Noise Reduction and Content Extraction from Web Pages Using DOM Based Page Segmentation

Sandeep Kaur¹, Abhishek Tyagi²
Department of Computer Science Engineering, Lovely Professional University Jalandhar, Punjab, India¹
Email: Sandeepkaur.1489@yahoo.com
Assistant Professor, Department of Computer Science Engineering, Lovely Professional University Jalandhar, Punjab, India²
Email: Abhishek.16857@lpu.co.in

Abstract

Web Page has large amount of information and the information in web pages is useful in real world applications. Web Page has some additional contents like hyperlinks, header footer, navigational panel; advertisements may cause the content extraction to be complicated. These additional contents in Web Pages are treated as noisy contents. A method is needed to remove the noise from web pages and extract only useful information. The DOM Based page Segmentation is used to discard the noisy content and extract the informative content from Web Pages. Initially a XML Web Page is converted into DOM tree, token is generated and noise is removed using DOM Based Page Segmentation which converts the page into blocks and regions. Performance of Web Content extraction is analysed based on complexity and efficiency on proposed algorithm.

Keywords:- Web Content Mining, Content Extraction, DOM Based Page Segmentations.

1. INTRODUCTION

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 [1]. 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.

2. **LITERATURE REVIEW**

**Pusdekar J et al. (2014)** presented a Vision Based approach. The Web Pages contains the information related to Web Page layout and font based on the visual information. The web page visual information extracted through a Visual block tree generated. The main block is show the all page and every block in structure represent in a rectangular region. Data records extracted from deep Web Page and the boundary of data is discover and extract them from deep web pages. The noise is filtered blocks are grouping based on similarity images, plain text, links [2].

**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 [3].

**Xuhong Zhang. (2013)** proposed a low cost vision based web page segmentation method which detecting the information block. The VIPS row column splitting and visual tree is generating and the noise nodes are removing from web pages or choosing an appropriate Degree of coherence value. Row Column splitting avoids the traditional clustering process in detecting information block [4].

**Shuang Lin et al. (2012)** described the extraction of contents from Web pages using Density – based approaches. Density based approaches cannot easy manage those pages which contain small contents and more noises. A tool called Block Extractor was developed based that identifies contents in three steps, looks for all Block-Level Elements (BLE) & Inline Elements (IE) blocks, which are designed to roughly segment pages into blocks. Computes the densities of each BLE&IE block and its element to eliminate noises, removes all redundant BLE&IE blocks that have emerged in other pages from the same site [5].

**LiY 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 outlinks 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 [10].

**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) presented 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 [7]. 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 the rules for removing advertisement from internet pages. ADEATER achieves high level of accuracy result [9].

3. Techniques of Content Extraction and Noise Removing From Web Pages:

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 [8] 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:

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.

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

Fig.1 DOM Tree Example
3.2.2 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. The Text Density technique can be defined as follows:

\[ TD_i = \frac{C_i}{D_i} \]

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.

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.

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 then that page is mostly navigation page or large segments are mainly content page and the content page is not divided into many small blocks [10].

3.2.1 Outlinks/pagesize: The ratio of outlinks in navigation page is more than content page. These are calculated as following:

\[ \text{OL/PS} = \frac{\text{Numberoflinks}}{\text{PageSize}} \]

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.

\[ \text{AnchorTD} = \frac{\text{AnchorText}}{\text{NonAnchorText}} \]

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:
  \[ \text{ATD} = \frac{\text{SizeOfText}}{\text{NumberOfLines}} \]

- 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

PROPOSED DOM BASED PAGE SEGMENTATION TECHNIQUE

Web Pages contain many information parts and noisy parts. The noisy parts can harm the web content mining.

DOM tree structure effects performance and does not perform well with lots of menus and short description or hyperlinks. Another problem is that various Web Pages are in HTML code and mostly traditional solutions based on analysing the HTML code. HTML is not Exclusive Web pages programing language, and many new languages have been introduced e.g. XHTML and XML. So that an efficient algorithm require which will be overcome all these problems. To overcome this problem new algorithm is used DOM Based Page Segmentations. Page Segmentation approaches used to segment the web pages into blocks, and DOM builds XML documents into tree like structure The DOM Based Page Segmentation technique observation that web pages are segmented into small blocks The parent and main block represent the all page structure. Blocks in the pages are shown in corresponds to a rectangular region. Removal of noise will be done based on these blocks and regions.

3. CONCLUSION

World Wide Web is a source of information where large amount of data is stored. The information present in the form of local and global noise. Mainly noisy data in web pages are like images, files, multimedia documents structure and unstructured data. Extracting useful information from these web pages is very complex task. Extra information from web pages like header footer, advertisements,
Navigational Bars which called noisy data needed to remove for extract the main content. A DOM Based Page Segmentation method is proposed for noise reduction and extraction of Web content from Web Pages. The navigational bar, Home page and short description noise is removed using Dom based page segmentation which convert the Web Pages into blocks and regions remove the noise and extract the information based on regions and blocks.

4. REFERENCES
1. Neetu Narwal, “Improving Web Data Extracting By Noise Removal,” IET.
4. Xuhong Zhang, Yanqing Zhang, Jing He, Frank Cobia, “Vision -based Web Page Block Segmentation and Informative Block Detecting,” 2013 IEEE/WIC/ACM International Conferences on Web Intelligence(WI) and Intelligent Agent Technology(IAT).
5. Shuang Lin, Jie Chen, Zhendong Niu (2012) “Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction”.