An Efficient Method of Web Page Noise Cleaning for Effective Web Mining

S. S. Bhamare
School of Computer Sciences
North Maharashtra University
Jalgaon (MS), India

B. V. Pawar
School of Computer Sciences
North Maharashtra University
Jalgaon (MS), India

ABSTRACT
In the huge network of World Wide Web, web pages contained large amount of information. Web researches are always requiring main content (e.g., an article text) from the web pages to be gathered, processed and stored quickly and efficiently. Mining the data on the Web has become a major task for locating useful information from the Web. The Web information's that are considered as useful information usually has huge amounts of noise data's such as navigation bars, links, advertisements, copyright notices etc. Performance of Web mining can be improved by identifying and removing noises from Web pages. In this paper new method is proposed for removing noise content tag and extracts the information of main content tag from web pages.

General Terms
Web Mining, Global Noises, Local Noises, DOM, Web Pages and WWW.

Keywords
WPNC, Noise Block, HTML Tag, White Listed tags, HDT, LDT, Black Listed Tags.

1. INTRODUCTION
The number of web pages along with data or information is continuously increased on internet. This data exists on web are in different types e.g., structured tables, semi structured web pages, unstructured texts, and multimedia files (such as Pictures or Images, video and audio's clips) this information on the Web is in heterogeneous form.

Web mining aims to discover useful information or knowledge from Web hyperlinks, page contents, and usage logs. Based on the primary kinds of Web data used in the mining process, Web mining tasks can be categorized into three main types: Web structure mining, Web content mining and Web usage mining. Web content mining extracts required useful information or knowledge from Web page contents [2].

Inner Content of web pages provide basic source of information used in many Web mining tasks, but with this useful information in Web pages is often accompanied by a large amount of noise content such as advertisements, navigation bars, links, and copyright notices. This information items are functionally useful for human internet browsers and necessary for the Web site owners, but this information hamper automated information collection system and Web mining tasks, e.g., information retrieval and information extraction, Web page clustering and Web page classification. Web page noise cleaning is the preprocessing step of Web documents to deal with such noisy information. In general, noise refers to less important, irrelevant or harmful information.

This paper focuses on the removal of web page noise (local noise) from web pages.

2. WEB PAGE NOISE
In web environment, Yi and Liu [3] categorized noise data into two groups such as global noise and local noise

Global noises are redundant or repeated Web pages over the Internet such as mirror sites and duplicated legal or illegal Web pages.

Local noises only related intra-page redundancy and exist in the Web page. There are different known categories of noise pattern within Web pages of any Web sites including banners with links, advertisements, directory list or navigational panel, copy right and privacy notice in each Web site.

Several Web pages contain these noise pattern together but most of noise patterns are organized by using web page interactive tags, sectioning & sectioning separating tags such as <SELECT>, <FIELDSET>, <FRAMESET>, <INPUT>, <TABLE> & <DIV>,. Additionally, anchor tag <A> and <IMG> tag are most frequently used to link another Web page or another Web site [4].

2.1 Web Page Data
In the study of web pages of different web sites. Typically webpages consists of HTML tags, the most important tags among the various tags available in the HTML web script are head, title and body. HTML tags are splitted based on the web page content, that are composed in the HTML web script.

Generally HTML tags can be divided into two categories such as [5],

[1]**Container or Main Content tag:** Used to design the layout of the web page, these container tags visually divided the web page into several content blocks. Common container or main content tags include <body>, , <table>, <tr>, <td>, <ul> and <form> etc.

[2]**Designer or Noise Content Tag:** Used to describe a segment of contain in the web page, these tags have no use of the layout but only use for picture or a hyperlink on the web pages, common designer or noise content tags include <a>, <img>, <font> and etc. Generally main content i.e. informative content are found in <BODY> tag and its corresponding tags. We are interested at most to extracts the information of <BODY> and there corresponding tags.

3. RELATED WORKS
Researchers have worked in this area for retrieving and extracting main content and removing noise data from different Web pages. Most of them have focused on detecting main content and informative blocks in Web pages; relatively list of the work has been done in this field such as,

Kushmerick [6] proposed some learning mechanisms to recognize banner ads, redundant and irrelevant links of Web pages. However, these techniques are not automatic. They require a large set of manually labelled training data and also domain knowledge to generate classification rules.

Kao et al. [7] enhances the HITS algorithm by using the entropy of anchor text to evaluate the importance of links. It focuses on improving HITS in order to find informative or useful structures in Web sites, though it segments Web pages into content blocks to avoid unnecessary authority and hub propagations, it does not detect or eliminate noisy contents in Web pages.

Kao, Ho, and Chen [8] InfoDiscoverer, proposed a new approach to find out informative contents from a set of tabular documents of a web site by dynamically selecting the entropy threshold. In this approach, first partitioned a page into several content blocks according to HTML tag <TABLE> in a Web page.

Most of the techniques needs for extracting the content structure of a web page. Researchers have considered using the html tag information and dividing the page based on the type of the tags. Some useful tags include <P> (paragraph), <TABLE> (table), <UL> (list), <H1>~<H6> (heading), etc.

Diao et al. [9] treats segments of web pages in a learning based web query processing system and deals with these major types of tags.

Kaasinen et al. [10] split the web page by some easy tags such as <P>, <TABLE> and <UL> for further conversion or summarization.

Wong et al. [11] defines tag types for page segmentation and gives a label to each part of the web page for classification. Besides the tag tree, some other algorithms also make use of the content or link information.

In [12], different supervised and unsupervised web page noise cleaning techniques like classification based, template based, segmentation based, SST based and CST based cleaning techniques are discussed.

4. PROPOSED WORK

The goal of this proposed work is to identify main content tags and noise content tags of web pages and remove the noisy data to improve the performance of web mining tasks.

Our empirical evaluation is done using 155 web pages of different categories from 3 web sites, CNNIBN, ABB News and Times of India and empirical outcome is discussed in this paper.

4.1 Overview of Cleaning Process

The noise cleaning process goes through different steps to identify and remove noise content information from web pages for effective web mining.

Following Figure 1. Shows overall flow of web page noise removal (Cleaning) process.

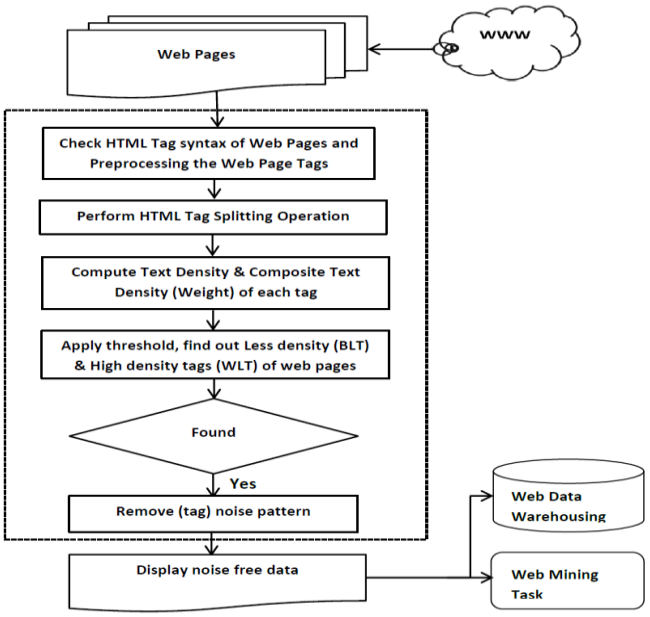


Figure 1. Process flow diagram for web page noise removal

4.2 Proposed Method

In this proposed method, first check the HTML syntax of web document because most of the HTML Web pages are not well-formed, through an HTML parser [13] correct the web page markup and then transformed it into HTML tag tree for further processing. This proposed method compute text density or tag weight of each tag including their corresponding tags of each page. Using best threshold value high density tags and less density tags of web pages are identified. Figure 2. Shows proposed method for web page noise cleaning (WPNC).

Begin

Step 1: S: Access multiple web pages

Step 2: For each page i

Load HTML Tags of web pages (used HTML DOM Parser).

Step 3: Check Web Page HTML tags

Step 4: Preprocessing the web page tags: Those tag that does not contain any text and invalid tags which are not related to the main content should be filtered out or removed.

Step 5: Separate [1]Container or Main Content Tags and [2]Designer or Description Tag based on importance of text of each WebPage.

Step 6: Create files {i.e. container tag file & designer tag file}.

Step 7: For each page i

TDi = Compute a Text density & composite Text Density of each tags (including corresponding tags) of container file. i.e

$$TDi = TCi / Ti$$

Step 8: For each page i

Tr = Using appropriate threshold, find out Less & High density tags,

If

$$TDi > Tr$$

then

HDT
Else
LDT

Step 9: HDT = Consider, high density tags named as White Listed Tags (I.e. Main Content Tags that should be extract and LDT = less density tags named as Black Listed Tags (i.e. noise content Tags) that should be removed.

Step 10: Finally receive noise free web page data.

End

Figure 2. Proposed Method for WPNC

4.2.1 Preprocessing the Web Page Tag

In preprocessing of web page tags, those tags that do not contain any text, as well as invalid tags such as <script>, <style>, <marquee> <meta>, <anchor> etc., which are not related to the main content of web page should be filtered out or remove first to improve the noise removal accuracy of web pages.

4.2.2 Splitting Operation

On a given set of web pages, splitting operation is carried out on HTML tags and used in web page cleaning process. Web page tags are split into two types i.e. [1]container or main content tag and [2]designer or noise content tag. Usually the content is composed in a Body tag of web page within the DIV and TD sub tags. These contents of web page are said to be informative content.

4.2.3 Compute Text Density

For this experiment, information is collected from web pages, importance of both main content and noise are taken into considerations [14].

Text Density: Text density will help to found the noise which is more formatted and contain a small text and main content usually lengthy and less formatted. Once an HTML Pages has been parsed, the number of characters per tag and tags that each node contains can be figured out.

Text density or Tag weight can be defined as,

$$TDi = TCi / Ti$$

Where,

TDi = Text Density of each tag (node)
TCi = Count No. of Characters of each tag (including corresponding tags)
Ti = Count No. of tags (including corresponding tags)

Composite Text Density: We find that noise in the web pages consists of hyperlinks by adding statistical information about hyperlinks per tag i.e. number of hyperlink character and number of hyperlink tag is called the Composite Text density.

4.2.4 Identify HDT and LDT for Content Extraction

A threshold Tr is used to identify High density tag (usually lengthy and less formatted i.e. HDT) and Less density tags (usually more formatted and contain a small text i.e. LDT) based on text density of tags of each page and determine main content tags and noise content tags of web pages. The best value of the threshold is found and used and useful content is extracted.

We categories these web page tags in to two types,

Number of Black Listed Tag (BLT) per Page: Less density tag element of each web page.

Number of White Listed Tag (WLT) per Page: High density tag element of each web page.

To identify HDT and LDT of web pages as below,

Less Density Tag and High Density Tag using appropriate Threshold :

If TDi > Tr
Then HDT
Else LDT

Where ,

TD_i = Text Density of each tag (node)
Tr = Threshold Value
HDT = High Density Tag
LDT = Low Density Tag

With the help of these tag information, to remove less density tags (i.e. Black Listed Tag) and its data because these tag supposed to be noise data tag or pattern. Generally web page main content tags can hold large text or characters as compared with noise content or pattern tags of web pages.

Following Table 1. Shows number of black listed tags (BLT) and noise ratio per page for nine sample web pages out of 155 web pages,

Table 1: Web page tag information

Web Site Name	Source Category	B L T per Page*	Total No. of Tags per Page	% Noise Ratio Per Page
CNNIBN	Main	340	540	62.96
	Sport	305	438	69.63
	Technology	400	521	76.77
ABB News	Main	1322	1480	89.32
	Sport	870	138	62.68
	Technology	820	1064	77.06
Times of India	Main	45	66	68.18
	Sport	56	96	58.33
	Technology	52	87	59.77

4.3 Performance Evaluation

This experiment provide us statistical information of web pages for performance evaluations. While working on tag information of web pages, this proposed method gives us web page tag data before and after web page cleaning. Web page data for nine sample web pages out of 155 pages as given below,

Table 2: Web page tag information before and after removal of noise for each page

Web Site Name	Source Category	Total No. of Tags per Page	% Noise Ratio Before Cleaning Per Page	% Noise Ratio After Cleaning Per Page	Noise Removal Accuracy Improvement per Page
CNNIBN	Main	540	62.96	24.44	38.52
	Sport	438	69.63	18.26	51.37
	Technology	521	76.77	26.87	49.90
ABB News	Main	1480	89.32	24.05	65.27
	Sport	1388	62.68	22.69	39.99
	Technology	1064	77.06	23.30	53.76
Times of India	Main	66	68.18	21.21	46.97
	Sport	96	58.33	21.85	36.48
	Technology	87	59.77	20.68	39.09

In this above table accuracy of noise removal is increased after cleaning the web page.

The accuracy of noise removal on set of web pages is also checked and has been computed as follow,

Noise Before Cleaning (NBC):

$$NBC = \frac{\sum_{i=1}^{n} PBi}{n}$$

Where,

n = Sample number of web pages

PB*i* = Percentage of noise before cleaning for each $_i^{th}$ web page

Noise After Cleaning (NAC):

$$NAC = \frac{\sum_{i=1}^{n} PAi}{n}$$

Where,

n = Sample number of web pages

PB*i* = Percentage of noise after cleaning for each $_i^{th}$ web page.

Noise Removal Accuracy Improvement (NRA):

$$NRA = NBC - NAC$$

Where,

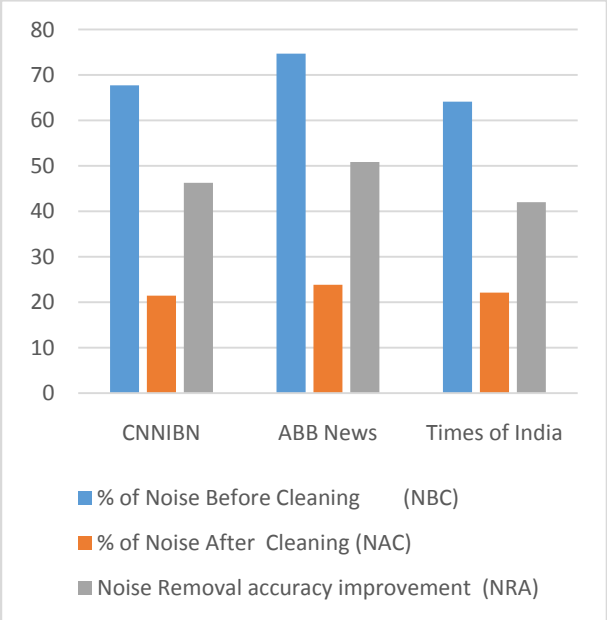
NBC = Noise Before Cleaning

NAC= Noise After Cleaning

The following table shows improvement in noise removal accuracy after cleaning the set of 155 web pages,

Table 3: Web Page tag information before and after removal of noise on set of web pages.

Web Page Source	% of Noise Before Cleaning (NBC)	% of Noise After Cleaning (NAC)	Noise Removal accuracy improvement (NRA)
CNNIBN	67.72	21.46	46.26
ABB News	74.70	23.84	50.86
Times of India	64.12	22.10	42.02



Graph 1: Ratio of noise tag information before and after noise tag removal on set of web pages.

In experimental results of proposed method, we observed that Noisy less Web Pages before Cleaning is **31.15** percent and Noisy less Web Pages after cleaning is **77.53** percent. Thus, in this experiment noise removal accuracy of web pages has been improved by **46.38** percent.

5. CONCLUSION

In this proposed method of web page noise cleaning text density is computed of each tag including their corresponding tag of each page. Through best threshold value, main content tag i.e. High density tag and noise tag i.e. Less density tag of web pages are identified and only the information of high density tags is extracted for further web mining tasks. Those tags that do not contain any text, as well as invalid tags which are not related to the main content should remove to improve the performance of noise cleaning from web pages. In this experiment noise removal accuracy of web pages has been improved by 46.38 percent. In the study of existing methods or technique of web page noise cleaning, not a single methods or technique can remove all types of noises from web pages

most of them work only to remove some specific types of noises from web page.

6. REFERENCES

[1] R. Kosala and H. Blockheel. Web Mining Research: A Survey. In SIGKDD Explorations, Vol. 2, No. 1, pp 1-15, 2000.

[2] Bing Liu, Web Data Mining (Exploring Hyperlinks, Contents, and Usage Data), Springer.

[3] L. Yi, B. Liu, and X. Li. Eliminating noisy information in web pages for data mining. In Proceedings of the International ACM Conference on Knowledge Discovery and Data Mining, pages 296–305, 2003.

[4] Thanda Htwe. Cleaning Various Noise Patterns in Web Pages for Web Data Extraction, International Journal of Network and Mobile Technologies ISSN 1832-6758 Electronic Version VOL 1 / ISSUE 2 / NOVEMBER 2010 © 2010 INTI University College.(University of Computer Studies, Yangon, Department of Software Technology, tdhtwe80@gmail.com)

[5] Hu Fei, Yang Huaqian, Wei Pengcheng, Pu Changjiu, Lei Yang, Web Page Noise Reduction Algorithm Using Non-template Approach in International Journal of Digital Content Technology and its Applications(JDCTA)Volume6, Number20, November 2012

[6] Kushmerick, 1999] Nicholas Kushmerick. Learning to remove Internet advertisements. Agnets-1999, 1999.

[7] Kao et al., 2002] Hung-Yu Kao, Ming-Syan Chen Shian-Hua Lin, and Jan-Ming Ho, Entropy-Based Link Analysis for Mining Web Informative Structures. CIKM-2002, 2002.

[8] H. Y. Kao, J. M. Ho, and M. S. Chen, Wisdom Web intrapage informative structure mining based on document object model in IEEE Trans KDD, 2005.

[9] Diao, Y., Lu, H., Chen, S., and Tian, Z., TowardLearningBased Web Query Processing, In Proceedings of International Conference on Very Large Databases, 2000, pp. 317-328.

[10] Kaasinen, E., Aaltonen, M., Kolari, J., Melakoski, S., and Laakko, T., Two Approaches to Bringing Internet Services to WAP Devices, In Proceedings of 9th International World-Wide Web Conference, 2000, pp. 231-246.

[11] Wong, W. and Fu, A. W., Finding Structure and Characteristics of Web Documents for Classification, In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Dallas, TX., USA, 2000.

[12] S. S. Bhamare, Dr. B. V. Pawar "Survey on Web Page Noise Cleaning for Web Mining" in International Journal of Computer Science and Information Technologies (IJCSIT) Volume 4 Issue 6, Nov-Dec. 2013, ISSN: 0975-9646.

[13] The HTML DOM Parser Library Version 2.0, [Online] Available: http://thehtmldom.sourceforge.net

[14] Dandan Song, Fei Sun, Lejian Liao.‖ A hybrid approach for content extraction with text density and visual importance of DOM nodes‖. In the proceedings of Springer Knowl Inf Syst, DOI 10.1007/s10115-013-0687-x, Verlag London 2013.

[15] Jing Li and C.I. Ezeife. Cleaning Web Pages for Effective Web Content Mining School of Computer Science, University of Windsor, Windsor, Ontario, Canada N9B 3P4, cezeife@uwindsor.ca, http://www.cs.uwindsor.ca/_cezeife.

[16] YI L. et LIU B. (2003), "Web Page Cleaning for Web Mining through Feature Weighting", in Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03).

[17] A. Rahman, H. Alam, and R. Hartono. Content extraction from html documents. In 1st Int. Workshop on Web Document Analysis (WDA2001).

[18] B.D. Davision. Recognizing Nepotistic links on the Web. Proceeding of AAAI 2000.

[19] Hu Fei, Li Ming, Ma Yan" Eliminating Noisy Information in Web Pages based on Source Code Shrinking", International Journal of Advancements in Computing Technology (IJACT), Vol.4, No. 18, October 2012.