International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 08, August-2022

Framework for Web Page Noise Removal for Effective Web Mining

S. S. Bhamare
School of Computer Sciences
Kavayitri Bahinabai Chaudhari North Maharashtra University
Jalgaon (M.S) India.

***Abstract*— In huge network of World Wide Web, information on web pages is considered as very useful information that information contains big amounts of noise or irrelevant data such as navigation bars, links, advertisements, copyright notices etc. It is more important to distinguish important or useful information from noisy or irrelevant content that may misguide the user's interest. Performance of Web mining task can be enhanced by identifying and removing noise from Web pages. The main objective of this research paper is to discuss the research work has been done in this area and propose an efficient framework for web page noise cleaning (WPNC) for effective web mining.**

***Keywords*— Noise, Web Page, Tag, HTML, Features, WPNC**

I. INTRODUCTION

Removal of Noise from Web pages is mostly related to web content mining which is the main branch of web mining. It offers an approach for eliminating noise from web pages for the purpose of improving the accuracy and efficiency of web content mining. The existing methods or techniques are not capable of removing all kinds of noise. Most of them have focused on detecting main content blocks in web pages. It is observed that noise free information retrieval from web pages is related to feature selection, feature weighing, block splitting, duplicate block elimination, noise block and finding important blocks in the web content mining field to improve subsequent mining tasks by filtering irrelevant or useless information.

Many researchers have proposed several methods and techniques for retrieval of main content from web pages. Most of the methods work on general noise and detect only specific noisy items from web pages. These methods require similar structure of web pages. The main objective of this research is to propose an efficient framework for web page noise cleaning which remove all kind of noises from all types of web pages.

II. LITERATURE SURVEY

Cleaning noisy page data is an important task and a researcher have worked in this area for retrieving and extracting main content and eliminates noisy data from different Web pages. 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.

[04] in this work how to find a model to automatically assign importance values to blocks in a web page is investigated, the block importance estimation is defined as a learning problem. First, the VIPS (Vision-based Page Segmentation) algorithm is used to partition a web page into semantic blocks with a hierarchy structure. A feature vector for each block is constructed using the spatial features and content features. Based on these features, learning algorithms like SVM and neural network are used to train diverse block importance models for differentiating noisy or unimportant blocks from pages.

[05] proposes a new tree structure, called Style Tree, and features an algorithm how to construct a site style tree. The Style Tree Model is used to detect and eliminate noises in any web pages of the site. An information based measure is used to determine which element node is noisy. Experimental results show that noise elimination technique is able to improve the mining result significantly.

[06] proposes an intrapage informative structure mining system called WISDOM (Web Intrapage Informative Structure Mining based on the Document Object Model) which applies information theory to DOM tree knowledge in order to build the structure. This system WISDOM splits a DOM tree into many small sub-trees and applies a top-down informative block searching algorithm to identify candidate informative blocks. The structure is built through expansion of the set employing proposed merging methods. In Experimental results high precision and recall rates which validates WISDOM's practical applicability.

[07] In this work, by taking the advantage of the HTML structure of web and n-gram technique for partial matching of strings, an n-gram based algorithm for mining web content outliers is proposed. To save time, the optimized algorithm uses only data captured in <Meta> and <Title> tags. Experimental results indicate that the proposed n-gram-based algorithm is capable of finding web content outliers. In addition, using texts captured in <Meta> and <Title> tags give the same results as using text embedded in <Meta>, <Title>, and <Body> tags.

[08] proposes an algorithm to extract the structure of a Web site using hyperlink analysis. It identifies and filters noise hyperlinks by patterns of Web pages that these hyperlinks have connected, instead of patterns of the hyperlinks. The initial results show that the proposed algorithm has a great improvement on both precision and recall ratio.

[09] proposes a novel approach to automatically extract main contents from web pages. Compared with existing studies, the method may determine whether a web page contains main contents, and then extracts the main contents without using DOM-Tree and template. Main contributions contain: (1) Introduction of a new concept of Block and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 08, August-2022

suggesting a method to partition web pages into blocks. Main contents and noise contents may be well partitioned into various blocks. (2) Bringing about a concept of Web Page Block Distribution and studying its features. Based on Block Distribution, it can be effectively determined whether the web page contains main contents, and then extract main contents via outlier analysis.

[10] in this proposed method, web pages are processed as images. And then, all the image features can be resiliently used as the standard for measuring analogy of noise blocks. As a result, noise blocks and information blocks can be distinguished after measuring similarity, and the reduction of noise is realized. The results of experiments indicate that this method gives accuracy and reliability and it can support joint measurement of multiple image features.

[12] proposed algorithm takes visual and non-visual characteristics of a web page into account and is able to remove noisy parts from three major categories of pages which contain user-generated content (News, Blogs, Discussions). Based on a manually generated corpus consisting of various topics, domains, and templates, it demonstrates the learning abilities of this algorithm examining its effectiveness in extracting information and its usage as a rule-based classifier for web page type detection in a realistic web setting.

[13] proposes a novel method to filter web pages and retrieve the actual content of a web page. This research work also proposes an approach for removing the noise from a given web page which will improve the performance of web content mining. At first, the web page information is divided into various blocks which are then tokenized to separate the informative content from noise. It presents an algorithm for removing noise from the web page and automatically extracts important web contents.

[14] proposes a system that removes boilerplate and extracts main contents. This has two phases: Feature Extraction Phase and Clustering Phase. The system classifies the noise or content from HTML web page. Content Extraction algorithm describes how to get high performance without parsing DOM trees.

[15] proposes a simple priority-assignment based approach with a view to differentiating main contents of the page from noise. This proposed technique, it first makes partition of the whole page into a number of disjoint blocks using HTML tag based technique. Next, the priority level for each block based on HTML tags priority is determined while considering aggregate priority calculation. This assignment process gives a priority value to each block which helps rank the overall search results in online searching. The blocks with higher priority are termed as informative blocks and preserved in database for future use, whereas lower priority blocks are considered as noisy blocks and are not used for further data searching operations.

III. PROPOSED FRAMEWORK FOR WEB PAGE NOISE CLEANING AND ITS WORKING

It is essential to remove noise information from web pages for improving the various web mining tasks. Noise removal from web pages is not easy task. Many of researcher have work on this topic but they have limitation in identification and removal of noise from web pages such as, mostly these methods are not performed on dynamic structures of web pages. Most of the techniques and methods worked on similar structure of web pages. To take an advantage of this problem, a new framework for identification and removal of web page noise is proposed as shown in Figure 3.1.

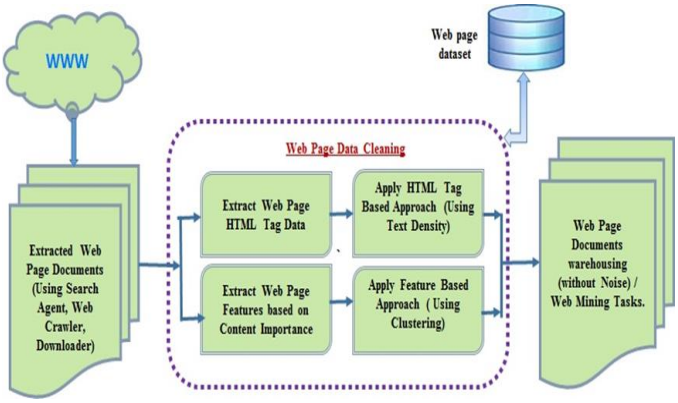

Figure: 3.1

This proposed framework as shown in Figure 3.1 consists of three major phases through which we can identify and remove noise from web pages.

In the first phase, retrieval of web pages from World Wide Web is used for the preparation of web page tag data set. In the second phase, this web page tag information and web page features are used as input to go through the web page noise identification and removal process. In the third phase, noise free web documents used for performance evaluation of system are obtained.

First Phase: - In the first phase of the system, a web page data set was prepared. For experimental results of this proposed system news domain set of different categories of Web Pages such as Sports, Technology and Main Pages from three news web sites, CNNIBN, ABB News and Times of India are used. These Web pages (or Web documents) are extracted through search engines by giving query. Here, web scrapping method is used to extract Web page HTML tag content information (or HTML source code) from each web page.

The web documents thus obtained are pre-processed. Web pages are constructed through their source codes, source codes are text files with an ".html" suffix and HTML commands / tags. Different types of HTML tags are used to design or construct any web page.

All the information in a web page is not of equal importance. Navigational bar, copyright notice, links; Advertisements etc. are considered as noise part of web pages and not important for web mining.

This framework consists of two proposed methods which include HTML tag based and Web page feature based methods. These proposed methods help to identify noise from web pages to remove for efficient web mining.

Second Phase: - In the second phase of the proposed framework, web page tag information and web page location features weights are used to identify and remove web page noise for effective mining operations. In web page tag

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 08, August-2022

information, actual data is enclosed within a pair of open and a close tag, a web page block is a portion of web page enclosed within an open-tag and its matching close-tag. Generally, HTML tags can be divided into two categories i.e., container or informative and description or decoration tags. Main content i.e., informative content is found in <BODY> tag and its corresponding tags. The important tasks are extracting the information of <BODY> and there corresponding tags. To prepare web page tag information data set, web scrapping method is used to extract this source tag information of each web page.

In web page feature based, web page has several features including information location, occupied area, and its contents. The position of contents and importance of contents perform a crucial role in identification of noise in web pages. Weights are assigned to these features according to their contents and location importance for separating noise and informative contents.

Third Phase: - Finally in third phase, noise free web page information is obtained for efficient web mining operations.

IV. FUTURE WORK

The Implementation and empirical evaluation part of these proposed framework of two methods i.e., HTML tag based and Web page feature based methods for identification and removal of noise from web pages is in progress and soon it is communicated for publication. This proposed method is implement using open source python library.

V. CONCLUSION

The existing methods and techniques of web page noise cleaning are analyzed and compared. These techniques and methods are site dependent, and performance is based on similarities of web pages. It is observed that the noisy blocks of a web page in a given web site usually share some common contents and/or presentation styles with other pages. While the main content blocks of the web page are often diverse in their actual contents and presentation styles. To improve the performance of web content mining operations, a new framework is proposed which identifies and removes noise from web pages. This framework consists of two proposed methods for noise removal. The main issue with the web pages is, they do not have a similar structure and the web page data is in heterogeneous form. The position of contents and importance of contents perform a important role in identification of noise in web pages. Web page noise identification and removal is the preprocessing task of web mining. Proposed framework of methods for web page noise cleaning are capable of improving the results.

REFERENCES

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

[2] N. Kushmerick. Learning to remove Internet advertisements, AGENT-99, 1999.

[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] Song. Liu, Wen, Ma,Learning Block Importance Models for Web Pages, WWW 2004, May 17-22, 2004, New York, NY USA. ACM.

[5] Zhao Cheng-Ii, Yi Dong-Yun, A Method of Eliminating Noises in Web Pages by Style Tree Model and Its Applications, Wuhan University Journal of Natural Sciences Vol.9 No.5 2004.

[6] Hung-Yu Kao, Jan-Ming Ho, Ming-Syan Chen WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 5, May 2005

[7] Malik Agyemang, Ken Barker, Rada S. Alhajj , Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams ACM Symposium on Applied Computing-2005

[8] Feng Li, Extracting Structure of Web Site Based on Hyperlink Analysis, 978-1-4244-2108-4/08/$25.00 © 2008 IEEE

[9] JiangtaoQiu, , Changjie Tang, Kaikuo Xu, and Qian Luo, Web Contents Extracting for Web-Based Learning , ICWL 2008, LNCS 5145, pp. 59–68, 2008.© Springer

[10] Haitao Yao, Zhiyi Yin, Fuxi Zhu, Changsheng Gong, The Noise Reduction Method of Web Pages Based On Image Features, 978-1-4244-4507-3/09/$25.00 ©2009 IEEE.

[11] Hu Fei, Yang Huaqian,, Wei Pengcheng, Pu Changjiu, Lei Yang, Web Page Noise Reduction Algorithm Using Non-template Approach, International Journal of Digital Content Technology and its Applications(JDCTA) Volume6,Number20,pp 556-561,November 2012.

[12] Nikolaos Pappas, Georgios Katsimpras, EfstathiosStamatatos, Extracting Informative Textual Parts from Web Pages Containing User-Generated Content, I-Know'12, September 05-07, 2012, Graz, Austria Copyright 2012 ACM 978-1-4503-1242-4/12/09.

[13] SurabhiLingwal, Noise Reduction and Content Retrieval from Web Pages, International Journal of Computer Applications (0975 – 8887) Volume 73– No.4, pp 24-30, July 2013.

[14] Pan Ei San, Boilerplate Removal And Content Extraction From Dynamic Web Pages, International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.4, No.6, December 2014.

[15] RaselKabir, ShailyKabir, and Shamiul Amin, Isolating Informative Blocks From Large Web Pages Using Html Tag Priority Assignment Based Approach, Electrical & Computer Engineering: An International Journal (ECIJ) Volume 4, Number 3, September 2015.

[16] 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.