

**To evaluate and enhance DOM Tree algorithm for noisy  
Data detection in web content mining**

A Dissertation submitted

By

Sachit Mahajan

(11301681)

To

**Department of Computer Science & Engineering**

In partial fulfillment of the Requirement for the

Award of the Degree of


**Master of Technology in Computer Science**

Under the guidance of

**Ms. Maneet kaur**

(May 2015)

# Approval page

 **LOVELY PROFESSIONAL UNIVERSITY**  
Empowering Education. Transforming India.

School of: Computer Science & Technology

---

**DISSERTATION TOPIC APPROVAL PERFORMANCE**

Name of the Student: <u>Sachit Mahajan</u>	Registration No.: <u>11301681</u>
Batch: <u>2013-15</u>	Roll No.: <u>RK2307032</u>
Version: <u>August-2014 (Sem-3)</u>	Parent Section: <u>K2307</u>
Details of Supervisor:	Designation: <u>D.P</u>
Name: <u>Maneet Kaur</u>	Qualification: <u>M-T</u>
Id: <u>15709</u>	Research Experience: <u>3yrs</u>

---

SPCIALIZATION AREA: Database Mining [pick from list of provided specialization areas by DAA]

PROPOSED TOPICS:

- Web mining for removal of data redundancy by using Dom tree, and its enhancement.
- Web mining for noise reduction by using concept of Dom tree and its enhancement
- currency & classification

[Signature]

FAC Remarks:

first topic approved

1st/11/17

APPROVAL OF PAC CHAIRPERSON: [Signature] Date: 11/11/17

\*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 FAC approval will be retained by the student and must be attached in the Project/Dissertation final report.  
\*Last copy to be submitted to Supervisor.

## **ABSTRACT**

Internet is continuously developing and improving and become prime source of knowledge and information such that amount of data present will keep on increasing day by day. Information can be of different type like text audio video .The web pages that consist of mixture of information containing central part which user intent to seek, copyright part, navigational panel, advertisement part etc. For end user only part of information which is desired is important rest all is considered as noise which degrades the information and can harm the web mining. In our work we will focus on detecting and removal of redundant noise data to from the web page and improve the performance of web mining. Here basic focus is on local noisy data and for the detection of the noisy data the concept of structure tree is used called Dom (document object model) tree. It bifurcate the page in the form of nodes and from where the content with information and noise can be detected.

## **ACKNOWLEDGMENT**

It is a pleasure of mine to find myself penning down these lines to express my sincere thanks to all my coordinators to give me this opportunity of preparing this project, to enhance my professional as well as my technical practice. I express my deep sense of gratitude to my DISSERTATION MENTOR Ms.Maneet Kaur to give me knowledge about the topic and the concept related to this research. Without this guidance I could not even imagine to complete my dissertation on time.

My deepest gratitude is to all my teachers for always boosting my moral and providing me encouraging environment. In the last, I would like to thank my parents, without whom nothing was possible.

Sachit Mahajan

## **DECLARATION**

I hereby declare that the dissertation proposal entitled, **To enhance and evaluate Dom tree algorithm for noisy data detection in web content mining** submitted for the M.Tech 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-**

01-may-2015

**Investigator-** Sachit Mahajan

**Reg No-** 11301681

## **CERTIFICATE**

This is to certify that **SACHIT MAHAJAN** has completed M.tech dissertation proposal titled **To enhance and evaluate Dom tree algorithm for noisy data detection in web content mining**, 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 other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.tech Computer Science & Engg.

**Date:**

01-may-2015

**Signature of Advisor**

**Name:** Maneet Kaur

**UID:** 15709

# TABLE OF CONTENT

<b>Abstract.....</b>	<b>i</b>
<b>Acknowledgement.....</b>	<b>ii</b>
<b>Declaration.....</b>	<b>iii</b>
<b>Certificate.....</b>	<b>iv</b>
<b>Table of Contents.....</b>	<b>v</b>
<b>List of Figures .....</b>	<b>vi</b>
<b>Chapter 1 Introduction.....</b>	<b>1-15</b>
1.1 Web mining.....	5-7
1.2 Web page segmentation.....	7-8
1.3 Informative content extraction .....	8-9
1.4 HTML based segmentation .....	9-15
<b>Chapter 2 Review of Literature.....</b>	<b>15-18</b>
<b>Chapter 3 Present work .....</b>	<b>19-21</b>
3.1 Problem formulation.....	19
3.2 Objective of research.....	20
3.3 Research Methodology.....	20-21
<b>Chapter 4 Results and Discussion .....</b>	<b>22-35</b>
4.1 Introduction to MATLAB.....	22
4.2 Introduction to the Enhanced Technique.....	23
4.3 Solution Implementation.....	27-35
<b>Chapter 5 Conclusion and Future Scope.....</b>	<b>36</b>

<b>Chapter 6</b>	<b>References .....</b>	<b>39</b>
<b>Chapter 7</b>	<b>Appendix.....</b>	<b>40</b>



## LIST OF FIGURES

Figure 1 Knowledge Discovery Process.....	2
Figure 2 An HTML document node tree.....	11
Figure 3 Visual Method.....	12
Figure 4 Interface of MATLAB.....	23
Figure 5 Proposed flow Diagram.....	25
Figure 6 The Original web page.....	27
Figure 7 Block wise segmentation of the page.....	28
Figure 8 Web Page after preprocessing.....	29
Figure 9 Advertisement scheme Interface.....	30
Figure 10.Interface representing existing algorithm complexity.....	31
Figure 11 Graphical representation of existing algorithm complexity.....	32
Figure 12 Interface representing enhanced algorithm complexity.....	33
Figure 13 Graphical representation of enhanced algorithm complexity.....	34
Figure 14 Comparison between two graph.....	35

# CHAPTER 1

## INTRODUCTION

---

Data mining is takes on a vital part in piles of the field such as market-basket analysis, classification, etc. In data mining, frequent item sets have substantial role that is used for finding out the co-relations among the arena of database. Data mining is also known as Knowledge Discovery in Database (KDD). Association rule is grounded on discovering patronize item sets. Association rules are often utilized by retail stores to supervise in inventory control, forecasting, marketing, advertising, flaws in telecommunication network. Data Mining, also popularly acknowledged as Knowledge Discovery in Databases relate to the non-trivial extraction of inexplicit, former unknown and likely useful information from data within databases. While data mining and knowledge discovery in databases (or KDD) are patronize and treated as equivalent word, data mining is in reality a component part of the knowledge discovery procedure (Gopal Pandey, 2013).

Data Mining is define as extracting the information from the large set of data. It can also define as data mining is mining the information from data. In the field of Information technology, it has enormous amount of data variable that require being bitter into useful information. This information further can be used for various applications like market analysis, customer retention, production control, fraud detection, science exploration etc (Ashish Jain, 2009).

It can be applicable to any kind of data repository. There is different kind of algorithms and techniques are available for different types of data. Data mining is studied for different databases like object-relational databases, relational database, data ware houses and multimedia databases etc.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning:** It is a stage in which noise data and irrelevant data are removed from the collection. It is also identified as data cleansing
- **Data integration:** At data integration stage, multiple data sources, as well as heterogeneous, may be combined in a common source.
- **Data selection:** In this step, the data applicable to the analysis is decided on and retrieved from the data collection.
- **Data transformation:** It is also known as data consolidation. In this phase in which the selected data is changed into forms appropriate for the mining procedure.
- **Data mining:** Data Mining is the crucial step in which clever techniques are used to extract patterns potentially useful.
- **Pattern evaluation:** In this step, severely interesting patterns representing acquaintance are recognized based on given measures.
- **Knowledge representation:** It is the final phase in which the exposed knowledge is visually represented to the user. This essential step in which they use visualization techniques to help users understand and take the data mining results.

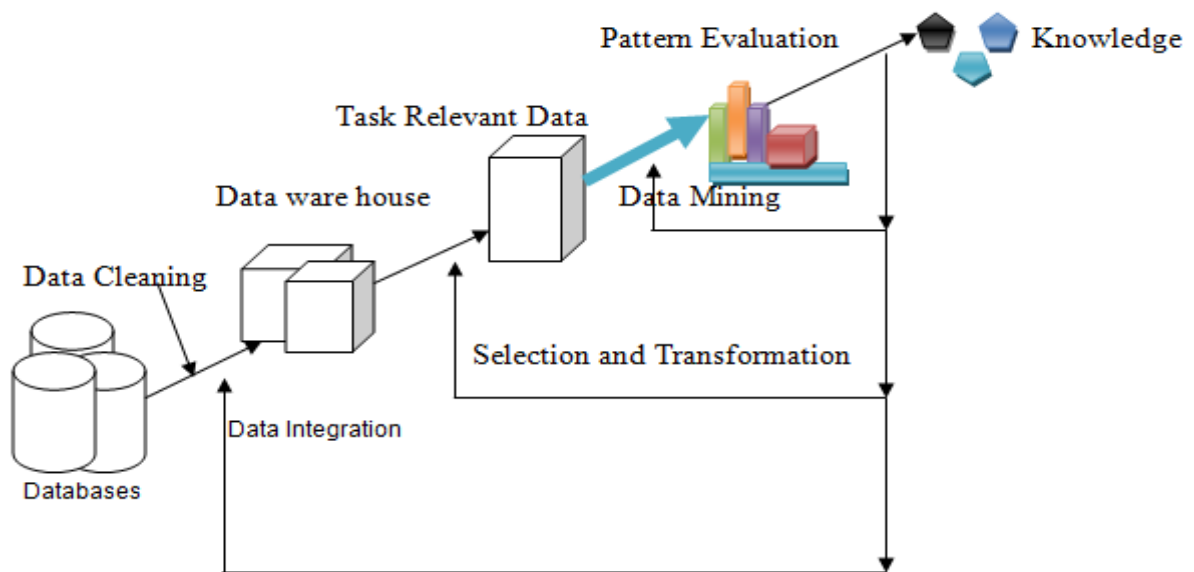


Figure 1: Knowledge discovery process

**Data Mining System:** The data mining systems are classified into various classifications according to their criteria. These classifications are as follow (Manne, 2011):

**1. Classification according to the type of data source mined:**

In this type of classification data mining system is categorized according to the data to be handled such as time series data, multimedia data, spatial data etc.

**2. Classification according to the data model:**

In this type of classification data mining system is categorized according to the data to be handled such as relational database, object oriented databases, relational database etc.

**3. Classification according to the king of knowledge discovered:**

In this type of classification data mining system is categorized according to the functionalities of the system and knowledge discovered like association, characterization, clustering, classification etc.

**4. Classification according to the mining techniques to be used:**

In this type of classification data mining system is categorized according to the data analysis approach to be used in the system. These approaches are like machine learning approach, neural network, genetic algorithm and visualization and data ware house oriented. It also take consider of user coverage and degree of user interaction.

**Issues in Data Mining:** Data mining algorithms symbolize techniques that have every so often existed for many years, but have only recently been practically as reliable and scalable tools that time and again break older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed.

Some of these issues are addressed below. Note that these issues are not exclusive and are not ordered in any way (Ford Lumban, 2009).

**Security and social issues:** Security is an important issue with any data collection that is shared and proposed to be used for deliberate decision-making. Moreover when data is composed for customer profiling, user behavior's understanding, correlating personal data with other information, etc. A large amount of responsive and confidential information about persons or companies is gathered and stored. This becomes divisive given the confidential nature of some of this data and the potential illegal access to the information.

**Performance Issues:** It refers to the following issues:

- **Efficiency and scalability of data mining algorithms:** In order to efficiently remove the information from huge amount of data in databases, data mining algorithm must be efficient and scalable
- **Parallel, distributed, and incremental mining algorithms:** The factors such as huge size of databases, wide distribution of data and complexity of data mining method inspire the growth of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed parallelism. Then the result from the partitions is combined. The incremental algorithms, updates databases without having mine the data again from scratch (K. Rajkumar, 2009).

The contents of Web pages are the primary focus of Web mining applications, such as the use of search engines and topic detection. Unfortunately, Web pages such as surrounded by many noises. According to Gibson et al., about 40%-50% of the data on the Web are noises. In addition to noises, the heterogeneity of pages and demands for automation and efficiency make it difficult to extract contents from pages. If Web pages were written according to a common template, we could easily extract content simply by writing a regular expression

However, this quickly becomes impractical when dealing with hundreds of Web pages that are generated from different templates. One way to complete this task is to segment Web pages into blocks and then detect the contents from the block set. In this way, Web pages are represented in HTML Document Object Model (DOM). According to HTML specifications and programming practices, in a node tree, related data will always be gathered as brothers and unrelated data will be separated. In other words, contents are likely to be placed in the same branch and the same layer of the node tree, while noises are likely to be separated. This makes it possible to extract content by Web page segmentation. Once pages have been segmented into blocks, the extraction can be done in each block, and the precision of extraction will be increased. Another way to extract contents is based on the fact that contents always contain large numbers of characters and need relatively few characters to depict them. Tags, while noises always contain only a few characters but need relatively large numbers of characters to depict their tags. The approaches based on this observation are known as density-based approaches. They are always easy to implement, and they are quite efficient as well. However, they have trouble managing pages with short contents and long noises.

Due to the high growth in the field of internet technology, everyone around seeks the information from internet. But it is not necessary that information we seek is valid one rather than the useful information on the web among websites, it consists of usual information such as copyright notices, header, navigation panels, banner ads, etc. These information are useful for human viewer and important for the Web site owners, but they can harm the task of information collection and Web data mining, e.g. Clustering of web page, classification of web page and retrieval of information what so is to be done to extract the valid/informative content from the web page. Web pages contain Div, Table or other HTML block which basically consist of the major content part among the sites so, method to extract the informative part among the sites is applied on these part basically

## **1.1 Web mining**

Web mining is the application of data mining techniques to extract knowledge from Web data including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

### **1.1.1 Web Content Mining**

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may contain of audio, text, images, video, or structured records such as lists and tables. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision the application of these techniques to Web content mining has not been very rapid.

### **1.1.2 Web Structure Mining**

The structure of a typical Web graph consists of Web pages as nodes and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

**Hyperlinks:** A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Webpage. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that

connects two different pages is called an Inter-Document Hyperlink. There has been a significant body of work on hyperlink analysis, of which provides an up-to-date survey.

**Document Structure:** In addition the content within a Web page can also be organized in a tree-structured format, based on the various XML and HTML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

### **1.1.3 Web Usage Mining**

Web Usage Mining is the application of data mining method to discover interesting usage patterns from Web data in order to understand and better serve the requires of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered.

## **1.2 Web Page Segmentation**

Several approaches have been explored to segment a web page into blocks or regions. In the DOM-based Segmentation scheme an HTML document is showed as a DOM tree. Useful tags that may represent a block in a page include UL (for list), P (for paragraph), TABLE (for table), H1~H6 (for heading), etc. DOM in general offers a useful structure for a web page. But tags such as TABLE and P are used not only for content organization but provides layout presentation also (Chaw Su Win, 2013).

In many cases DOM tends to reveal presentation structure other than content structure and is often not accurate enough to discriminate different semantic blocks in a web page. Another way of page segmentation is based on the layout of web page. In this a web page is generally separated into 5 regions: right, top, down, left and center.



The drawback of this approach is that such a kind of layout template cannot be fit into all the web pages. Furthermore the segmentation is too rough to exhibit semantic coherence. Compared with the above segmentation Vision-based Page Segmentation (VIPS) excels in both an appropriate partition granularity and coherent semantic aggregation. By analysis useful visual cues based on DOM structure a tree-like vision-based content structure of a web page is obtained.

The granularity is controlled by the Degree of Coherence (DoC) which represents how coherence each block is. VIPS (Vision-based Page Segmentation) can efficiently keep related content together while separating semantically different blocks from each other. Visual cues such as size, font and color are used to detect blocks. Each block in VIPS (Vision-based Page Segmentation) is represented as a node in a tree. The root is the whole page, inner nodes are the top level coarser blocks, children nodes are obtained by partitioning the parent node into finer blocks and all leaf nodes contains of a flat segmentation of a web page with an appropriate coherent degree. The stopping of the VIPS (Vision-based Page Segmentation) algorithm is controlled by the Permitted DoC (PDoC), which plays a role as a threshold to indicate the finest granularity that we are satisfied. The segmentation only stops when the DoCs of all blocks are not smaller than the PDoC.

### **1.3 Informative Content Extraction**

Informative Content Extraction is the process of finding the parts of a web page which contain the main textual content of this document. A human user nearly naturally performs some kind of Informative Content Extraction when reading a web page by ignoring the parts with additional non-informative contents, such as navigation, functional and design elements or commercial banners at least as long as they are not of interest. Though it is a relatively intuitive task for a human user, it turns out to be difficult to determine the main content of a document in an automatic way.

Several techniques deal with the problem under very different circumstances. For example Informative Content Extraction is used extensively in applications, rewriting web pages for presentation on small screen devices or access via screen readers for visually impaired users.

Some applications in the fields of Information Extraction, Information Retrieval, Web Mining and Text Summarization use Informative Content Extraction to preprocess the raw data in order to improve accuracy. It becomes obvious that under the mentioned circumstances the extraction has to be performed by a general approach rather than a tailored solution for one particular set of HTML documents with a well-known structure T. Gottron,(2009).

## **1.4 Html Based Segmentation Methods**

These approaches are based on analyzing web page without the need for rendering it. That means selected method is either based on inspecting the HTML code directly or (more often) traversing the DOM tree and evaluating information gathered from it. They are usually much faster than visual based methods. Quality and speed of this technique is given by used heuristics. Below are described some examples of algorithms which illustrate what heuristics can be used.

### **1.4.1 Text Based method**

The text-based methods differ from the other two in that they do not at all take the tree structure of the HTML into account. They only look at the text content and analyze certain textual features like e.g. the text-density or the link-density of parts of a page. These approaches are grounded in results from quantitative linguistics which indicate that statistically text blocks with similar features are likely to belong together and can thus be fused together. The optimal similarity threshold depends on the wanted granularity and needs to be determined experimentally.

#### **Advantages**

- Fast, since the DOM does not need to be built.
- Easier to implement, since no DOM access is necessary.

- Comparative performance to other approaches.

### **Disadvantages**

- Do not work with pages built via JavaScript (or you have to serialize the DOM in that case first).
- Do not take structural and visual clues into account.
- Recursive application on sub-blocks requires (arbitrary) changes to the text-density threshold.

### **1.4.2 DOM Based method**

The last group of segmentation approach is based on general traversal of the DOM tree and finding the content with usage of various heuristics. Some of related works might not even be directly solving the segmentation by traversing the DOM tree. WISH algorithm: In Hong et al. introduce their method for traversing the DOM tree and selecting relevant content. They target so called data records which are pieces of a web page which are repeating them, but with a different content.

An example of such record can be a category or search results listing. Their algorithm is divided in several steps. In the first step, they extract content candidate nodes using BFS-based algorithm. Data records are defined as tags on the same level of DOM tree, containing repetitive children sequences and having similar parent. The parent is denoted as data region. In case no nodes on a particular level of BFS satisfy the definition, the next tree levels inspected. Output of the first stage is a list of all data regions identified on the page. Other stages only filter results the algorithm gained in the first stage. Following observations are used for filtering

#### **Heuristics:**

1. Relative to the whole page, data records (and subsequently the whole data region) have large size

2. Data Records are usually repeated more than three times on a page
3. A regular expression can be devised for description of data record. Since all data records share the same template, it will apply on all of them
4. Data Records usually consist of a small amount of HTML tags after the list of data regions is filtered, every data region has to be assigned its relevancy score. The storing function described in determines the size of area taken by data records by counting characters and images each data record contain. Elements representing free space are taken into account as well. Data region with the best score is considered to be the main content of the page. This algorithm is the best example of how can different heuristics is used for page segmentation and classification.

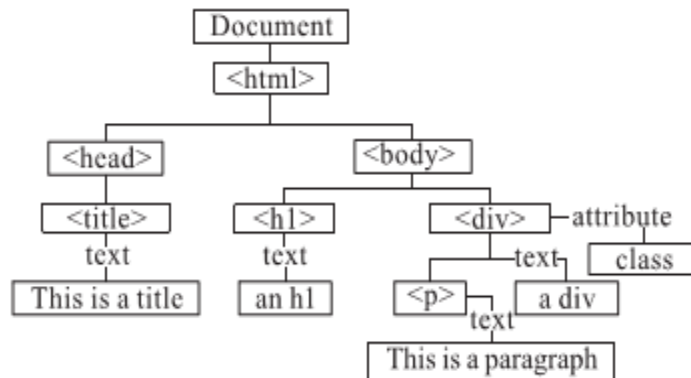


Figure2: An HTML document node tree.

### Advantages

- Easy to implement, since one only needs to parse the HTML code and not render the tree.
- Efficient to run because no browser engine is involved, thus suitable for segmenting large numbers of pages.

- Take the structural information into account.

### Disadvantages

- Based on the assumption that the HTML document reflects the semantics of the content, which is not necessarily true.
- There are many different ways to build the HTML document structure while the semantics stay the same.
- Disregard styling and layout information by design.
- Do not work with pages built via JavaScript (or you have to serialize the Dom in that case first).

### 1.4.3 Visual method

Visual approaches work the most similar to how a human segments a page, i.e. they operate on the rendered page itself as seen in a browser. They have thus the most information available but are also computationally the most expensive. They often divide the page into separators, such as lines, white-space and images, and content and build a content-structure out of this information. They can take visual features such as background color, font size and type and location on the page into account.

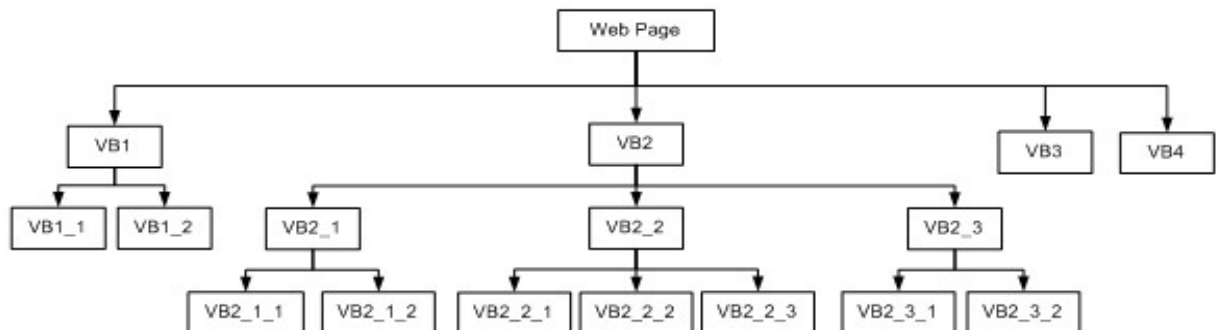


Figure3: Visual Method

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. However, Internet web pages typically contain a large amount of non-informative content such as advertisements, search and filtering panel, headers, footers, navigation links, and copyright notices, etc. Most clients and end-users search for only the informative content and the need of Informative Content Extraction from web pages becomes evident (Chaw Su Win, 2013).

To extract the informative content of the web page correctly, the informative content and the non-informative content of the web page must be known clearly. The informative content is the main content of the web page that gives some information to the user e.g. articles about technology, health or education, etc. The non-informative content of the web page contains fixed description noise such as site logos, copyright notices, privacy statements, etc. service noise, irrelevant services, such as the weather, stock or market index, etc. navigational links, advertisements, header, footer and so on. To distinguish between the informative and non-informative content in a web page, it needs to segment the web page into semantic blocks. There are several kinds of methods for Web Page Segmentation (2013).

**Least Recently Used:** The above discussed algorithm like DOM tree has some disadvantages like it has high complexity and time consuming process. To overcome this problem a new algorithm is used in page replacement that is least recently used (LRU). A good approximation to the best possible algorithm is based on the surveillance that pages that has been greatly used in the last. A small number of instructions will most likely be a lot used all over again in the next the minority. On the other hand, pages that have not been used for long time will probably remain unused for a long time. This idea suggests a practicable algorithm. When a page fault occurs, throw out the page that has been unused for the longest time. This strategy is called LRU paging (Pandey, 2009).

Although LRU is theoretically realizable, it is not cheap. To fully apply LRU, it is essential to keep a linked list of all pages in memory, with the most recently used page at the front and

the least recently used page at the rear. The complexity is that the listing must be updated on every memory reference. Finding a page in the list, deleting it, and then moving it to the front is a very time consuming operation, even in hardware also.

There are further many ways to implement LRU with special hardware. Let us consider the way first which is very simple. This method requires equipping the hardware with a 64-bit counter,  $C$ , that is automatically incremented after each instruction. In addition, each page table entry must also have a field large sufficient to include the counter. After each memory reference, the current value of  $C$  is stored in the page table entry for the page just referenced (zubi, 2002). When a page fault occurs, the operating system examines all the counters in the page table to find the lowest one. That page is the least recently used.

There is no single cache algorithm which will for all time do well because that requires perfect knowledge of the future. The dominance of LRU in VM cache design is the result of a long history of measuring system behavior. Given real workloads, LRU works appealing well a very large fraction of the time. However, it is not very hard to make a reference string for which FIFO would have superior performance over LRU.

Consider a linear sweep from side to side a large address space much larger than the available page able real memory. LRU is based some assumption like "what you have touched recently you are likely to touch again", but the linear sweep completely violates that assumption. This is reason that there are some operating systems allow programs to recommend the kernel about their reference behavior. Consider its one example is "mark and sweep" garbage collection typified by classic LISP interpreters (Ford, 2010).

Another example is the symbol table in an assured antique macro processor (STAGE2). The binary tree is searched from the root for each symbol, and the string evaluation is being done on a stack. It turned out that plummeting the available page frames by "wiring down" the root page of the symbol tree and the bottom page of the stack made a huge improvement in the page fault rate. The cache was small, and it churned fiercely, always approaching out the two most regularly referenced pages because the cache was smaller than the inter-reference distance to those pages (V. Prasad, 2003). So a little cache worked better, but ONLY because those two page frames stolen from the cache was used wisely.

The overall result of LRU is the standard because it is typically pretty high-quality for real workloads on systems that are not hideously overloaded and that is supported by years of careful measurements. On the other hand, we can surely find cases where different performance will be superior. This is the reason for importance of measuring (Elizebth, 1993).



## CHAPTER 2

# REVIEW OF LITERATURE

---

**K.Rajkumar (2012):** The paper comprises of a method to segmentation the web page by using Dynamic web page segmentation (DWS) technique. Under this, the segments of web page document are done on the bases of layout based segment and reappearance based technique. With the analysis of reappearance tag from Document object model (Dom) tree structure segmentation in the web page document is done and the segmentation is called reappearance base page segmentation. Determining and reviewing tag, it give segment web-pages and the layout of tag is like <DIV>, <TABLE>, and <FRAME> tags. In dynamic web page segmentation, if the segmentation consists of reappearance tag it means that segmentation is based on reappearance. Else it will segment based on web layout based, under which segmentation is based upon the region like header, footer navigation, context etc. The above mention technique also works with the mobile view of the web pages after the segmentation blocks with valid hyperlink will be displayed on the mobile and from that user select hyperlinks based on his preference and requirement

**Xuhong Zhang Yanqing Zhang (2013):** This paper proposed a new technique whose purpose is to extract the valid content from the web documents. This is done to get the useful data out of unwanted one. The technique is based on visual based page segmentation to detect the informative context with in a low budget cost. It introduces the row column splitting which provide a valid segmentation result as compare to the traditional clustering technique. Firstly the proposed system use splitting to bifurcate the given set of data then HTML Parser whose purpose is to extract the valid content, is used to build DOM (Document Object Model) tree from which the valid / informative content out of the main content block can b easily extracted. The proposed system can define the ranking of the documents and also extracts the top ranked documents as per the need of the query of user because Web page typically contained many information blocks along with the irrelevant

one. Apart from the main content blocks the other non informative block which are called noisy blocks which can seriously harm Web data mining.

**Jinbeom Kang (2010):** The paper comprises of new tech technique of Web page segmentation in which proposed by identifying repetitive tag and the segmentation based on it is determined as repetition based page segmentation, the pattern with the repetitive tag is called key patterns .This key pattern in Document Object structure of a page is used. The identification is based on Repetition-based Page Segmentation (REPS) algorithm which detects key patterns in a page. Lots of experiments is performed on Web sites show that Repetition-based Page Segmentation (REPS) play a vital role in contribution to improve the correctness of Web page document segmentation. This ultimately affects the performance of data extraction by granting access to the system to identify and recognize the informative blocks among the whole context of Web page.

**Libing Wu, Yalin Ke (2011):** In this paper they proposed block gathering based page segmentation algorithm which focus on mobile web page segmentation. As the page is divided into tags, these tags are gathered in this and called as gather node. On the basis of this gathered node, block gathering base segmentation is implies. This has great effect on the mobile web page segmentation. As there is rapid advancement in the technology, so mobile phone and other portable device play huge and important role in accessing the valid content from the internet and so as to get the valid data in such portable devices new segmentation technique is proposed and implies, Because traditional method cannot effectively manage pages that consist of short contents and long noises.

**Midhun Mathew (2013):** In this paper feature extraction technique is introduced along with the Dom tree. This is done to get the informative part out of the web and to eliminate noisy information. Thus clean web page is obtained which is playing its high role in improving web mining results. This paper particularly focuses on elimination the local noise and gets the

valuable data. Feature DOM tree is introduced. The technique follow 3 stages, in the 1<sup>st</sup> phase feature selection is done, that is feature DOM tree is formed. In the 2<sup>nd</sup> phase noise is marked and in the last phase the noisy data is removed from the content and valuable data remains which can be used to draw conclusion, gain knowledge and so on. Feature DOM-based method and reduce the drawbacks of previous works in Web Page Segmentation.

**Li liu, Junfang shi (2010):** This paper provides an overview of the extraction of informative part from the web on the basis of DOM tree and ontology. Information present over web is dynamic, irregular and huge. Due to large amount of information present on web, it is difficult to search the informative part from web. Based on the two techniques, the interested informative part is extracted. Ontology generates the rules for information extraction. Dom tree is used to extract the informative part as well as noisy area can also be detected. The result of the effectiveness can be determines based on the precision and recall value of the information that is extracted from the web page.

**Weicheng Ma, Xiuxia Chen (2012):** Information present over web is large, irregular and typical in nature, so as to access information special web agent / web spiders / web crawlers are there which provide information from the requested site / server to the host. To meet the effective needs of the today's web environment advancement in the web crawler is done with the use of Dom.

**Hengyu Lai, Yali Wang (2014):** With the rapid development of internet over a short period of time. Information becomes vast, rich and complex, so to extract information from internet is worth investing. Two methods are proposed for such, 1<sup>st</sup> method is width priority analysis method and 2<sup>nd</sup> method is depth priority analysis method which is based on Dom tree and applied to Html document. These methods provide valid information from the vast world of internet. These are used to fetch the useful data from the unwanted data

**Xing xie ,wei-ying ma (2005):** This paper determine that there is huge expansion of people who search the web, when they are on the move. Though conventional search engines can be directly visited from mobile devices with web browsing capabilities, the information is not as accurately accessible from a handheld device as it is from desktops. In this paper, a block importance model is employed to assign importance values to different segments of a web page, in order to extract and present more condensed search results to mobile users. Based on the block importance different levels of detail have been .A set of user study experiments have been carried out to compare commercial service on typical mobile devices.

**S.S.Bhamare, Dr.B.V.Pawar:** This paper determines web page noise cleaning is one of the new research areas of study for removing the noise patterns of web pages for effective web mining. The World Wide Web contains large amount of web pages which are accessible by users. With conventional data or text, Web pages generally contain a large amount of noise information that is not part of the main contents of the web pages, e.g., advertisement banners, navigation bars, and disclaimer/copyright notices. The main objective of this area is removing such irrelevant information (i.e. Web Page Noise or Local Noise) in Web pages that can seriously harm Web mining task such as clustering and classification etc. The main purpose of this paper is to review and discuss the major research work that has been done in this area and identifying the challenges and issues in this area.

## CHAPTER 3

# PRESENT WORK

---

### 3.1 Problem Formulation

The web pages are managed in the certain fixed manner. The users are required to extract the useful information from the web page. The user's useful information is the useful data for the users and other information is noisy one. The user extracts the useful information from the web page on the basis of web page template. Data mining on the Web thus becomes an important task for discovering useful knowledge or information from the Web. However, useful information on the Web is often accompanied by a large amount of noise such as banner advertisements, navigation bars, copyright notices, etc. Although such information items are functionally useful for human viewers and necessary for the Web site owners, they often hamper automated information gathering and Web data mining e.g. Web page clustering, classification, information retrieval and information extraction. Web noises can be grouped into two categories according to their granularities: The user extracts the data from the web page using the content page holder tag. The concept of DOM trees has been used to extract the useful information. The DOM tree technique comes under the technique of web segmentation. In the DOM-based segmentation approach, an HTML document is represented as a DOM tree. Another intuitive way of page segmentation is based on the layout of webpage. In this way, a web page is generally separated into 5 regions: top, down, left, right and center. Another application of block importance is on web page classification. The main problem in the HTML parse is that HTML parser is quite slow. The second problem in DOM trees is that it requires a lot of time to construct the DOM trees which will reduce the efficiency. In this thesis, we work on to remove the noisy data from the web page. To enhance the efficiency and to increase the response novel algorithm will be proposed.

## 3.2. Objective of Research

1. To study the various existing techniques for web page segmentation
2. To identify the disadvantage of DOM tree technique for web page segmentation
3. To propose novel technique to remove the noisy data from web page using web segmentation technique
4. To implement the proposed technique and compare the results with the existing technique of DOM trees

## 3.3. Research Methodology

A Web page typically contains many information blocks. Besides, the content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements. These blocks that are not the main content blocks of the page, we call them as noisy blocks. We show that the information contained in these noisy blocks can seriously harm Web data mining. Thus eliminating these noises is of great importance. To eliminate the noise from the web page efficient algorithm is required which is faster than DOM parser. In this work, we work on to remove the non required advertisements. This can be done with the use of least recent used algorithm. This algorithm is the page replacement algorithm in which the least recently used pages are removed from the web page. In this work, we use the concept of web usage. The web usage mining will tell us the recent used advertisement link. On the basis of this information LRU algorithm will work.

**Least Recently Used:** To fully apply LRU, it is essential to keep a linked list of all pages in memory, with the most recently used page at the front and the least recently used page at the rear. The complexity is that the listing must be updated on every memory reference. Finding a page in the list, deleting it, and then moving it to the front is a very time consuming operation, even in hardware also. The overall result of LRU is the standard because it is typically pretty high-quality for real workloads on systems that are not hideously overloaded

and that is supported by years of careful measurements. On the other hand, we can surely find cases where different performance will be superior.

LRU is very easy to implement. It is not simple as compare to FIFO but its complexity is less than optimal page replacement.

# CHAPTER-4

## RESULTS AND DISCUSSION

### 4.1 Introduction to MATLAB

MATLAB stands for matrix laboratory. It is a multi-prototype numerical computing environment and fourth generation programming language. It is used for matrix manipulation, plotting of function and data. With the help of its programming capabilities it provides tool which is very useful for all areas of science and engineering.

GUI toolbox allow advanced matlab programmer to provide graphical user interface to their program.

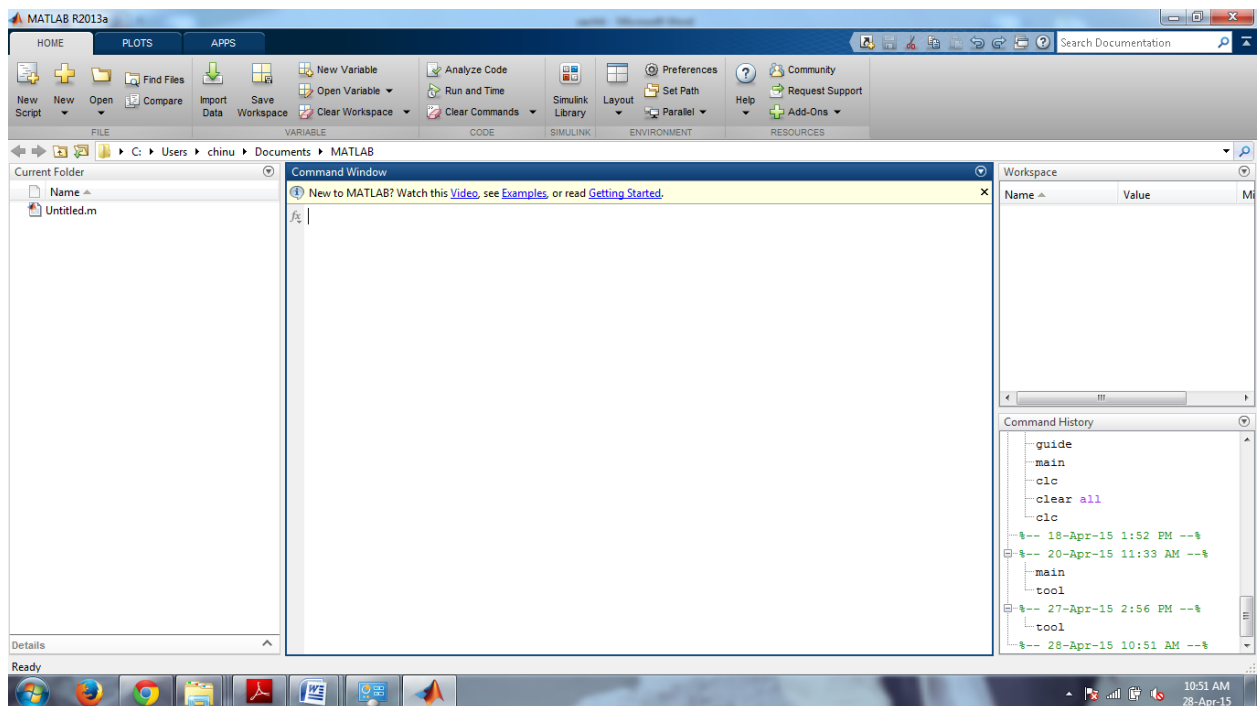


Figure 4: Interface of matlab



## **4.2 Introduction to the Enhanced Technique**

A Web page consists of blocks that contain informative data. Apart from the main content all the other part of the web page are referred as a noisy blocks or noisy data. To detect these noises among the web page is of great importance.

Steps involved:

- 1: Input the web content to the parser.
- 2: Apply Dom tree algorithm.
- 3: Since the web page is bifurcate into number of nodes so to detect the noisy content LRU algorithm is applied.
- 4: The complexity of the noise in the web content is detected.

#### 4.2.1. Proposed flow diagram:

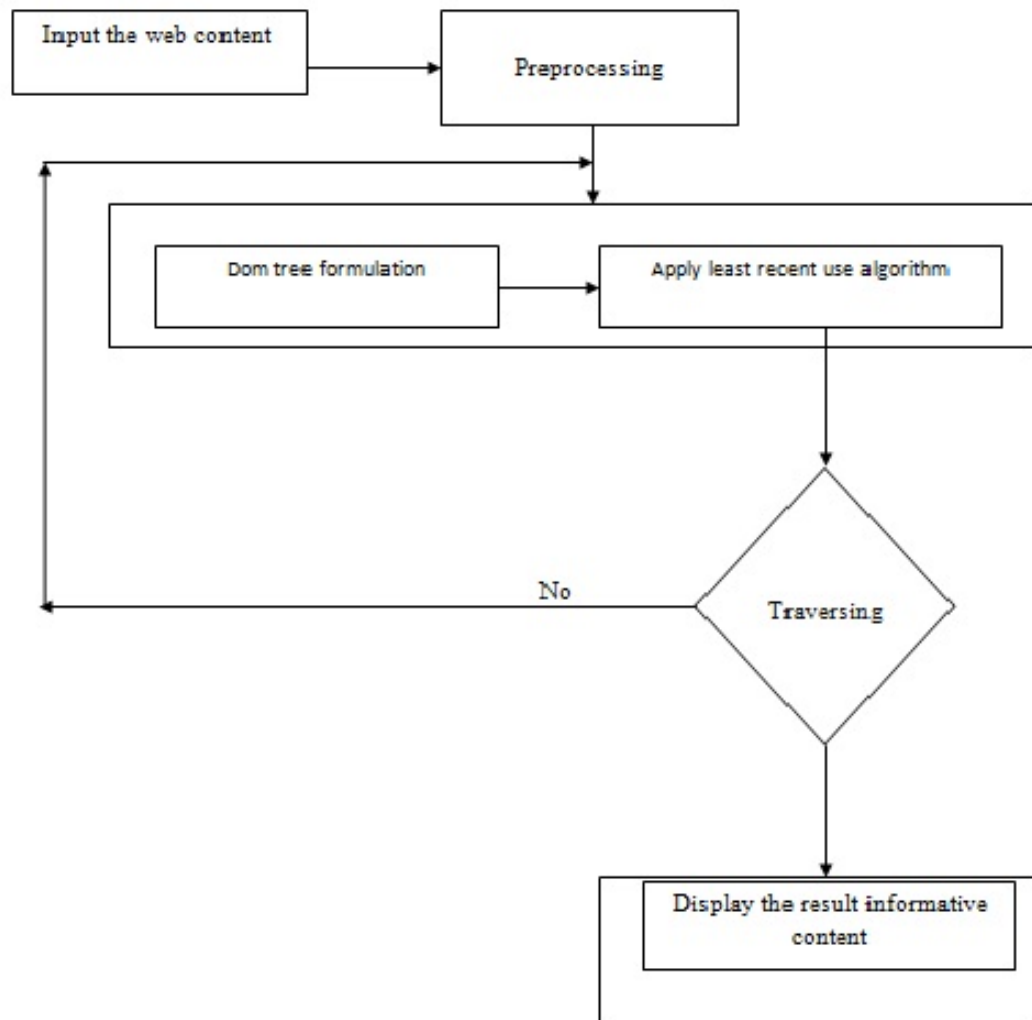


Figure 5: Proposed Flow Diagram

#### 4.2.2 Algorithm:

1. If the referenced page is in memory:
2.  $P$  = position of the referenced page in the LRU queue;
3. If  $P > W$ :
4. If  $N > 0$ : /\* If the execution state is sequential operating mode \*/
5.  $INERTIA = 0$ ; /\* Finish the sequential operating mode \*/
6. If  $P > W+1$
7.  $TC = TC + N$ ; /\* Increase the tendency confirmation (error) \*/
8.  $N = 0$ ;
9.  $W = P$ ; /\* Increase the working area size \*/
10. Move the referenced page to the head of the LRU queue.
11. Else (the referenced page is not in memory):
12. If memory is full:
13. If  $W \leq L$ :
14.  $INERTIA = INERTIA + 1$ ;
15. If  $W \leq C$ :
16.  $W = C + 1$ ;
17. If  $INERTIA \geq W+TC$ : /\* Sequential operating mode \*/
18. If  $N < M$  or  $N < 50$ :
19.  $N = N + 1$ ;
20. If  $TC > C$ :
21.  $TC = TC - 1$ ;
22. Remove page at the  $W+1$  position in the LRU queue;
23. Else ( $INERTIA < W+TC$ ): /\* Sequential tendency \*/
24. Remove page at the  $M$  position in the LRU queue;
25. Else ( $W > L$ ): /\* LRU tendency \*/
26.  $INERTIA = 0$ ;
27.  $W = 0$ ;
28.  $N = 0$ ;
29. Remove page at the  $M$  position in the LRU queue;
30. Insert the referenced page at the head of the LRU queue.

### 4.3 Solution Implementation

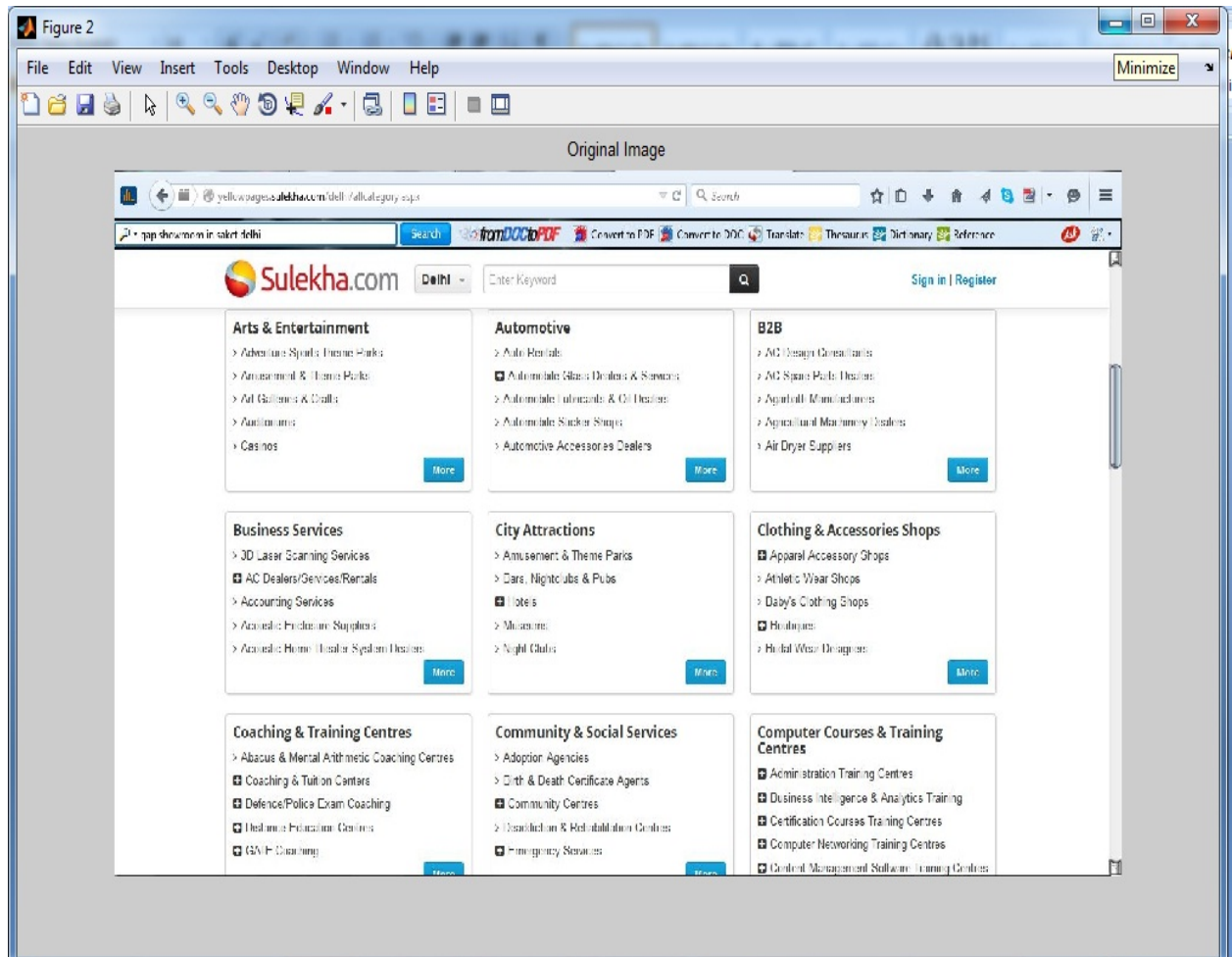


Figure 6: The original web page

Figure 6 illustrate the web page on which the action is applied to detect the noise out of it. That is irrelevant data out of the web page.

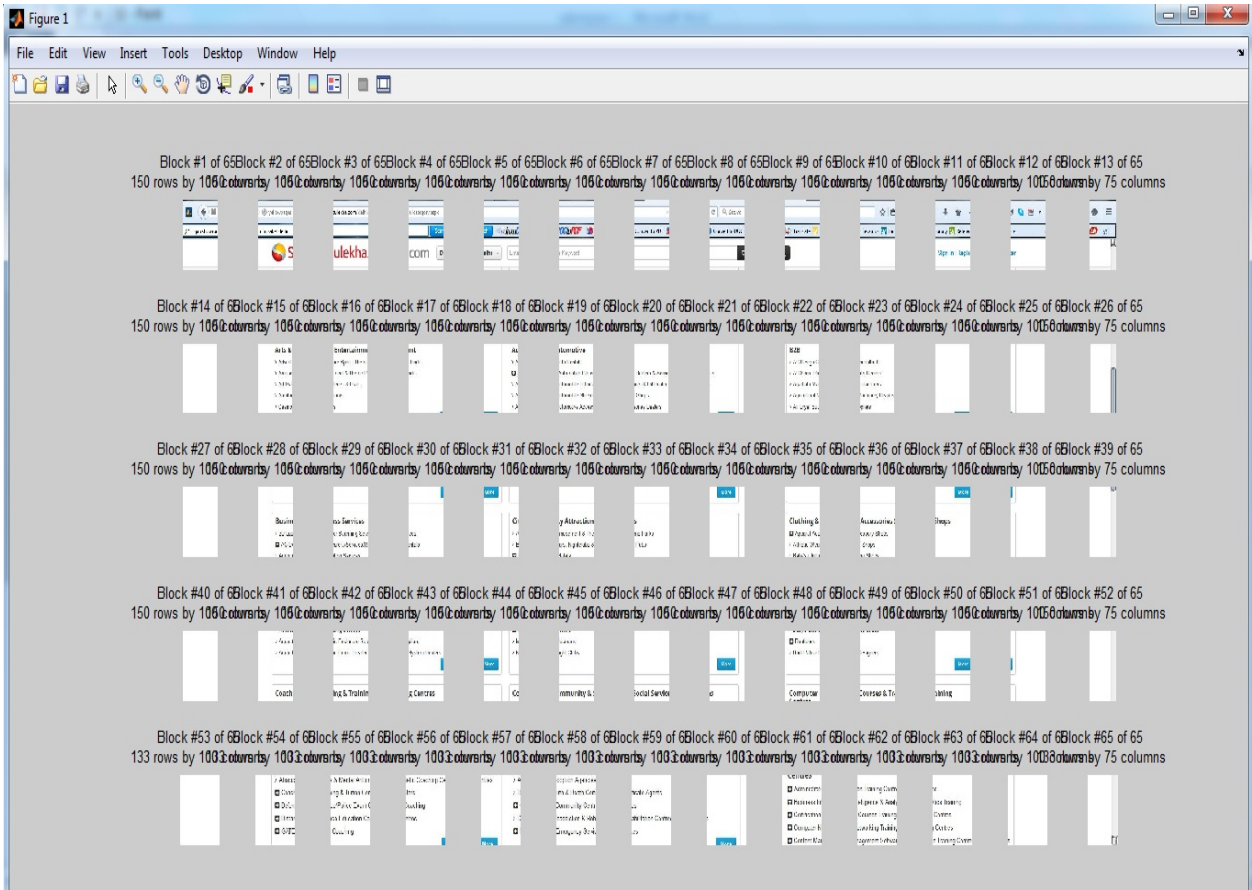


Figure 7: Block wise segmentation of the page

Figure 7 illustrate the block wise segmentation of the web page based upon certain length and breadth; this is done to remove the unwanted things such as color, blocks etc

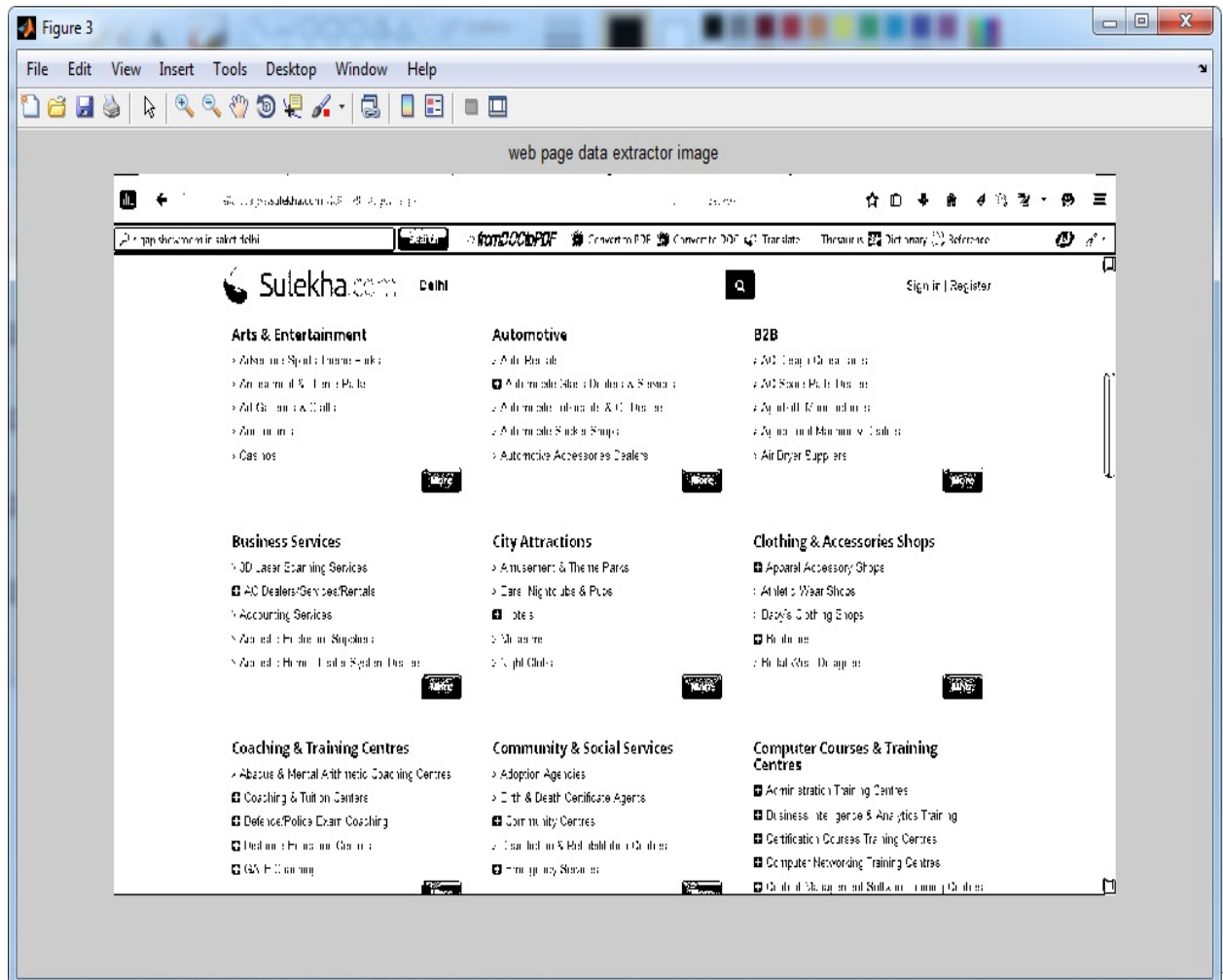


Figure 8: Web Page after preprocessing

Figure 8 illustrate the processed page after preprocessing is done, from this page noise is detected and detected noise is used by Dom tree algorithm for detecting complexity and the enhance Dom tree for detecting complexity.

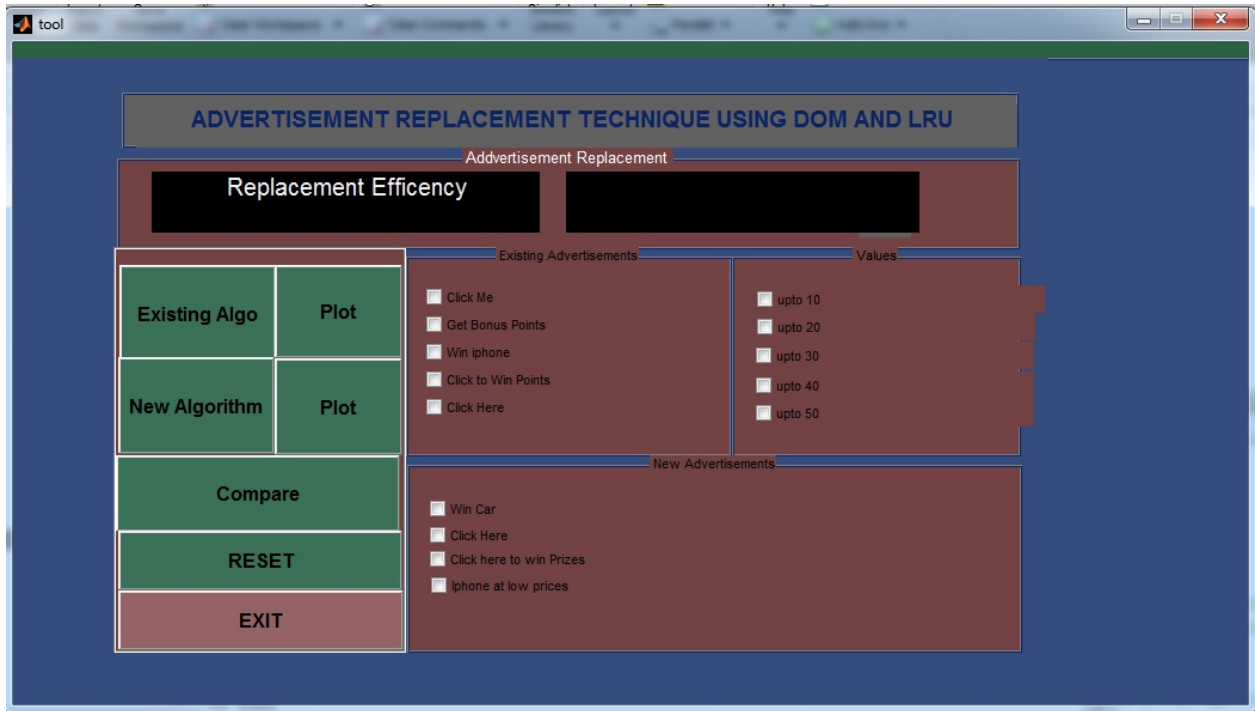


Figure 9: Advertisement Scheme Interface

As illustrated in the figure 9, the interface is been shown in which the user can select the advertisements available and which advertisement you want to replace. The two schemes are used for advertisement replacement i.e. DOM Tree and LRU

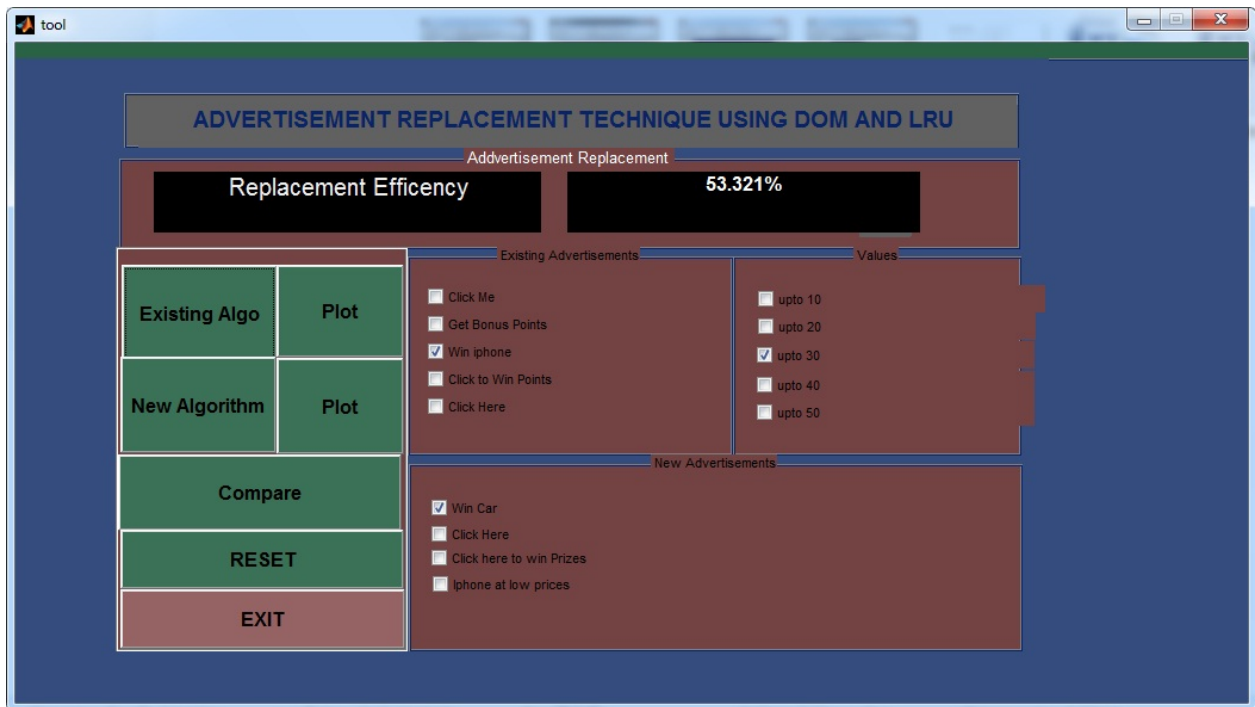


Figure 10: Interface representing existing algorithm complexity

As illustrated in the figure 10, the user selected the click me advertisement which user wants to replace with win car advertisement and with Dom tree complexity is been shown



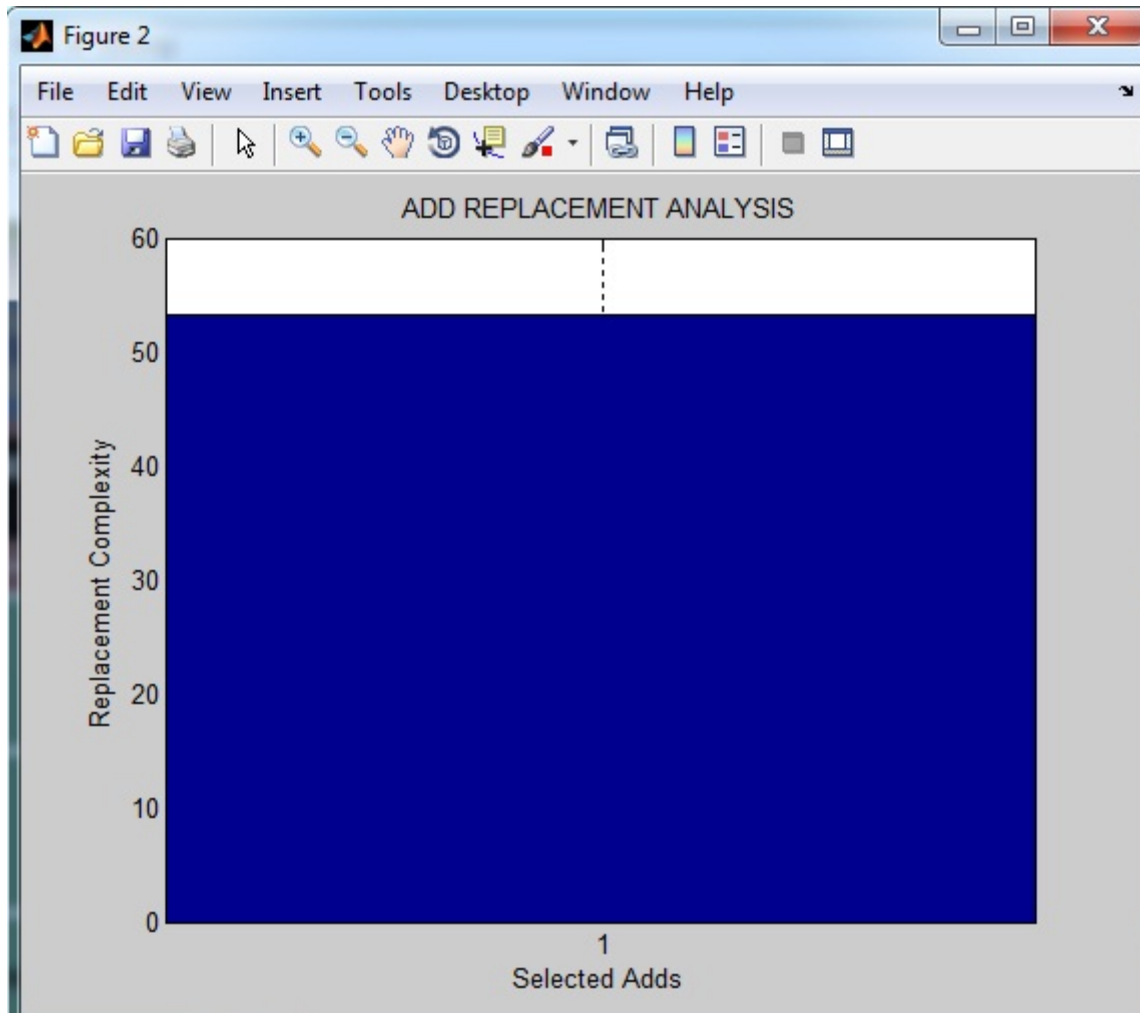


Figure 11: Graphical representation of existing algorithm complexity

As illustrated in the figure 11, the user selected the click me advertisement which user wants to replace with win car advertisement and with Dom tree complexity is been shown and graph is plotted to compare the results graphically

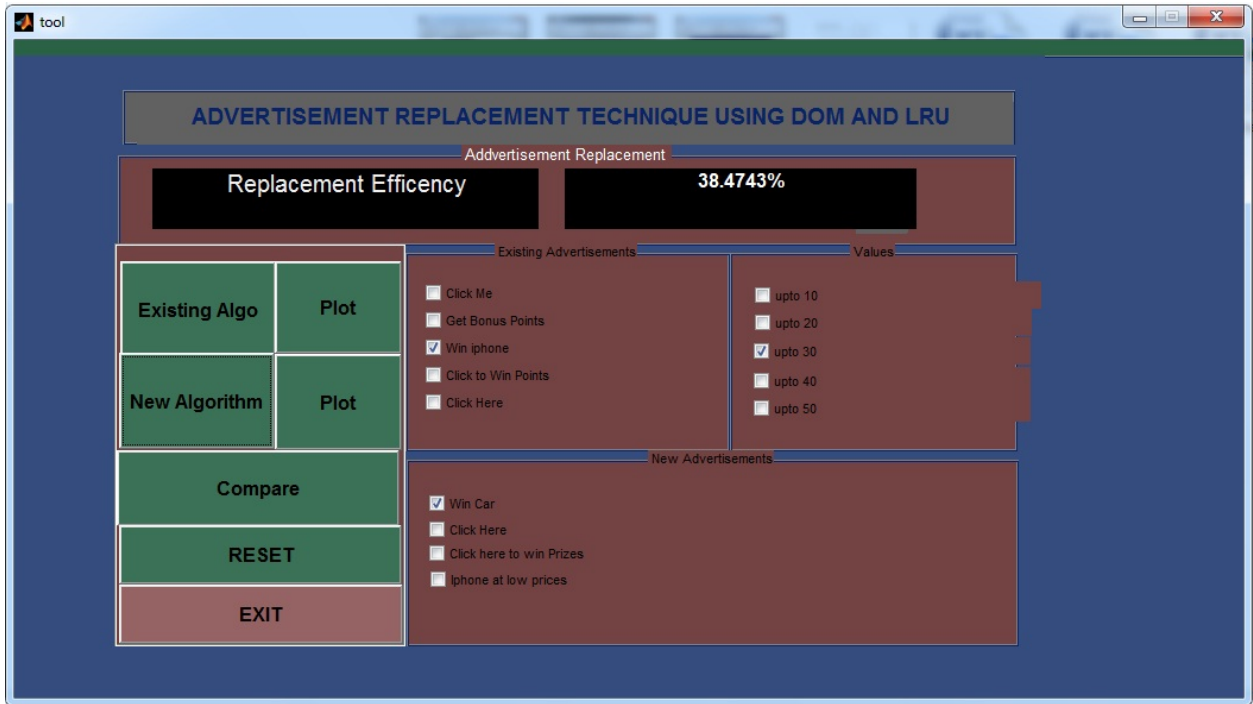


Figure 12: Interface representing enhanced algorithm complexity

As illustrated in the figure 12, the user selected the click me advertisement which user wants to replace with win car advertisement and the enhanced Dom tree complexity is been shown which is the reduced value

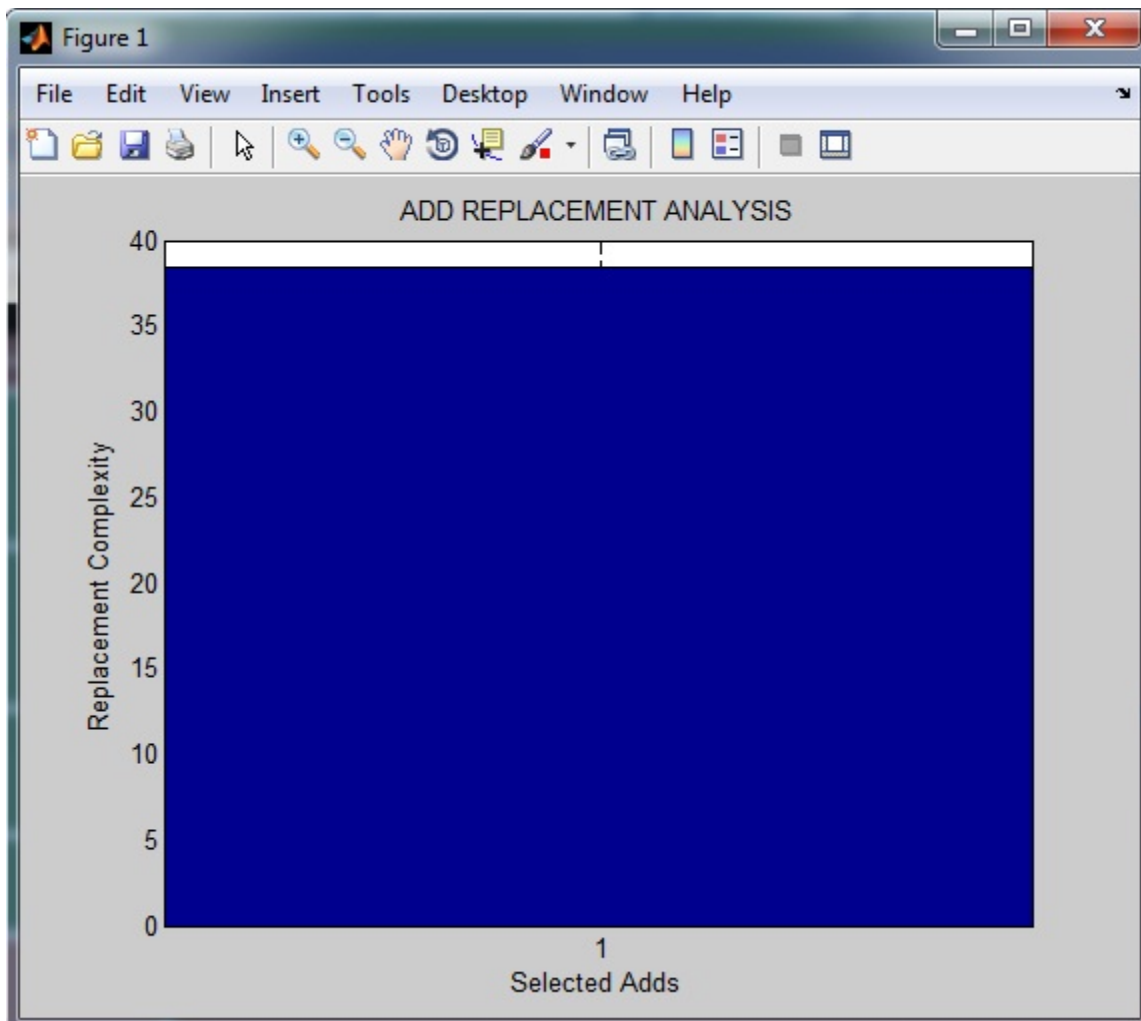


Figure 13: Graphical representation of enhanced algorithm complexity

As illustrated in the figure 13, the user selected the click me advertisement which user wants to replace with win car advertisement and with LRU complexity is been shown and graph is plotted to compare the results graphically

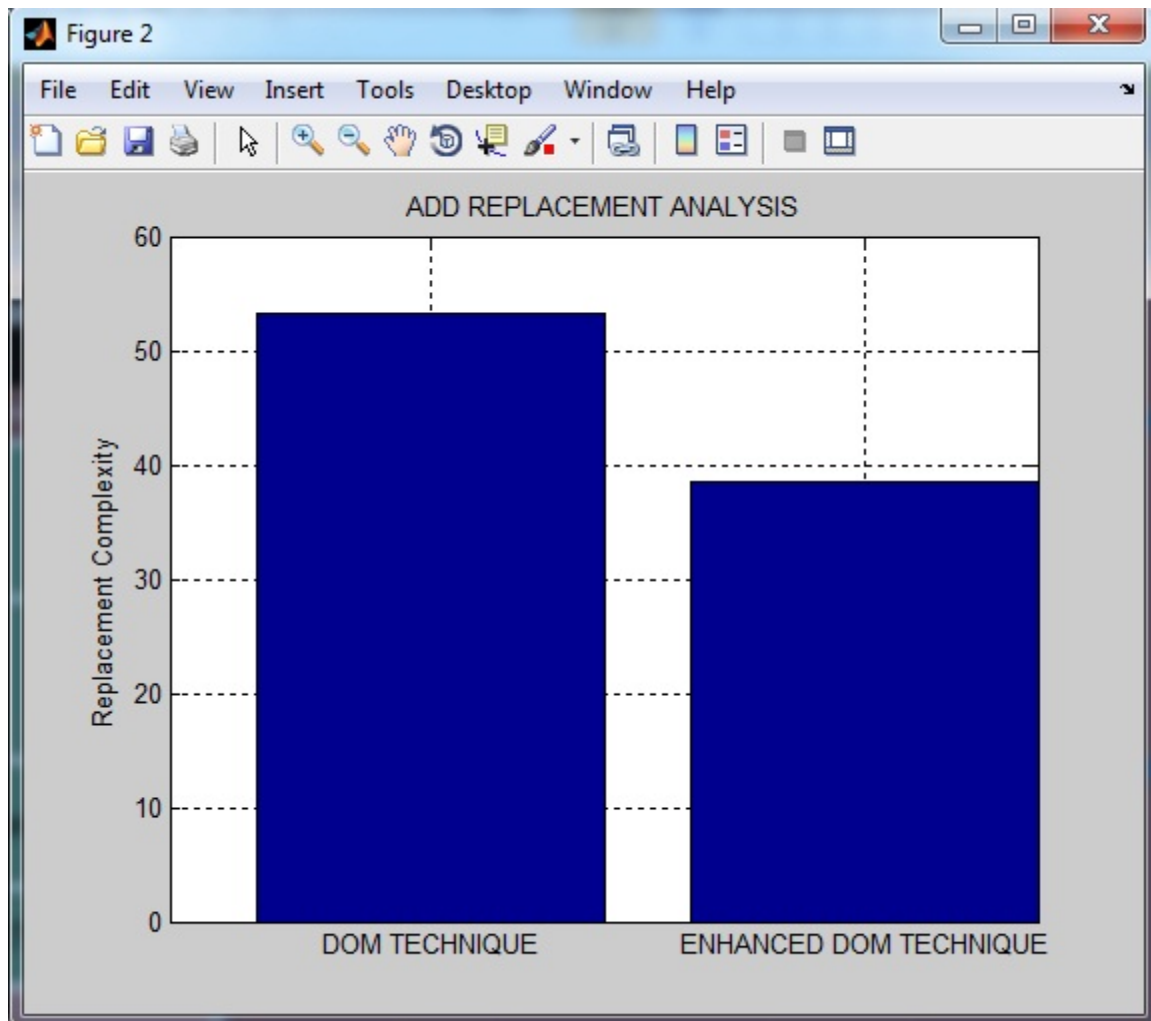


Figure 14: Comparison between two graph results

As illustrated in the figure 14, the user selected the click me advertisement which user wants to replace with win car advertisement and with DOM tree and LRU complexity is been shown and graph is plotted to compare the results graphically

# CONCLUSION AND FUTURE SCOPE

---

The webpage contains various type of data like advertisement, static, dynamic data. In the webpage some data will be categorized as the noisy data. In this work, advertisements are defined as the noisy data which are not used by the users from long period of time. In the previous time DOM tree algorithm has been implemented to detect noisy data from the WebPages. In this work, enhancement will be proposed in DOM tree algorithm to improve performance of algorithm in term of noisy detection. In future work, further enhancement will be proposed in DOM tree algorithm to improve noise detection rate and reduce detection time.

## CHAPTER 6

### REFERENCES

---

Chaw Su Win, Mie Mie Su Thwin (2013)” *Informative Content Extraction By Using Eifce*” International Journal Of Scientific & Technology Research Volume 2, Issue6.

Deng Cai (2003)” *VIPS: a Vision-based Page Segmentation Algorithm*” Microsoft Research Microsoft Corporation One Microsoft Way Redmond, WA 98052

ELIZABETH J. O’NEIL AND PATRICK E. O’NEIL, “ An Optimality Proof of the LRU-K Page Replacement Algorithm” Journal of the ACM, Vol. 46, No. 1, January 1999

Ford Lumban Gaol (2010) “*Exploring The Pattern of Habits of Users Using Web Log Sequential Pattern* “Second International Conference on Advances in Computing, Control, and Telecommunication Technologies.

H Lai, Y Wang,” *The Research and Implementation of Web Information Extraction Technology Based on Multi-level Pages*” ,IEEE ISSC 2014/CIICT 2014

J Kang, J Yang & N Choi, “*Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices*”, IEEE Transactions on Consumer Electronics, Vol. 56, No. 2, 2010.

Jan Zelený (2010)“*Web Page Segmentation And Classification*” Journal of Data and Knowledge Engineering.

K Rajkumar & V.Kalaivani, “*Dynamic Web Page Segmentation Based on Detecting Reappearance and Layout of Tag Patterns for Small Screen Devices*”, IEEE 2012.

Li liu & J shi ,”*Web Information Extraction Algorithm based on Ontology and DOM Tree*” , IEEE 2010.

Manne suneetha (2011) “*Clustering of Web Search Results using Suffix Tree Algorithm and Avoidance of Repetition of same Images in Search Results using L-Point Comparison Algorithm*” *PROCEEDINGS OF ICETECT*

Miguel Dar'io Duss'an-Sarria\_(2009) “*A recommendation-based web content mining model for an university community*” international conference on Knowledge discovery and data mining

Rui Xie (2012) “*Lexicon Construction: A Topic Model Approach*” International Conference on Systems and Informatics (ICSAI )

S. Ajoudanian and M. Jazi, (2009) “*Deep Web Content Mining,*” World Academy of Science, Engineering and Technology 49.

S.Das , M. Mathew , “*Eliminating Noisy Information in Web Pages using featured DOM tree*”, Foundation of Computer Science FCS, New York, USA Volume 2– No.2, May 2012.

S.Karthikeyan (2011) “*Removing Non-informative Blocks from the Web Pages*” ICCCT

S.S Bhamare, Dr.B.V. Pawar ,” *Survey on Web Page Noise Cleaning for WebMining*” IJCSIT

Shuang Lin, Jie Chen, Zhendong Niu(2012) .“*Combining a Segmentation-Like Approach And A Density-Based Approach In Content Extraction*” TSINGHUA SCIENCE AND Technologyissn11007-02141105/1811pp256-264 Volume 17.

Sumaia Mohammed AL-Ghuribi and Saleh Alshomrani(2013) “*A Comprehensive Survey on Web Content Extraction Algorithms and Techniques*”

Swe Swe Nyein (2011) “*Mining Contents in Web Page Using Cosine Similarity*”.

V. Bharanipriy and V. Prasad, (2011)” *Web Content Mining Tools: A Comparative Study,*” International Journal of Information Technology and Knowledge Management, Volume 4, No. 1, pp. 211-215. January-June .

W Ma, X Chen, “*Advanced deep web crawler based on Dom*” , IEEE Fifth International Joint Conference on Computational Sciences and Optimization, 2012.

X.Zhang, Jing(Selena) He & F.Cobia, “*Vision-based Web Page Block Segmentation and Informative Block Detection*”,IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT), 2013.

Xing xie ,wie-ying ma ,”*Efficient Browsing of Web Search Results on Mobile Devices Based on Block Importance Model*” IEEE computer society 2005

Yalin Ke+, Y.He & Nan Liu,” *A Block Gathering Based on Mobile Web Page Segmentation Algorithm*” , IEEE Intl. Conf. on Trust,security,privacy in computing and communication, Publication Year: 2011 .

Zubi, Z.(2002).”Using Some Web Content Mining Techniques for Arabic Text Classification”.ISSN 1790-5109.pp.73-84.



## **CHAPTER 7**

## **APPENDIX**

---

### **ABBREVIATIONS**

DOM: Data Object Model

VIPS: Vision Base Page Segmentation

LRU: Least Recent Used