ISSN:0975-9646

Survey on Web Page Noise Cleaning for Web Mining

S. S. Bhamare, Dr. B. V. Pawar

School of Computer Sciences
North Maharashtra University
Jalgaon, Maharashtra, India.

***Abstract*—Web Page Noise Cleaning is one of the new research area 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.**

***Keywords*— WWW, Web Page Cleaning, Noise Block, DOM Tree, Web Mining, Web pages.**

I. INTRODUCTION

The rapid expansion of internet has made World Wide Web (WWW) a popular place for disseminating information. This provides very useful and helpful information. Current estimates shows that there are more than seven billion Web pages in WWW. The web mining is the application of data mining techniques to automatically discover and extract patterns or information from World Wide Web (WWW) documents and services. In Web mining, collection of data is an important task, specifically for Web structure mining and content mining, in this crawling a large number of target Web pages are involved. We should also note that web users play an important role in the information extraction or knowledge discovery process on the web since the World Wide Web is an interactive medium for this.

In the WWW, the inner content of Web pages is providing basic information or source of information used in many Web mining tasks. But unfortunately, this useful information in Web pages is often accompanied by a various type of noise such as advertisement banners, navigation bars, and disclaimer/copyright notices. Although such noise information items/blocks are useful for web browsers and also necessary for the owners of web site, noise also affect automated information gathering or collection and Web mining tasks, e.g., information retrieval and extraction, Web page clustering and Web page classification. So, In general, web page noise refers to redundant, incoherent with main content or harmful information. Web page cleaning is the preprocessing step or task of Web documents to deal with such noisy information.

II. WEB PAGE NOISE

In the World Wide Web, Noise on the web pages are not the part of the main content and this irrelevant information (called web page noise) in web pages can really affect web mining task.

Lan Yi [12] has grouped noise data of Web documents into two categories according to their granularities:

Global noises: These are noises on the WWW with large granularity, which are usually no smaller than individual pages. Global noises include mirror sites, legal/illegal duplicated Web pages and old Web pages to be deleted, etc.

Local (intra-page) noises: These are noisy regions/items/pattern/blocks within a Web page of web site. Local noises are usually incoherent with Web pages main contents. Such noises include banner ads, copyright notices navigational guides, decoration pictures etc should be removed for effective web mining.

Web mining task decomposed into the subtasks, namely, Resource finding, Information selection and pre-processing, Generalization and Analysis. Apart from these tasks enumerated under Web mining, another task viz. '**cleaning**' be applied in web content mining with objective of removing redundancy(i.e. global noise) or Web Page Noise(i.e. local noise).

III. MAJOR TECHNIQUES OR METHODS ADOPTED

In this area of Web Page noise cleaning some of the techniques, methods and approaches are already developed and used for improving the result of web mining task. Such as,

Classification Based Cleaning method is simple method of Web page cleaning to detect specific noisy items (e.g., advertising images, nepotistic hyperlinks, etc.) in Web pages by adopting some pattern classification techniques. The classification based method is supervised and semi-automatic. All existing classification based cleaning methods simply adopt decision tree classifier to detect noisy items in Web pages.

Decision tree classifier is a classic machine learning techniques are used in many research fields. The ID3 (Iterative Dichotomiser 3) algorithm and the C4.5 algorithm are two widely used decision tree methods till now. The decision tree classifier technique can be adopted to detect certain kind of noisy items (e.g., images and linkages) in Web pages. For example, **Davisons work [18]** proposed decision tree based system is, namely **AdEater** that detects and cleans advertising images in Web pages. The AdEater system first defines features for images in

Web pages. Also **Paeks work [19]** trains the decision tree classifier to recognize banner advertisements.
The main steps of decision tree based Web page cleaning are as below:

1. Define nominal features for the target type of item (e.g.,
   Images, linkages, etc.)
2. Build decision tree based on (noisy and non-noisy) sample items and extract rules
3. Determine noisy items from non-noisy ones by created
   decision tree or rules

Image and linkages are not the only types of items in Web pages. To build decision trees for each type of item is inefficient and inapplicable in practice. The classification method is not site dependent and easily clean web pages from other web sites also.

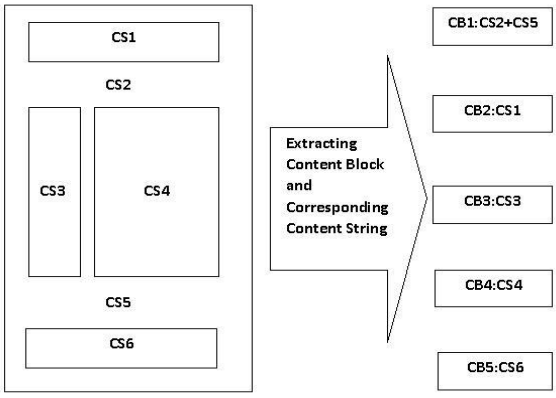
In Lin and Ho [5], a segmentation based cleaning method is supervised cleaning method proposed to detect informative content blocks in Web pages based on the observation that a Web site usually employs one or several templates to present its Web pages. A set of pages that are presented by the same templates is called page cluster. Assuming that a Web site is a page cluster, this work classifies the content blocks in Web pages into informative ones and redundant ones. Informative content blocks are the distinguished parts of the page whereas redundant content blocks are common parts.
The segmentation based cleaning method discovers informative blocks in following four steps:

i) **Page segmentation** step extracts out each <TABLE> in the DOM tree structure of a HTML page to form a content block. The rest contents which are not contained in any <TABLE> also form a special block.

ii) **Block evaluation** step selects feasible features (i.e, terms) from blocks and calculates their corresponding entropy values.

iii) **Block classification** step decides the optimal block entropy threshold to discriminate the informative content blocks from redundant content blocks.

iv) **Informative block detection** step simply classify content blocks into informative ones and redundant ones according to the decided optimal threshold.



Above figure shows the content blocks extracting from a sample page, where each rectangle denotes a table with child tables and content strings. Content blocks CB2, CB3, CB4 and CB5 contain content strings CS1, CS3, CS4 and CS6 correspondingly. The special block CB1 contains strings CS2 and CS5 which are not contained in any existing blocks.

In **Bar-Yossefs and S. Rajagopalan work [6], a template based cleaning method** is a unsupervised and automatic noise cleaning method is proposed to detect templates whereas the templates found are viewed as local noisy items/data in Web page. Basically, the template based cleaning method first partitions Web pages into pagelets and then detects frequent templates among the pagelets. Page partition step segments all Web pages into logically coherent pagelets. In the template based cleaning method, Web pages are assumed to consist of tiny pagelets.
The pagelet is defined as follows:
Definition (pagelet): An HTML element in the parse tree of a page p is a pagelet if (1) none of its children contains at least h hyperlinks; and (2) none of its ancestor elements is a pagelet.
In template Detection step finds those frequently occurred pagelets in different Web pages as templates. A collection of pagelets is called template.

Figure shows pagelet examples in the main page of Yahoo!.

There are two algorithms for template detection:
The first one is the local template detection algorithm which is suitable for the document sets that consist of small fraction of documents from the larger universe. The local template detection algorithm in fact only satisfies the first requirement of template definition.
The second algorithm is the global template detection algorithm which is suitable for template detection in large subsets of the universe. It requires the detected templates to be undirected connected by hyperlinks.
The template based cleaning method in [6] is not concerned with the context of a Web site, which can give useful clues for web page cleaning. In template based cleaning, the partitioning of a Web page is pre-fixed by considering the

number of hyper-links that an HTML element has. This partitioning method is very simple and useful for a set of Web pages from different Web sites, while it is not suitable for Web pages that are all from the same Web site because a Web site typically has its own common layouts or presentation styles, which can be exploited to partition Web pages and to detect noises.

SST based cleaning technique is a "partially" supervised cleaning technique, based on the analysis of both the layouts and the actual contents (i.e., texts, images, etc.) of the Web pages in a given Web site.

This approach mentioned in L. Yi, B. Liu [12] builds Site Style Tree in simple manner, which is generalized DOM tree presentation of related pages. The DOM (Document Object Model) tree structure is widely used to model individual HTML and XML Web pages. The tree structures of Web pages are useful for detecting and removing Web page noise.

To construct the Site Style Tree, the system needs to learn the whole web site to detect the common presentation style and content. They identified the common presentation style and content and then compressed them into site style tree. Noisy elements in the tree are detected based on entropy calculation over set of features. Some Web sites are structured with dynamic Web pages and their content and presentation style are not common. It is difficult to detect different noise patterns for those Web sites by using above technique. These techniques are less successful in identifying noise patterns which vary from expected patterns.

Building a style tree (called site style tree or SST) for the pages of a Web site is fairly straightforward. We first build a DOM tree for each page and then merge this DOM tree into the style tree in a top-down fashion. It gives the algorithm for detecting and eliminating noises given a SST and a new page from the same site.

The algorithm basically maps the DOM tree of the page to the SST, and depending on where each part of the DOM tree is mapped to the SST, to find whether the part is meaningful or noisy by checking if the corresponding element node in the SST is meaningful or noisy. If the corresponding element node is neither noisy nor meaningful, we simply go down to the lower level nodes of tree.

Feature weighting based cleaning method is an improved version of SST based method. This method is an unsupervised and automatic noise cleaning method. The approach mentioned in Yi & Liu [10] discuss a tree-based approach that combines features based on HTML tree structure, content, and visual representation. To construct a Compressed Structure Tree (CST Tree); elements within the tree of every document are combined (compressed) if their child elements share the identical tag names, attributes, and attribute values. Based on the number of different presentation styles for an element, the weight (importance) of this element is determined. The resulting weights are then utilized in follow-up tasks (e.g. classification). While this approach performed well in the follow-up classification task; it depends on the availability of a relatively large number of documents from a limited number of sources. Yi & Liu used documents from distinct sources. The documents from each source had very similar structure which made the use of the compressed structure tree a viable option.

These above major Web Page Cleaning techniques or methods go through following four main steps:

a) Page segmentation manually or automatically segments a Web page into small blocks focusing on coherent subtopics.
b) Block matching identifies logically comparable blocks in different Web pages.
c) Importance evaluation measures the importance of each block according to different information or measurements.
d) Noise determination distinguishes noisy blocks from non-noisy blocks based on the importance evaluation of blocks.

IV. Other Related Work

Many 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. Although cleaning noisy data is an important task, relatively list of the work has been done in this field such as,

Kushmerick [7] 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. [8] enhances the HITS algorithm of 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 [9] InfoDiscoverer, was proposed an approach to discover informative contents from a set of tabular documents of a Web site by dynamically select the entropy threshold. The system first partitioned a page into several content blocks according to HTML tag <TABLE> in a Web page. The system is not applicable general Web pages which is consisted using tag <DIV>.

Debnath et al. [11] proposed an approach similar to the one in [12]. They also select portions of web pages, called blocks, which have importance level above a given threshold. However, while Yi [12] defined the notion of importance based on features of the whole target site summarized in style trees; Debnath estimate the importance for each individual block. Two distinct strategies are proposed. One is based on occurrence of similar blocks among the pages of the site and another is based on a specific predefined set of desired features that must be present on blocks.

The need of techniques for extracting the content structure of a web page. Many researchers have considered using the tag information and dividing the page based on the type of the tags. Useful tags include <P> (paragraph), <TABLE> (table), <UL> (list), <H1>~<H6> (heading), etc. **Diao et al. [14]** treats segments of web pages in a learning based web query processing system and deals with these major types of tags.

Kaasinen et al. [15] split the web page by some easy tags such as <P>, <TABLE> and <UL> for further conversion or summarization.

Wong et al. [16] 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.

Gupta et al. [20] have proposed a DOM-based content extraction method to facilitate information access over constrained devices like PDAs. They implemented an advertisement remover by maintaining a list of advertiser hosts, and a link list remover based on the ratio of the number of links and non-linked words.

Cai Deng et al. [21] VIPS makes full use of page layout features and some heuristic rules to partition the web page at the semantic level. The main limitation of this approach is that performing visual rendering and segmentation of web pages is resource intensive.

Tripathy, Singh [22] proposed cleaning technique that is based on the analysis of both the layouts and the actual contents (i.e., texts, images, etc.) of the Web pages in a given Web site. Thus, in first task of proposed technique is to find a suitable data structure to represent both the presentation styles (or layouts) and the actual contents of the Web pages in the site. They propose a Pattern Tree (ST) to capture those frequent presentation styles and actual contents of the Web site. The site pattern tree (SPT) provides us with rich information for analyzing both the structures and the contents of the Web pages and used an information based measure to evaluate the importance of element nodes in SPT so as to detect noises to clean a page from a site, they simply map the page to its SPT.

Thanda Htwe et al. [23] propose the mechanisms to eliminate multiple noise patterns in Web pages to reduce irrelevant and redundancy data by applying Case-Based Reasoning technique to detect multiple noise patterns in current Web page and also present back propagation neural network algorithm for matching current noise with storing noise patterns for noise classification, and then they remove this noise pattern in current page for content extraction.

YongZhang, Ke Deng [24] they proposed a new approach of the web page purification based on improved DOM and statistical learning. They produced Block Tree structure that is very helpful for applications such as web page classification, information retrieval and information extraction and also using of statistical learning, it makes easily to find the main content block of the web page.

Guohua Hu, Qingshan Zhao [25] in this approach, they first introduce a new tree structure, called style tree, to capture the common layouts (or presentation styles) and the actual contents of the pages in a Web site. Then propose an information based measure to determine which parts of the style tree indicate noises and which parts of the style tree contain the main contents of the pages in the Web site. To clean a new page from the same site, they simply map the page to the style tree of the site and according to the mapping, can decide the noisy parts and delete them.

V. RESEARCH ISSUES AND CHALLENGES

Web page cleaning is not an independent research topic because the Web page noise is task dependent which is always related to Web tasks. Therefore, the categorization of Web page noise and the Web page cleaning are two critical tasks to improve the Web mining results and to help many Web page content based tasks, e.g., information retrieval, information extraction, Web data warehousing, etc.

The main issue with existing Web page cleaning methods is that they do not recognize or find all kinds of Web page noise patterns to remove. The most of the methods work only on some specific kind of noises (such as, advertising images or nepotistic hyperlinks etc.) to remove for web page cleaning, they worked only on general noises are found on web pages. Some Web sites are structured with dynamic Web pages and their content and presentation style are not common. It is difficult to detect different noise patterns for those Web sites by using above technique. These techniques are less successful in identifying noise patterns which vary from expected patterns. So, these techniques & methods are site dependent and performance is based on similarities of web pages.

This research area presents new challenges related with above issues. Try to overcome these issues by developing such method or framework that can work on most of the noise patterns found on web pages and also works on web pages of different web sites that means this technique & method are not site dependent and performance is not based on similarities of web pages.

VI. CONCLUSION

The Web page noise and the research of Web page cleaning is a newly proposed topic. The pre processing task Web page cleaning is also related to data cleaning in the data mining field where text files or databases are preprocessed to improve subsequent mining tasks by filtering irrelevant or useless information. But, these traditional techniques cannot be directly used to do Web page cleaning. However we should also note that current Web page cleaning methods still cannot perfectly clean

In this paper, various web page cleaning techniques, methods and approaches have been reviewed and discussed. Most of the cleaning techniques & methods work on general noise and some of them detect only specific noisy items of web pages. Not a single technique or method removes all types of noise patterns of web pages. These techniques & methods are site dependent and performance is based on similarities of web pages.

REFERENCES

[1] R. Kosala and H. Blockheel. Web Mining Research: A Survey. In SIGKDD Explorations, Volume 2, Number 1, pages 1-15, 2000.

[2] 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.

[3] 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

[4] Y. Yang, and H. J. Zhang, HTML page analysis based on visual cues, in Proceedings of the Sixth International Conference on Document Analysis and Recognition, pages 859– 864, Washington, DC, USA, 2001.

[5] S.H. Lin and J.M. Ho. Discovering informative content blocks from Web documents. In Proceeding of SIGKDD-2002, 2002

[6] Z. Bar-Yossef and S. Rajagopalan. Template Detection via Data Mining and its Applications, In Proceedings of the 11th Internation World-Wide Web Conference (WWW 2002), 2002.

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

[8] 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.

[9] 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.

[10] 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).

[11] S. Debnath, P. Mitra, and C. L. Giles. Automatic extraction of informative blocks from webpages. In ACM Symposium on Applied Computing, pages 1722–1726, 2005.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] V. Bharanipriya1 & V. Kamakshi Prasad2 WEB CONTENT MINING TOOLS: A COMPARATIVE STUDY International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215

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

[19] S. Paek and J. R. Smith, Detecting Image Purpose in World-Wide Web Documents, SPIE/IS&T Photonics West, Document Recognition, January, 1998.

[20] Gupta, S., Kaiser, G., Neistadt, D. and Grimm, P., DOMbased Content Extraction of HTML Documents, In the proceedings of the Twelfth World Wide Web conference(WWW 2003), Budapest, Hungary, May 2003.

[21] Cai Deng, Yu Shipeng, and Wen Jirong, et al. VIPS: a Vision Based Page Segmentation Algorithm[R] , Microsoft Technical Report:
(MSR-TR-2003-79) ,2003.

[22] A. K. Tripathy, A. K. Singh "An Efficient Method Of Eliminating Noisy Information In Web Pages for Data mining" in Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04) 0-7695-2216-5/04 © 2004 IEEE

[23] Thanda Htwe, Khin Haymar Saw Hla "Noise Removing from Web Pages Using Neural Network" in iccae2010 ,978-1-4244-5586-7/10/$26.00 C 2010 IEEE Volume 1.

[24] YongZhang, Ke Deng "Algorithm of Web Page Purification Based on Improved DOM and Statistical Learning" in 20 I 0 International Coriference On Computer Design And Appliations (ICCDA 2010)

[25] Guohua Hu, Qingshan Zhao " Study to Eliminating Noisy Information in Web Pages