Web Data Cleaner : Relevant data filtering tool using Web Data Extraction

Shraddha Borate, Chaitrali Kulkarni, Akshata Waghmode

[1] *BE, Computer Department, International Institute of Information Technology, Maharashtra, India*
[2] *BE, Computer Department, International Institute of Information Technology, Maharashtra, India*
[3] *BE, Computer Department, International Institute of Information Technology, Maharashtra, India*

ABSTRACT

Nowadays internet usage is vast. It contain large information in the form of text, audio, video etc. This information should be necessary or unnecessary, hence removing irrelevant information from web is known as noise. The noise can be in the form of Ads, Texts, Images etc., information degrades performance of web content mining. Web content mining is used for discovering the useful information from the web page. It's important to eliminate the noisy data from web pages. This paper tells about the new approach called ENDW (Eliminating Noisy Data in Web pages) which is based on query keyword and Dom tools to eliminate the noisy data from the web pages. Thus this approach will be helpful to the user for getting the required information from the web pages, which is reliable and efficient.

Keyword: - *Web mining, WDE (Web Data Extraction), Noise Elimination, Web Extraction, Filtering, DOM tree.*

1. INTRODUCTION

As we know that in today's day-to-day life usage of internet is wide, everyone is familiar to internet. On the web page there are various type of noise data like images, videos, advertisement, animated images, this noisy data disturbs whole attention of user while searching any information, filling important form. Therefore we need to eliminate this noisy data so that user gets the required data. To eliminate this noisy data there are many existing approaches of web mining which, has disadvantage that it extracts the data only from single URL.Web Mining is the process of extracting the required data from web pages using crawler .Web crawler is nothing but an automated program which scans through internet pages to create an index of the data. We proposed the new approach to extract the required information from three different URLs. Web page is designed by using various HTML tags and these tags are of two different types as follows:

- Positive tags: These tags contain useful information. i.e. ***
- Negative tags: These tags do not contain any information. i.e. <script>**</script>

In our system we maintain a pointer to scan the whole document, document is nothing but source code of the web page and it eliminates this unwanted tags which does not contain any type of information.

2. EXITING SYSTEM

Existing system technique is based on analysis of layouts as well as the actual contents of the web page in a given website for eliminating noisy information. Initially, tag based filtering method based on regular expression is applied. But filtering does not ensure removing all the noisy information present in the web page. To remove remaining noisy information we need to make structural analysis of the web page along with some crawled web pages of the same website. This noise removal technique based on the analysis that in a website the noisy blocks usually presence with some similar contents and layout instruction, whereas the main content blocks often varied with their actual contents and presentation style. Based on the observation, the analysis of both contents and structure had been done in structural analysis phase. The Existing method had been organized in two stages: Filtering based on Regular Expression and Structural analysis of the crawled web pages after filtering. The structural analysis had been accomplished in two steps: a) extraction of body level tags from filtered web pages and b) comparison of extracted body level tags among all the filtered web pages.

3. PROPOSED SYSTEM

The main advantage of proposed system over the existing system, that it is Multi-URL System.
This software extracts the web page's source code from backend when user enters the appropriate URL. Then basic filtering is applied on that source code to eliminate the negative tags. Positive tags contain all useful information therefore Dom tree will be generated of this useful information which contains nodes and links. Nodes contain HTML tags or text. Rule based filtering is applied on this generated Dom tree [8], which gives us required information. By fetching Query user can get relevant data. Query is nothing but a keyword of which user wants the information [1]. Relevant data is displayed on the dummy web site. This Multi-URL System allows the user to enter at most three URL, therefore it makes easy for user to get the data from three different websites.

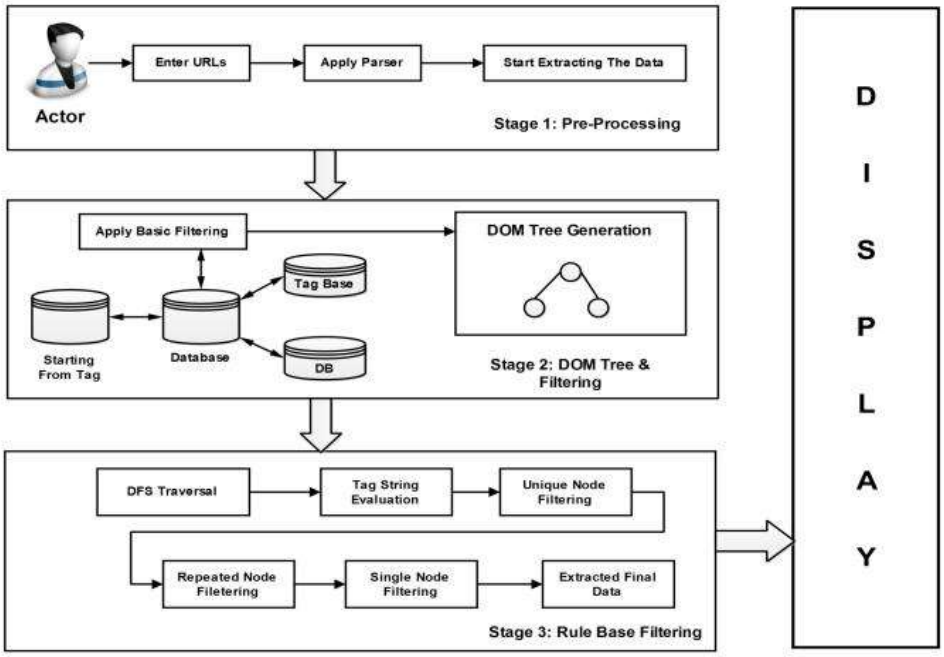

Fig -1: Architecture Diagram

The proposed method has been organized into following stages:

- **Pre-Processing**: User will enter the URL and the web pages will be extracted from backend.
- **Dom tree and Filtering**: In this stage, basic filtering will be applied and DOM tree generation will be done.
- **Rule base Filtering**: In this stage, DFS traversal is carried out in which every node is traversed and then single node, unique node, repeated node will be done and therefore when will get the required data.
- **Result**: After applying filtering on DOM tree generation it will display result on the dummy website. That result is nothing but required data.

4. ALGORITHM

4.1 Filtering

Input: web page
Output: return filtered web page
Step 1: Get HTML code of the web page
Step 2: Create a pattern using regular expression for removing contents enclosed by negative tags (except anchor tag)
Step 3: Pattern P = {p0: "<style.*?>.*? </style>",

p1: "<script.*?>.*? </script>",
p2: "<select.*?>.*? </select>",
p3: "<noscript>.*? </noscript>",
p4: "<!--.*? -->",
p5: "<link.*?>",
p6: "
",
p7: "<hr>"}
Step 4: for each Pattern pi in P do
if pi matches with the HTML code of the web
page then remove the code block
end if
end for
Step 5: Create a Pattern using regular expression for removing contents enclosed by noisy anchor tags
Step 6: Pattern anchor= ".*? "
Step 7: extract all the hyperlink references using pattern
anchor from HTML code of the web page
Step 8: Set T➔ URL of the web page
Step 9: for all extracted hyperlink references hi do
if returned shift s = 0 then remove the content enclosed by <a /> containing hi
end if
end for

4.2 Extraction

Input: Filtered web pages
Output: Relevant Information
Step 1: Generate DOM tree for html code of web
pages
Step 2: Apply DFS traversal on DOM tree
Step 3: Start from root node, move towards left node and goto child node
Step 4: if web page contains bk =database tags dk then remove entire node
else
{
if unique node then (Compare with siblings)
add as relevant data
else
remove
end if

if duplicated data k in node p (Repeated data)
remove duplicated data
else
consider relevant data
end if

if single node
then remove data in single node end if
}
end if
Step 5: Repeat step from step 2 until you traverse all the nodes.
Step 6: for each body level tags xk in a web page k
for each body level tags xk+1 in a web page k +1 remove data
end for
end for

5. EXPERIMENTAL RESULT

Input: Website with noisy data.

Fig -2: Original Website with Noise

Output: Website with noise free data

Fig -3: Website without Noise

6. PERFORMANCE ANALYSIS

Table -1: Accuracy Table

Websites	Total Words	Noisy Words	Noise Removed	Percentage of Accuracy
Horoscope(Gemini)	6750	5000	4500	90%
Horoscope(Taurus)	5000	4000	3500	87%
Career Guidance	2000	1500	1475	98%
Biography	9300	7000	4700	67%
Naukri	1800	1500	1300	86%

New ABC	8000	5000	3700	74%
New XYZ	7000	6000	5800	97%
Yahoo	12500	10000	8000	80%
Sport	18000	15000	9200	61%
Film Industry	37000	7000	6500	93%

7. CONCLUSION

This paper is proposed to detect and remove local noisy elements from web pages. Web page content extraction is more vital to retrieve the content of the web pages, particularly in unstructured web. The proposed technique uses the DOM tree parsing to remove the noise and irrelevant information. The system will extract the content dynamically from the different structured web pages such as blogs, forums, articles etc. By using filtering techniques as well as Dom tree generation it removes the noisy data from the web pages. This helps the user to read the information efficiently and reliably from web pages without any disturbance**.**

8. FUTURE SCOPE

It is not always possible that the relevant data will be in the form of text only, but it can also be in the forms of videos and images. Therefore it is possible that software can be developed which gives the relevant videos and images along with the text. Software that will give the noise free data from n number of websites can also be developed in future. This Software "Web Data Cleaner" can also be worked on Android Phones by building its Android Application.

9. ACKNOWLEDGEMENT

It gives us great pleasure in presenting the project on `Web Data Cleaner: Relevant data filtering tool using Web Data Extraction'. We would like to take this opportunity to thanks Head of Computer Engineering Department for giving us all the help and guidance needed, also really grateful to them for their kind support and valuable suggestions which were very helpful.

10. REFERENCES

[1] Ying-Kui Wang, Qian-Mao Tan, "A Query Keywords Based Approach for Noisy Data Elimination" Second International Conference on Business Computing and Global Informatization, 2012 IEEE.

[2] Dr.Anna Saro Vijendran, C Deepa, "LBDA: A novel framework for extracting content from web pages" International Conference on Advanced Computing and Communication Systems (*ICACCS* -2013), Dec. 19 – 21, 2013, Coimbatore, INDIA. 2013 IEEE.

[3] P. Sampatht, C. Ramesht, T. Kalaiyarasit, S. Sumaiya Banut, G. Arul Selvan2, "An Efficient Weighted Rule Mining for Web Logs Using Systolic Tree" IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012.

[4] Andrés Sanoja, Stéphane Gançarski, "Block-o-Matic: A Web Page Segmentation Framework" 2014 IEEE.

[5] B.Aysha Banu.,M.E.,(Ph.D).,Dr.M.Chitra.,M.E.,Ph.D.,"A Novel Ensemble Vision Based Deep Web Data Extraction Technique for Web Mining Applications" 2012 IEEE Intemational Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

[6] Rajni Sharma, Max Bhatia, "Eliminating the Noise from Web Pages using Page Replacement Algorithm", Rajni Sharma et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3066-3068.

[7] Neetu Narwal, "Improving web data extraction by noise removal", Asst. Prof., Maharaja Surajmal Institute, Affiliate College of GGSIP University, Research Scholar, Lingayas University.

[8] Bhavdeep Mehta, Meera Narvekar, "DOM Tree Based Approach for Web Content Extraction",2015 International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17, Mumbai, India.

[9] Dingkui Yang, Jihua Song,"Web Content Information Extraction Approach Based on Removing Noise and Content-Features",2010 International Conference on Web Information Systems and Mining.

[10] Wigrai Thanadechteemapat, Chun Che Fung,"Improving Webpage Content Extraction By Extending A Novel Single Page Extraction Approach: A Case Study With Thai Websites", Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012.

[11] Haitao YAO, Zhiyi YIN, Fuxi ZHU, Changsheng GONG,"The Noise Reduction Method of Web Pages Based On Image Features",School of Computer Wuhan University Wuhan, China 2009 IEEE.

[12] Gupta S,Kaiser G, Neistadt D, "DOM-based content extraction of HTML Documents". In: Proceedings of the 12th International Conference on
WWW. Budapest, Hungary 2003.

[13] Gupta S, Kaiser G, Stolfo S. "Extracting context to improve accuracy for
HTML content extraction". In: Pro ceedings of WWW'05. New York, NY, USA, 2005: 1114-1115.

[14] Yi L, Liu B, Li X. "Eliminating noisy information in web pages for data mining". In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2003: 296-305.

[15] Shine N. Das, Pramod K. Vijayaraghavan, Midhun Mathew – "Eliminating Noisy Information in Web Pages using featured DOM tree". In: International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA.

[16] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Extracting content structure for web pages based on visual representation," in Web Technologies and Applications, ed: Springer, 2003, pp. 406-417.

[17] A. H. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira, "A brief survey of web data extraction tools", ACM Sigmod Record, vol. 31, pp. 84-93, 2002.

[18] J. Wang and F. H. Lochovsky, "Data extraction and label assignment for web databases", in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 187-196.

[19] L. Fu, Y. Meng, Y. Xia, and H. Yu, "Web content extraction based on webpage layout analysis", in Information Technology and Computer Science (ITCS), 2010 Second International Conference on, 2010, pp.