A Survey on Web Content Extraction and Noise Reduction from Webpage

Charmi Patel, Prof. Hiteishi Diwanji
Information Technology Department, L. D. College of Engineering, Ahmedabad, Gujarat, India

ABSTRACT

A Web Page has large amount of information. Only some 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. This irrelevant data is available with original content which is known as noisy data of website. This paper discusses various approaches for extracting informative content from web pages and removes noisy data.

Keywords: Content Extraction, Text Density, Visual Importance, DOM Tree Generation, Noisy data

I. INTRODUCTION

Web Mining is a Data Mining technique to automatically discover and extract information from World Wide Web. Web Mining is used to capture relevant data about consumer, individual user and several others. The contents of Web pages are the primary focus of Web mining applications [1]. Web Mining decomposed into Resource Discovery, Information Selection & Pre-processing, Generalization and Analysis. We can classify web mining in 3 types according to its mining techniques that is web structure mining, web content mining, web usage mining.

A user is mainly interested in the original content of web page so, the process of identifying and fetching main content blocks from a web page is called content extraction. The term content extraction was found by Rahman[2]. The content extraction is very useful for pre-processing the data in many fields such as web mining, recommendation system, decision making, expert system, knowledge discovery and so on. It is also useful to special tasks such as false advertisement detection, demand forecasting, and comment extraction on product reviews[3]. The DOM Based page Segmentation is used to discard the noisy content block and extract the informative content block from Web Pages. Initially a XML or HTML Web Page is converted into DOM tree 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 of the method. For content extraction firstly DOM tree is generated. HTML attributes, Tag pattern generation, Subject detection, Node density, Visual information, text density etc. are used for precise content extraction and removing noisy data. In this survey paper we are discussing above techniques in detail.

II. METHODS AND MATERIAL

A. Dom Tree Generation

Document Object Model (DOM) [4] is a standardized, platform-independent and language-independent interface for accessing and updating content, structure and style of any web documents. We can generate DOM tree for each HTML page where tags are internal nodes and the detailed text and images are leaf nodes. For example,

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

```
</BODY>
</HTML>
```

Dom tree for above HTML code is given below:

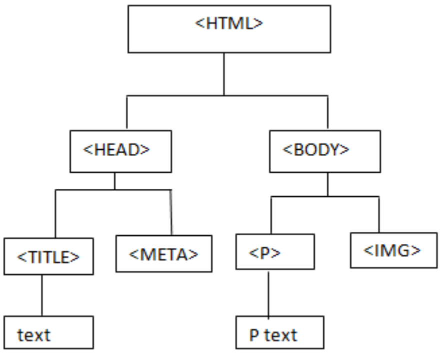

Figure 1 : DOM tree

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

B. Literature Survey

There are various techniques used for content extraction and noise removal. Each method has different percentage of content extraction and noise removal. According to the type of any website different content extraction techniques are applied for efficient and precise content extraction and noise removal.

1. SUBJECT DETECTION AND NODE DENSITY[3]

In this technique, before extracting the content we automatically detect the subject on data intensive pages of any e-commerce web sites. Main topic of data-intensive pages of e-commerce web sites is taken as subject of it. In subject detection algorithm we assign weight to each node in DOM tree by using tag name, key words used in meta tag and title tag, display properties, some CSS properties like font weight, font size. At last node which has highest weight value is taken as subject node.

Before extracting original content of web page we find data rich region of it using node density. In e-commerce web site, the node in DOM tree that contains the product detail which keep only the needed information in that page is called data rich region. In this technique node which contains the content data by defining link nodes as noise nodes which are mentioned in CECTD-DS [5] is taken as data rich region. After taking subject node as input, it assigns the current node as the subject node. Then, it reaches the parent node of the current node for deciding whether the node is the data rich region by using the threshold.

2. TEXT DENSITY AND VISUAL IMPORTANCE OF DOM NODES[6]

It is found that noise in a web page is highly formatted and contains less text and more hyperlinks and original content is simply formatted and it is lengthy too. Here, Text Density ($T\ Di$) is the ratio of its Char Number(Ci) to its Tag Number(Ti).

$$T\ Di = Ci/\ Ti$$

The node which get higher text density is commonly contain long and simply formatted text and highly formatted nodes containing less, brief text have low text density value. It is useful for determining whether a part of a web page is meaningful or not. From research it is found that most of noise in web page consist more hyperlinks. We argue that a node with too many hyperlinks and less text is less important, thus getting a low-density value; and a node that contains much non-hyperlink text and few hyperlinks is more important, and receives a high density value.

Traditionally it is found that main content of any web page is located at central part of it. For content extraction Relative displaying positions and sizes of DOM nodes are considered as useful visual information. For each leaf node in the DOM tree structure (which corresponds to an innermost tag encompassing text only), the Visual Importance considers its relative displaying size and location [6]. By combining composite text density and visual information we get hybrid text density which is used for efficient content extraction.

3. WORD TO LEAF RATIO WITH LINK ATTRIBUTES[7]

In previous technique characters are used for finding text density. Here instead of characters words are used. As we discussed above, blocks which have more links and less text is less informative. So, adding word to leaf ratio

to the text link and anchor text ratios gives more efficient content extraction technique.
After constructing DOM tree and removing noisy data word to leaf ratio is calculated by using below equation:

$$WLR (n) = tw (n) / l (n)$$

Here, tw (n) = number of words in the node n

l (n) = number of leaves in the sub tree of node n

After that text link ratio (TLR) and link text ratio (LTR) is obtained and using that weight is calculated for each node by using below equation:

$$W(n) = (TLR (n) + LTR (n)) / a$$

Here, a is normalizing factor.

Relative position(R(n)) of each node is also calculated. At last the node which has higher value of 0.7*R(n)+0.3*W(n) is taken as more informative and content of it is extarcted.

4. TAG PATTERN GENERATION AND HTML ATTRIBUTES[8]

This approach is used for news website content extraction. This approach find data rich region of news website using HTML attributes and pattern.

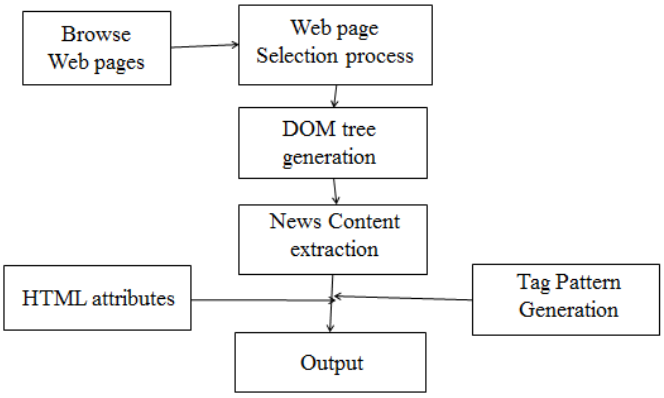

Figure 2 : System Architecture

Here user selects any news webpage as per his requirement during browsing internet. From that webpage DOM tree is generated and for each tag of that DOM tree all attribute values are retrieved using HTML attribute generation algorithm. Using Tag pattern generation algorithm some predefined pattern is generated for news website. So output of attribute generation algorithm is given as input to this algorithm and content rich area is fetched. Experiments show that using this technique news extraction is performed with higher accuracy.

5. BLOCK LEVEL ELEMENTS AND INLINE ELEMENTS BLOCK[9]

This technique describes the extraction of contents from Web pages using Density based approaches. Density based approaches cannot easily manage those pages which contain small contents and more noises. A tool called Block Extractor was developed. Based on that it identifies contents in three steps:

1) Looks for all Block-Level Elements (BLE) & Inline Elements (IE) blocks, which are designed to roughly segment pages into blocks. Here elements which are shown as block margins in a new line with independent height and width are taken as BLE and elements that are displayed in line with margins, width, and height inherited from BLE are taken as IE. BLE can contain other BLE, text or IE. IE can contain other IE and text only.
2) Computes the densities of each BLE&IE block
3) Eliminate noises, removes all redundant BLE&IE blocks that have emerged in other pages from the same site.

Working of BLE&IE blocks is same as segmentation technique. After finding these blocks sub tree is generates and density based approach is applied and redundancy from each web page is removed.

Table 1 List of methods discussed

Sr No.	Technique	Method
1	Subject Detection and Node Density	Select content reach region based on subject node having maximum weight and node density using CECTD-DS.
2	Text density and visual importance of DOM nodes	Select content reach region based on maximum value of hybrid text density.
3	Word to leaf ratio with link attribute	Select content reach region based on weight and relative position of node.
4	Tag pattern generation and	Select content reach region based on HTML

	HTML attributes	attribute value and pattern.
5	Block level elements and inline elements	Remove redundancy using density of BLE and IE blocks.

III. CONCLUSION

In this paper, we have reviewed different techniques for informative content extraction and noise removal. Here, text density and visual information is main criteria to find content rich area and take other content as noise and remove it. For precise and efficient content extraction and noise removal we can work on visual importance and also find a way to detect malicious URL from webpage for efficient noise removal.

V. REFERENCES

[1] Shuang Lin, Jie Chen, Zhendong Niu, “Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction” ,TSINGHUA SCIENCE AND TECHNOLOGY, ISSNll1007-0214ll05/18llpp256-264 Volume 17, Number 3, June 2012

[2] A.F.R.Rahman, H.Alam and R.Hartono, “Content extraction from HTML documents”, International workshop on Web Document Analysis, pp. 7-10, 2001.

[3] Warid Petprasit and Saichon Jaiyen, “Web Content Extraction Based on Subject Detection and Node Density”, 978-1-4799-6049-1/15/$31.00 ©2015 IEEE

[4] W3C Document Object Model (2009) Website. http://www.w3.org/DOM

[5] F. Sun, D. Song, and L. Liao, “DOM Based Content Extraction via Text Density,” Special Interest Group on Information Retrieval, ACM, 2011

[6] Dandan Song, Fei Sun, Lejian Liao, “A hybrid approach for content extraction with text density and visual importance of DOM nodes” , Springer-Verlag London 2013

[7] Aanshi Bhardwaj, Veenu Mangat, “A Novel Approach for Content Extraction from Web Pages”, 978-1-4799-2291-8/14/$31.00 ©2014 IEEE

[8] Yogesh W. Wanjari, Vivek D. Mohod, Dipali B. Gaikwad, Sachin N. Deshmukh, “Automatic News Extraction System for Indian Online News Papers”, 978-1-4799-6896-1/14/$31.00 ©2014 IEEE

[9] Shuang Lin, Jie Chen, Zhendong Niu, “Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction”, TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214ll05/18llpp256-264 Volume 17, Number 3, June 2012