Methods for Removing Noise from Web Pages: A Review

Maya John[1], Dr. Jayasudha J.S[2]

[1] Research Scholar, Department of Computer Science and Engineering, Noorul Islam University, Tamil Nadu, India.

[2] Professor, Department of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, Kerala, India

---***---

Abstract - *Internet has emerged as the largest repository of information pertaining to any area. It has been observed that web is noisy in nature i.e. a web page contains information that is not relevant to the user. Web page noise hinders effective web data mining. Hence it is highly essential to eliminate such type of noises. Identification and elimination of noisy information in web page is a pre-processing in variety of applications such as classification of web pages, information retrieval, displaying web pages in devices like Personal Digital Assistant (PDA) and mobile phones, web page clustering etc. Many techniques rely on the concept of Document Object Model (DOM) tree to eliminate noise from web pages. Majority of the users are interested in the informative content of a web page. So primary content of web pages are to be separated from other content sections. Web pages can be segmented into web page blocks from which the informative contents are extracted. Eliminating non content blocks from web pages reduces the storage and indexing requirements. A brief overview of various techniques used to remove noise from web pages is discussed in this paper.*

***Keywords*: Noise, DOM tree, Web pages, Blocks, Block importance.**

1. INTRODUCTION

A typical web page contains relevant piece of information as well as some distracting contents which may be of no use to the user. Informative web contents [1] include multimedia data, structured document like XML document, semi structured information such as HTML document and plain text which is unstructured in nature. Navigation bars, copyright information, advertisements [2] etc. are examples for non-informative information in web page. These non-informative contents are referred to as web page noises. Web noises are widely classified as local noise and global noise [3], [4]. Local noise is also known as intra page noise, it deals with noisy elements inside a web page. Various local noises are navigation links, advertising segments, unnecessary images etc. Global noise include large granular objects like mirror sites, duplicated web pages and old versioned web pages [1]. Eliminating noises from web page improve the performance of web page clustering, classification, content mining, summarization etc. Removing web page noise manually is a tedious and expensive task hence automatic noise elimination techniques are required. A potential application of noise removal is that if the download rate is less, a user can easily view atleast the essential contents of the web page excluding noise. In most of the web pages the main content is present in the middle block. The location, occupied area of web page etc. of the main content play an important role in distinguishing noise patterns from main content of web page [5].

2. WEB PAGE NOISE REDUCTION TECHNIQUES

Several techniques have been developed to identify the noise in web pages and to eliminate them. Many noise removal techniques are based on DOM tree which represents the layout of a web page. Based on threshold values the noisy content in web pages are found out from DOM tree. Techniques like natural language processing (NLP), filters, artificial neural network (ANN) etc. are used along with DOM tree representation of web page to eliminate noise. Other major techniques used to remove din from web pages include structural analysis and regular expression based method, layout based detachment approach (LBDA), pattern trees and heuristic based method, text density approach, image features, web page segmentation, featured DOM tree etc. The aforesaid techniques for noise removal are discussed in this section.

2.1 Based on structural analysis and regular expression

Dutta et. al. [6] proposed a structural analysis and regular expression based method for removing noise from web pages. The two main steps in this method include tag based filtering based on regular expressions and structural analysis of web page. Based on the content of HTML tags, tags are classified into positive tags and negative tags [7]. Contents of positive tag form the useful part of web page while negative tags contain information that is not useful. Negative tags include <a>, <style>, <link>, <script>, <hr>, <br> etc. The contents inside the negative tags are removed using patterns which are created using regular expressions. After filtering operation the web page is devoid of banner advertisements, images obtained from other websites, mirror sites etc. Inorder to remove noises like navigation panel, menu bar etc. structural analysis of the web pages crawled from the website is carried out. Obviously the aforesaid noise is present in every page of a web site and has same content and presentation style. During the structural analysis phase the page structure and information present in the crawled web pages are compared. Specialty of noise block is that it has same presentation style and content in all crawled web pages of a site.

2.2 Layout Based Detachment Approach(LBDA)

Deepa and Vijendran [8] devised a layout based detachment approach for extracting the content from web pages. The major steps involved in this technique are structure analysis, tag tree parsing, block acquiring page segmentation and content extraction. Structural analysis of the web page is carried out to find the tags accessible in web page, child tags and tags over inner block. Web page in HTML format is converted to XML format. DOM tree is created based on the XML file generated. HTML parser creates independent tag tree corresponding to each web page linked to a web site. The tag trees are then integrated into a single tree. Block Acquiring Page Segmentation (BAPS) technique is employed for content extraction. The unwanted tags include tags that are not closed and tags that lack child node. The unwanted tags are removed using BAPS technique. After data is extracted, boundary of blocks are eliminated to get the required information.

2.3 Based on visual block tree

Narwal [9] developed a web page noise elimination system based on visual block tree. Visual block tree is constructed for each web page using web page segmentation algorithm. Based on the visual block tree pattern tree is generated. A pattern tree consists of pattern node and information node. Each pattern node stores information regarding the layout and order of nodes at a given level. Noise elements are differentiated from main content by applying the following heuristic rules:

i) Important nodes have variety of presentation styles.
ii) Unique pattern nodes at specified level across web pages are deemed to be important.
iii) Noisy content of a web site shows consistency across web pages of that web site.

To identify the important content and noise element of a web page two similarity measures namely style importance metric and similarity count metric is used. Style importance metric measures the number of styles applied on a given node. Similarity count metric specifies how many times nodes are repeated across web pages.

2.4 Based on nX1 table and XSL display

Raheja and Katiyar [10] proposed nx1 table and XSL display method to reduce the noise in web pages. In this approach web pages are developed in the form of a table having n rows and one column. The data is inserted in the corresponding row in form of internal table. Each internal table is assigned an attribute. The main content of the web page has attribute value content, other internal tables may have attribute values like link, header, footer. The web page is converted into XML format. The XML document is displayed using XSL and the filter feature of XSL is used to extract the content of web page.

2.5 Text density approach

Eldirdiery and Ahmed [11] suggested a text density approach for detecting and eliminating noisy information present in web documents. HTML page is used as input for text density approach. The web page to be processed is divided into multiple blocks. Only the valid blocks are processed further. Blocks which contain large number of blank characters, symbols etc. are considered as invalid blocks. Noisy blocks are distinguished from non-noisy blocks by using threshold value which is computed automatically. Text density of each block is computed. Text density refers to number of words in a block. The neighboring blocks are compared by using ε value which is computed based on text density. The ε value is then compared with the threshold value. If the value of ε is greater than or equal to threshold, then block having less number of lines is considered as noisy block. In this approach only single web page is required to identify noisy content of a web page and there is no need for generating DOM representation of the web page.

2.6 Noise reduction by removing advertisements

Advertisement is an example of noisy content in a web page. HTML tag differentiator technique was developed by Hetal and Parmar [12] to remove image advertisement from web pages. Popular advertisement types in web pages are banner advertisement, video, pop up etc. In this technique DOM structure of the web page is first generated and <img> tag inside <a> tags are analyzed. Details like name of the image file, alternate text, aspect ratio etc. are gathered. The relevant features collected are given as input of predefined rule based image classifier. The rules used by classifier include domain name, difference rule, dimension rule, well known ad-provider rule and related keyword rule, advertising by scripting, dynamic advertisement rule and flashy plug in removal rule. Based on the aforesaid rules the classifier differentiates ads from other web page contents.

2.7 Using neural network

Htwe and Hla [13] developed a method based on case base reasoning(CBR) and neural network to eliminate noisy content in web pages . It's a fact that most of the noise patterns are structured by using tags like <TABLE>, <DIV>, <FRAMESET>, <SELECT>, <INPUT> etc. In this approach, CBR is used to detect the noise patterns. CBR is a machine learning approach which makes use of past experiences to solve future problems. The experiences are stored as cases and form the basis for taking decision. Each case consists of past experience and solution. To solve a problem a case base is searched for similar problem. Case base is a collection of individual cases. The different noise patterns in web sites are identified and they are stored as case in form of DOM tree. The web pages are to be converted into well formed document form before processing. According to threshold level the DOM tree structure of the web page is divided into several sub trees. The case base is then searched for similar existing noise pattern. Artificial neural network (ANN) is used for matching the given pattern. The subtrees are converted into standardized numeric representation by using the equation (1)

$x_i=S_n/T_n$ (1)

where x_i represents the input nodes at input layer, S_n is the number of occurrence of leaf nodes in sub-tree and T_n denotes the total number of leaf nodes in sub tree.

There are three classes of information namely data class, noise class and mixture of data and noise class. ANN is trained using back propagation algorithm. Here the occurrence of noise pattern is modelled using standard sigmoid activation function.

2.8 Based on image features

Yao et. al. [14] devised an image feature based technique to remove din from web pages. Same layer page noises are removed using image features. A set of same layer web pages is called as objective pages. The objective pages are passed through presentation component and converted to images. An image of a web page is divided into mutually exclusive sub images by using partition list generator. Color is the feature used to extract detail about web page sub image. Central moment of color histogram is used to calculate similarity between objective sub image and its corresponding template sub image. The major steps involved in noise reduction are calculation of global similarity degree, determination of sub image type and noise removal. Corresponding to each sub image color vectors are generated, one of the color vector is taken as eigen vector. The global similarity degree is computed using the color vectors and eigen vector. The sub images are analyzed by comparing the color similarity degree value with average one in same position. If the similarity degree between objective and template sub image is large then block is tentatively considered as noise otherwise as tentative information block. Value of SNS is calculated and if the preset ratio is greater than SNS, then template sub image is considered as information otherwise noise.

2.9 Identifying informative web content using web page segmentation

Dias and Gadge [15] proposed a web page segmentation based method to identify the important content of web page and hence eliminate noise. Set of web pages is the input for the web page segmentation technique. The web page is preprocessed by removing tags such as <a>, <b>, <script>, <span>, comments etc. The web page is then represented as DOM tree. Using only one depth child nodes [16], a sequence is generated from the DOM tree. Key pattern in the sequence is found out. Key pattern refers to the longest and most frequent repetitive pattern. Key patterns are matched with the sequence and the sub sequences are found out. Corresponding to the matching sub sequence, virtual node is created as root node. The children nodes of the virtual node are the elements of sub sequence. Importance of each block is computed by counting the number of important tags in each block. If the block importance of a given block is less than the threshold value then the block is deemed to be noisy and is eliminated.

2.10 Removing noise and duplicate contents

Sivakumar and Parvathi [1] proposed a technique to remove primary noises, duplicate contents and noisy information according to block importance from web pages. The primary noise includes copyright information, privacy notice, navigation bar, advertisements etc. Block splitting technique is used to remove the primary noises. In block splitting only the important content of a web page enclosed in div tag is considered. The main content is divided into several blocks. Simhash method [17] is employed to eliminate the duplicate contents. Simhash is a fingerprinting methodology where corresponding to each block a fingerprint is generated. The keywords in each block are identified and the frequency of each keyword in a block is found out, based on these details the fingerprint is generated. The collection of fingerprints of blocks is analyzed. A block is considered as duplicate block if its fingerprint is different from other fingerprints by at most l bit position where l denotes a small integer. Block importance is calculated based on keyword redundancy, linkword percentage and title word relevancy. Keyword redundancy refers to the percentage of redundant words in a block. The percentage of link words in a block is also computed. Percentage of keywords present in a block is referred to as title word relevancy. A block is considered to be important only if its block importance is greater than threshold value.

2.11 Using DOM and NLP

Joshi and Liu [18] developed a technique to extract text and images from web pages by using DOM analysis and NLP. In this method DOM structure is generated for each web page. HTML tag element is the information represented in each node of the DOM tree. Block element and style element are the two classes of HTML tag elements. Examples of block tags are <div>, <p>, <br>, <li> etc. while rest tags are considered as style tags. <div>, <p>, <br> tags are used to create paragraph in web pages. Rule based algorithm is used to determine which article sub tree is the main article.

Images in the web page are obtained from <img> tag. It is assumed that article images are present inside the article block. Caption of an image is obtained from text content of the first parent block element of the image. Semantic similarity between the caption of images and text in the article block are found out using NLP and cosine similarity. In this method Named Entity Recognition (NER), an NLP technique is carried out using GATE [19]. The extracted patterns of named entity present in two given blocks are then compared to determine the similarity in content. Even advertisements placed in the middle on an article can be identified using this technique.

2.12 Featured DOM tree method

Das et.al. [20] proposed a variation of DOM tree known as featured DOM tree to remove noisy contents from web pages. The three main steps involved in noise removal are featuring, modelling and pruning. DOM tree depicts the layout or presentation style of a web page and it is not sufficient to study the content or meaning of a web page. Hence featured DOM tree was developed which represents the feature set of individual blocks of a web page apart from the presentation style. The feature set is generated after web page preprocessing steps like tokenization, stop word removal, stemming etc. The featured DOM tree is created on the basis of optimal feature selection and feature weighing. In the modelling phase DOM tree is generated. Tags form the internal nodes of DOM tree and text, images, hyperlinks etc. are represented in leaf nodes. HTML web pages are analyzed from the <BODY> tag. Noise checking is carried out in the pruning phase. Minimum weight overlapping (MWO) technique is used to find out the feature set similarity. If the MWO value of a leaf node is less than the threshold value then the leaf node is marked as noisy. Bottom up traversal of the DOM tree is carried out to remove the noisy nodes. A parent node is marked as noisy if all of its children nodes are noisy in nature. The marking process is propagated up the tree and the marked portion of the tree is pruned.

2.13 Other noise removal methods

Oza and Mishra [21] proposed a method to remove web page noise based on DOM tree. They observed that at times web pages are not well formed. So the web pages are passed through HTML parser, which corrects the markup and generates DOM tree. The maximum depth of the DOM tree is found and a suitable threshold level is determined. The technique of linear regression analysis is used to find the relationship between the maximum depth and threshold level. According to the threshold level the DOM tree is split into multiple subtrees. Nodes of the DOM tree less than the threshold level is considered as noise and they are removed. Mehta and Narvekar [22] suggested a technique based on DOM tree and data filters to get rid of web page noise. DOM tree is the graphical representation of a web page and the nodes in the DOM can be manipulated by using the appropriate methods. The DOM tree contains both noisy and non noisy information. Using filters unwanted content of the DOM tree is removed. Filters contain user defined function which determines whether a node should be filtered or not. Abarna and Pradeepa [23] proposed a hybrid approach integrating region separation and information extraction can to reduce web page noise and redundancy of contents. Initially tag tree is constructed and tree pattern matching is done to obtain the data regions. Based on threshold from the data regions information regions are extracted.

CONCLUSION

The noisy content in web page hinders effective content extraction from web pages. Hence different techniques have been devised to detect and eliminate noise from web pages. Various methods employed to remove noise from web pages have been discussed in this paper. It is observed that majority of the methods employ DOM tree to eliminate noise. But DOM tree has some limitations hence improvised versions of DOM tree like featured DOM tree were developed. Apart from DOM tree analysis, structural analysis of web page, web page segmentation etc. can also be used to remove din from web pages.

REFERENCES

[1] P. Sivakumar and R. M. S. Parvathi, “An efficient approach of noise removal from web page for effectual web content mining,” Eur. J. Sci. Res., vol. 50, no. 3, pp. 345–356, 2011.

[2] S. H. Lin et. al., “Discovering Informative Content blocks from Web Documents,” In Proceedings of Eighth ACM SIGKDD International Conference knowledge Discovery and Data Mining, 2002, pp. 588–593.

[3] L. Yi et. al. , “Eliminating Noisy Information in Web Pages for Data Mining,” In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[4] L. Yi et. al., “Web page cleaning for Web mining through feature weighting,” In Proceedings of Eighteenth International Joint Conference on Artificial Intelligence, 2003.

[5] R. Song et. al., “Learning Important Models for Web Page Blocks based on Layout and Content Analysis,” ACM SIGKDD Explor. Newsl., vol. 6, pp. 14–23, 2004.

[6] A. Dutta et. al., “Structural Analysis and Regular Expressions based Noise Elimination from Web Pages for Web Content Mining,” In Proceedings of IEEE International Conference on Advances in Computing, Communications and Informatics, New Delhi, India, 2014, pp. 1445–1451.

[7] B. H. Kang and Y. S. Kim, "Noise Elimination from the Web Documents by Using URL paths and Information Redundancy," In Proceedings of International Conference on Information and Knowledge Engineering, Las Vegas, Nevada, US, 2006, pp. 26–29.

[8] C. Deepa and A. S. Vijendran, "LBDA : a Novel Framework for Extracting Content from web pages," In Proceedings of IEEE International Conference on Advanced Computing and Communication systems, Coimbatore, India, 2013.

[9] N. Narwal, "Improving web data extraction by noise removal," In Proceedings of IET 5th International Conference on Advances in Recent Technologies in Communication and Computing. Bangalore, India, 2013.

[10] N. Raheja and V. K. Katiyar, "A Noise Reduction Approach based on n x 1 Table and XSL Display Method for Efficient Web Data Extraction," Int. J. of Comput. Appl., vol. 64, no. 11, pp. 1–6, 2013.

[11] H. F. Eldirdiery and A. H. Ahmed, "Detecting and Removing Noisy Data on Web Document using Text Density Approach," Int. J. Comput. Appl., vol. 112, no. 5, pp. 32–36, 2015.

[12] J. G. Parmar and R. Hetal, "Removal of Image Advertisement from Web Page," Int. J. Comput. Appl., vol. 27, no. 7, pp. 1–5, 2011.

[13] T. Htwe and K.H.S. Hla, "Noise removing from Web pages using neural network," In Proceedings of 2nd International Conference on Computer and Automation Engineering, Singapore, 2010, pp. 281–285.

[14] H. Yao et. al., "The Noise Reduction Method of Web Pages Based On Image Features," In Proceedings of IEEE International Conference on Computational Intelligence and Software Engineering, Wuhan, 2009, pp. 1-5.

[15] S. Dias and J. Gadge, "Identifying Informative Web Content Blocks using Web Page Segmentation," Int. J. Appl. Inf. Syst., vol. 7, no. 1, pp. 37–41, 2014.

[16] J. Kang et. al., "Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices," IEEE Trans. Consum. Electron., vol. 56, no. 2, 2010.

[17] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," In Proceedings of the 34th Annual ACM symposium on Theory of Computing, 2002, pp. 380–388.

[18] P. M. Joshi and S. Liu, "Web document text and images extraction using DOM analysis and natural language processing," In Proceedings of the 9th ACM symposium on Document engineering, 2009, pp. 218-221.

[19] H. Cunningham et. al., "GATE: A framework and graphical development environment for robust NLP tools and applications", In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.

[20] S. N. Das et. al. , "Eliminating Noisy Information in Web Pages using featured DOM tree," Int. J. Appl. Inf. Syst., vol. 2, no. 2, pp. 27–34, 2012.

[21] A. K. Oza and S. Mishra, "Elimination of noisy information from web pages," Int. J. Recent Technol. Eng., vol. 2, no. 1, pp. 115–117, 2013.

[22] B. Mehta and M. Narvekar, "DOM Tree based approach for web content Extraction," In Proceedings of IEEE International conference on Communication, Information and Computing Technology, Mumbai, India, 2015, pp.1-6.

[23] R. Abarna and S. Pradeepa, "A Hybrid Approach for Extracting Web Information," Indian J. Sci. Technol., vol. 8, no. 17, 2015.