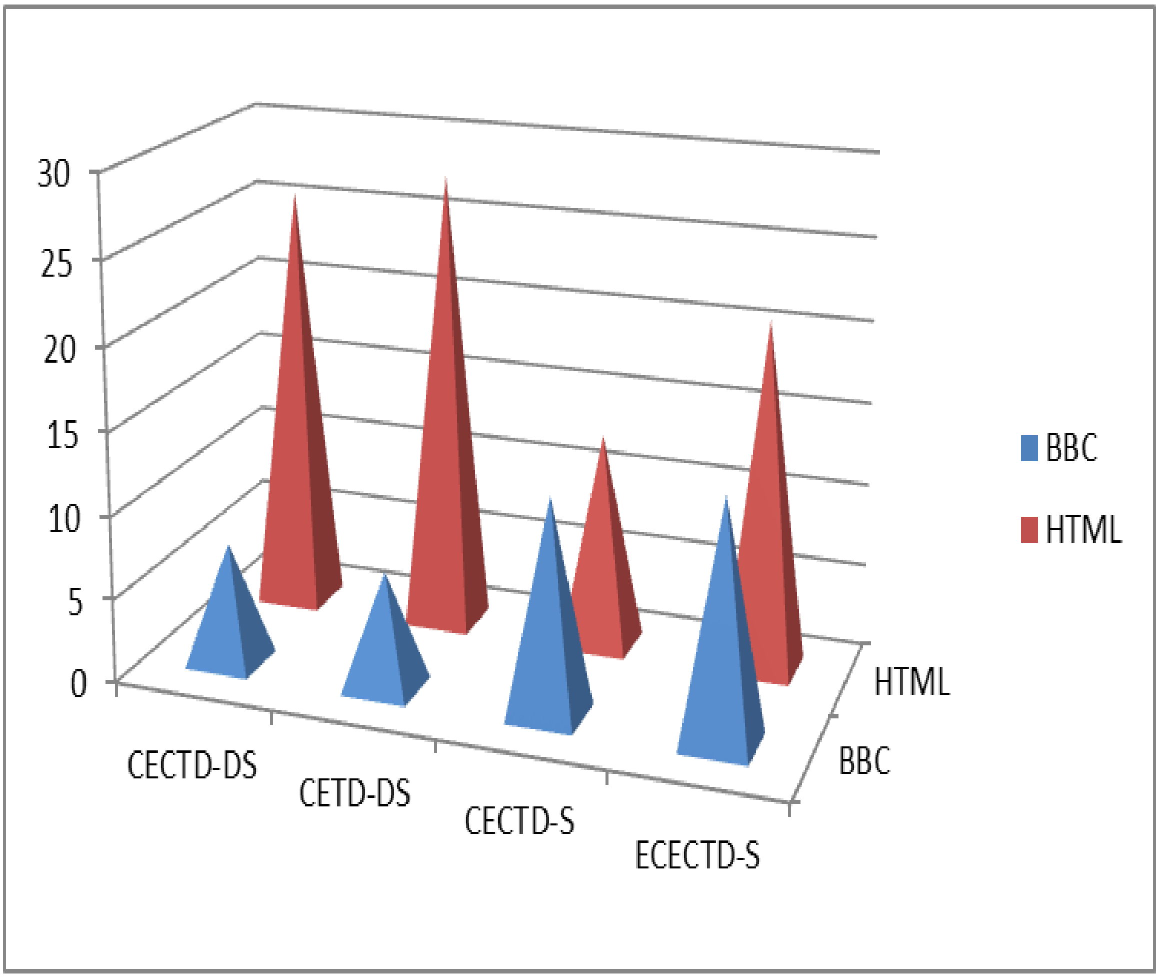


Fig.4.27 Result of four algorithms on BBC, HTML, Wikipedia Web pages

As shown in Fig.4.27 CECTD-DS, CETD-DS algorithms shows the text in web page based on Text Density and Density Sum and CETD-S, ECECTD-S show Noise based on advertisements and hyperlinks in Web Pages. As shown in graph in BBC Web page shows less text and high Noise in web page HTML Web Page show more text and low to moderate noise and Wikipedia show high text and less Noisy content in Web page. So our algorithms provide the different result based on different Web Page.

CHAPTER 5

CONCLUSION AND FUTURE WORK

CONCLUSION

A web page contains many types of noisy information, the user not interested to access this information and make the performance of PDA devices slow. The problem which is often occurs in web pages is that the main content mixed with unrelated content like navigational bar ,images, old versioned pages. To extract the text information including with original styling and information and eliminating images, figures, advertisements menu bars other unrelated text bars. A page segmentation text density and pattern matching approach is proposed to extract the content and analysis the composite text density Smoothing (CECTD-S) for detecting hyperlink images noise and Extended composite text density Smoothing with pattern matching (ECECTD-S) algorithm to detecting updated, non-updated hyperlinks and advertisements under the div tag. The result show that (ECECTD-S) algorithm and other algorithm noise score analysis. Using Page segmentation with composite text density algorithm just require a web page as input then returns a web page with extraction of text content with noise detection and also maintain the original structure of the web page. Our enhanced implemented technique provide the different and efficient result based on different Web Pages and detect the more noise and content in web page as compared to other techniques.

FUTURE WORK

As mentioned above Page segmentation technique content extraction and noise detection are carried out based on our proposed text density approach. There are still some issues to resolve in the future work. Content extraction technique some limitations not perform well on some hyperlinks and video clips or advertisements. In Future, to address this problem, develop the any other hybrid algorithms and extract the main information from web pages.

CHAPTER 6
REFERENCES

1. Anna Saro vijendran, C Deepa, "LBDA: A Noval Framework for Extracting Content from Web Pages", 2103IEEE International Conference on Advanced Computing and Communication System, Coimbatore, India.
2. Dandan Song, Fei sun, lejian Liao, "A hybrid approach for content extraction with text density and visual importance of DOM nodes", In the proceeding of Springer knowl Inf Syst, DOI 10.1007/s10115-013-0687-x,Verlag London 2013.
3. Jiawei Han, MichelineKamber, "Data Mining: Concepts and Techniques".
4. Kushmerick N (1999), "Learning to remove internet advertisements".
5. Li Yue, Dang Shou-bin, Zheng Xiang, Ma Bin-Hua, "Improving Navigation Page detection by Using DOM-Based Block Text Identification," In the proceeding of IEEE 2012 Tenth International Conference on ICT and Knowledge Engineering, 978-1-4673-2317-8/12.
6. Michal Marek 1, Pavel Pecina 1, Miroslav Spousta, "Web Page Cleaning with Conditional Random Fields," Cahiers du cental, 5(2007), 1.
7. Mrs.Bharati, M.Ramageri, "Data Mining techniques and Applications, "BharatiM.Ramageri/Indian Journal of Computer Science and Engineering", Vol. 1 No.4 301-305.
8. NeetuNarwal, "Improving Web Data Extracting By Noise Removal," IET.
9. Satish J.Pusdekar,Shaikh.Phiroj Chhaware , "Using Visual Clues Concept for Extracting Main Data from Deep Web Pages", In the proceeding of IEEE 2014 International Conference on Electronic Systems, Signal processing and Computing technologies, DOI 10.1109/ICESC.2014.39.
10. S.Balan, "A Study of Various Techniques of Web Content Mining Research Issues and Tool," International Journal of Innovative Research & Studies. Vol 2 Issue 5, May, 2005.
11. Shobhit Srivastava, Mohd.Haroon, Abhishek Bajaj, "Web Document Information Extraction using Class Attribute Approach", In the proceeding of IEEE 4th International Conference on Computer and Communication Technology 2013.
12. Shuang Lin, Jie Chen, ZhendongNiu (2012) "Combining a Segmentation-Like Approach and Density-Based Approach in Content Extraction", TSINGHUA SCIENCE AND TECHNOLOGY, ISSN 1007-0214 05/18 pp256-264, Volume 17, Number 3, June 2012.

13. Shumeet Baluja,"Browsing on Small Screens: Recasting Web-Page Segmentation into an Efficient Machine Learning Framework," WWW 2006, May 23-26, 2006, Edinburgh, Scotland. ACM 1-59593-323-9/06/0005.

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

15. Suresh Subramanian, sivaprakasam, "Efficient Algorithm for Duplicate Documents," International Journal of Soft Computing and Engineering (IJSCE), Volume -3, Issue-6, January 2014.

16. Tim Weninger,William h, Jiawei Han, "CERT-Content Extraction via Tag Ratios" ,In proceeding of ACM 978-1-60558-799-8/10/04, WWW 2010,North Carolina,USA.

17. Wu Qi, Xing-shu,Zhu Kai,WANG Chun, "Relevance-based content extraction of HTML documents", In proceeding of Springer j.Cent.South Univ,DOI:10.1007/s11771-012-1226-8,Verlag Berlin Heidelberg 2012.

18. Xuhong Zhang,Yanqing Zhang , Jing He, Frank Cobia, "Vision- Based Web Page Segmentation and Informative Block Detection", 2013 IEEE/WIC/ACM International Conferences on Web Intelligence(WI) and Intelligent Agent Technology(IAT), DOI 10.1109/WI-IAT.2013.194.

APPENDIX

GUI	Graphical User Interface
WWW	World Wide Web
DOM	Document Object Model
HTML	Hyper Text Mark-up Language
XML	Extensible Mark-up language
XHTML	Extensible Hyper Text Mark-up Language
TD	Text Density
CTD	Composite Text Density
HTD	Hybrid Text Density
VI	Visual Importance
OL	Out links
PS	Page Size
ATD	Anchor Text Density
VIPS	Vision Based Page Segmentation
IE	Inline Elements
BLE	Block Level Elements
IE	Information Extraction
PDA	Personal Digital Assistance
GAHWM	Genetic Algorithm for HTML Web Content Mining
CETD	Content Extraction Text Density
CECTD-S	Content Extraction Composite Text Density Smoothing

CETDPM	Content Extraction Text Density Pattern Matching
ECECTD-S	Extended Content Extraction Composite Text Density Smoothing