Mohinder Singh, Navjot Kaur / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 4, Jul-Aug 2013, pp.2671-2675

Retrieve Information Using Improved Document Object Model Parser Tree Algorithm

Mohinder Singh*, Navjot Kaur**
*(Department of Computer Science and Engg., SGGSW University, Fatehgarh Sahib.
** (Department of Computer Science and Engg. , Punjabi University, Patiala

ABSTRACT

The Data mining refers to mining the useful information from raw data or unstructured data. Whereas in web content mining the data is scattered or unstructured on web pages. Some time the user wants to retrieve only fix kind of data, but the unwanted data is also retrieved. The unnecessary information can be removed with this proposed work. The DOM Parser Tree Algorithm to filter the web pages from unwanted data and give the reliable output. The Document Object Model Parser Tree Algorithm fetches the HTML links. According to these Links the pages are accessed. Then the data with is useful for user, is send to the table. The DOM Parser Tree Algorithm works upon tree structure and we have used the table for output the results. As the results are shown in the table, the information displayed in the table is correct and reliable for the user. The user fixes the data which he/she wants to access time by time. The data dynamically fetched from that particular website or link. Currently the approach is implemented on limited field of experiment because of some limits of privileges. Hopefully the approach will be implemented on large experimental area.

Keywords – Data mining, DOM Parser Tree Algorithm, Web Content mining,

I. INTRODUCTION

In today's world the whole system is totally depended on computer system. The data stored in system can be of various types like text, video, audio, rich document and so on. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is used in almost every filed like business, army, finance, art etc. In the proposed work, the information from different websites is clustered in the table format. This is very useful for an individual business organization. We used the document Object Model that will allow programs and scripts to dynamically access and update the content, structure and style of documents. The document can be further processed and the results of that processing can be incorporated back into the presented page. HTML links are fetched by this application dynamically through official websites of related product. With the help of this algorithm, individual organization can retrieve their updated data; this process is pretty handy to analysis. The user specifies work related to his required information to the system. The web crawler takes its seed URL and searches for the relevant pages. Then DOM tree is generated of those pages. Now the irrelevant contents like advertisements, link lists to related articles, disclaimer information, user comments, navigational menus, headers, footers, copyright notices, and privacy policies, are removed using DOM parser algorithm which uses the DOM tree structure. Then the table from the extracted content is build using Document Object Model Tree.The flow from web crawler to DOM tree implementation plays a crucial part in the proposed work. Web crawlers download web pages by starting from one or more seed URLs, downloading each of the associated pages, extracting the hyperlink URLs contained therein, and recursively downloading those pages. Therefore, any web crawler needs to keep track both of the URLs that are to be downloaded, as well as those that have already been downloaded.

II. LITERATURE REVIEW

Dr. M S Shashidhara2, Dr. M. Giri ***[1]*** proposed new technique of web data extraction. It has three phases. In the first phase list of web documents are selected, second phase documents are pre-processed, in the final phase results are presented to users. Experimental results are compared with existing methods. Performance of proposed system is better than existing methods. Web mining is a class of data mining. Web Mining is a variation of this field that distils untapped source of abundantly available free textual information. The importance of web mining is growing along with the massive volumes of data generated in web day-to-day life. **Jing Li and C.I. Ezeife** ***[12]*** Classifying and mining noise-free web pages will improve on accuracy of search results as well as search speed, and may benefit webpage organization applications. Noise on web pages is irrelevant to the main content on the web pages being mined, and includes advertisements, navigation bar, and copyright notices. The few existing work on web page cleaning detect noise blocks with exact matching contents but are weak at detecting near duplicate blocks, characterized by items like navigation bars. They proposed, Webpage Cleaner, for eliminating noise blocks from web pages

for purposes of improving the accuracy and efficiency of web content mining. A vision-based technique is employed for extracting blocks from web pages. Important blocks are exported to be used for web content mining using Naive Bayes text classification. **Kamlesh Patidar(2011)*[2]*** Website plays a significant role in success of an e-business, e-books & Knowledge Discovery. It is the main start point of any organization and corporation for its customers, so it's important to customize and design it according to the visitors. Also, websites are a place to introduce services of an organization and highlight new service to the visitors and audiences. They proposed a prototype design in future as a search engine they proposed a algorithm web content mining using the database approach and multilevel Data tracking for digital library. **Qing Lu *[16]*** the link-based classification, unlabeled data provides useful information is provided in three important ways: first, it gives us additional information about the distribution of object attribute values second, links among unlabeled data in the test set provide useful information about classification and third, links between labeled (training) data and unlabeled (test) data also provide useful information that should not be ignored. When the classification problem is properly modeled, and we don't distort the data by removing links between the test and training and inference is used for collective classification, we are able to make use of all of the information that unlabeled data provides **G. N. Shinde *[3]*:** the researchers represent an efficient method for software development. They proposed MDA approach.MDA is a promising approach for software development. MDA using J2EE (Java to Enterprise Edition) is used to describe behavior of agents. JADE (Java Agent Development Environment) Framework provides a standard for developing MAS (multi-agent systems). Web Usage Mining, as well as Web Mining, is a new research field, which has a long way to go. For the Web-based data warehouse and data mining technology, the development of the Internet provides a broad application scope. With the rapid development of Internet, communications technology, the research of Web based data mining will be further in-depth and Web site design and so on. Also for Agent's Modeling Language (AUML) is being defined to effective implementation of defining agent roles **Shekhar Palta *[11]***: they proposed some areas where the unimportant words can be eliminated. The future work of this project involves a lot of improvements that can be done to improve the accuracy of the extracted text. The extracted text right now also includes the name of the author and the presence of dates and the likewise text that occurs as part of the news article and it becomes very difficult for the algorithm to prune out these phrases from the extracted text. So, a solution we have thought of is to use the concept of a suffix array to detect the regular and repeated occurrence of certain unimportant words over the set of extracted news article from the same web site. This has already been implemented by a colleague at Ask and we are working to integrate it with the present algorithm **R.Cooley *[13]*** has proposed some aspects of mining the information and pattern discovery.The term Web mining has been used to refer to techniques that encompass a broad range of issues. However while meaningful and attractive this very broadness has caused Web Mining to mean different things to different people and there is a need to develop a common vocabulary. Towards this goal they proposed a definition of Web mining and developed taxonomy of the various ongoing efforts related to it next we presented a survey of the research in this area and concentrated on Web usage mining. They provided a detailed survey of the efforts in this area even though the survey is short because of the area newness. They provided a general architecture of a system to do Web usage mining and identified the issues and problems in this area that require further research and development. **Niki R.Kapadia *[4]***: They proposed an approach for extracting web content structure based on visual representation. The resulted web content structure is very helpful for applications such as web adaptation, information retrieval and information extraction. By identifying the logic relationship of web content based on visual layout information, web content structure can effectively represent the semantic structure of the web page. An automatic top-down, tag-tree independent and scalable algorithm to detect web content structure is presented. They compared it with traditional DOM based algorithm as they got much more reliable partitioning. The algorithm is evaluated manually on a large data set, and also used for selecting good expansion terms in a pseudo-relevance feedback process in web information retrieval, both of which achieve very satisfactory performance. Recently, the developed algorithm is implemented on website with horizontal separation. The further work would be focused on development of improved code in order to take care of vertical separation also. Further, the various experiments of developed code will be carried out on different domain website such like education, shopping, social networking etc. to check accuracy of the code and to judge which consideration will give better results. This work will be useful in understanding VIPS algorithm, web content mining through partition based segmentation and further research directions in this area to computer engineering fraternity **Niki R.Kapadia *[5]*:** concludes that there are many existing methods to mine information. Hierarchical and partitioning methods are used commonly to mine information from the web. Each method has advantages and disadvantages. **G.Poonkujhali *[15]***: proposed new algo. using signed approach for improving the results

of web content mining by detecting both relevant and irrelevant web documents. They aimed at experimental evaluation of web content mining in terms of reliability and to explore other mathematical tools for mining the web content. Also, a comparative study of this algorithm with existing algorithms is to be done.

III. METHODOLOGY

In the proposed work we have used the document Object Model Parser Algorithm for extracting the useful data for specific user. In Figure 1 . we have shown the methodology of the proposed algorithm with the help of flow chart.

Input: we used selected URLs as an input

Output: Useful information in tables and execution time

Method:

- First read all the links one by one from the database using sql query
- Then extract the every link by using java http library
- Apply the DOM Parser Algorithm for extracting the useful data
- Links are fetched by the table
- Finally output in the tabular form In the methodology flowchart of DOM, our work is on 3rd step i.e. where DOM Parser Algorithm is applied.

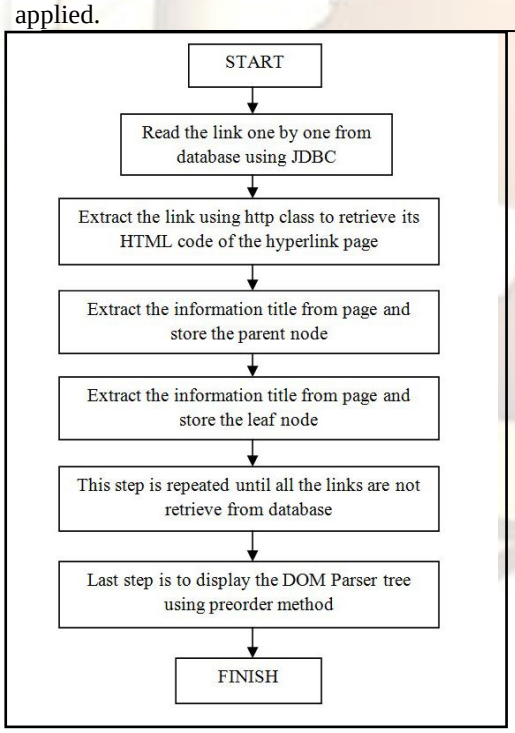

Figure 1: Flow chart for Proposed System

COMPARISON BETWEEN PREVIOUS WORK AND PROPOSED WORK: In the previous work the FP-Growth algorithm is used. The approach uses the log files for obtaining the frequent pattern. The major problem in this algorithm was that it fetches the frequent pattern. Some time unwanted data can also be frequent in particular website, in that case the algorithm not work as efficient as it should be.

In out proposed Document Object Model Parser Tree Algorithm, it works upon the data which is selected by the user single time. After selecting the required website, the user will be able to get the updated data or information. This algorithm works on dynamic pages or links, on which the information changes time by time.

DOM PARSER TREE ALGORITHM

Step 0: START

Step 1: First we add links as Right Add and Left Add

```
Data leftAdd;
String[] sdata = new String[10];
int portnum = 0;
Data rightAdd;
```

Step 2: Declare Class DST

```
int v = 0;
Data Root = null, headNext = null;
```

Step 3: Declare Function DataSet

```
for (int i = 0; i < 6; i++)
node.sdata[i] = s1[i];
node.portnum = num;
```

Step 4: Now Check Root

```
if (Root == null)
node.leftAdd = null;
node.rightAdd = null;
Root = node;
return Root;
```

Step 5: Read HTML from source links

```
URLConnection urlConnection;
DataOutputStream outStream;
DataInputStream inStream;
URL url;
```

Step 6: Connect the Database and Links

```
b1.addActionListener(this);
b2.addActionListener(this);
b3.addActionListener(this);

url = new URL(s);
urlConnection = url.openConnection();
((HttpURLConnection) urlConnection).setRequestMethod( "POST");
```

Step 7: END

IV. RESULT AND DISCUSSION

We run the project; it shows results as well as time taken by algorithm. We used three phase's cars, bikes and mobiles by using tables as output of our algorithm. All the unwanted data is cleaned by DOM Parser Tree Algorithm. The output is shown as below. In Figure 2 it displays the results of Car's module. Text box represents the time taken by the algorithm to execute the whole process. Same as Figure 3 and 4 respectively represents the Bike's and Mobile's module. All the links are dynamically fetched by the algorithm. Any kind of updating, modification, altering etc. made in the related website can be accessed with the help of proposed algorithm.

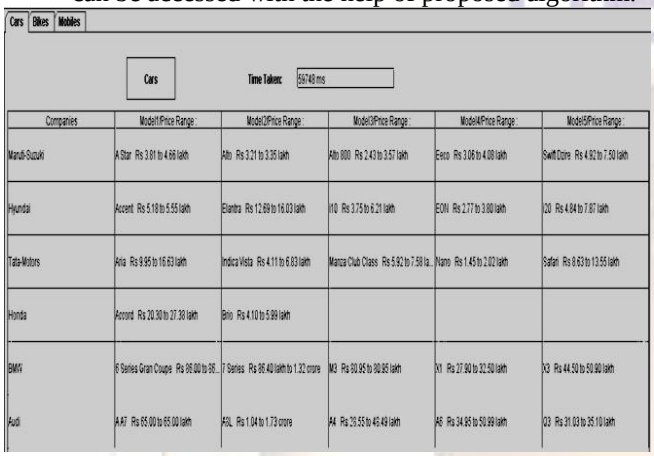

Figure 2: Car's Experimental Result

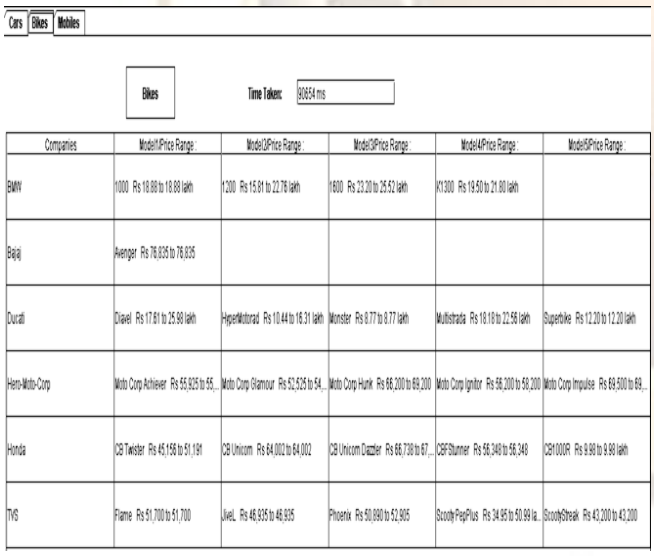

Figure 3: Bike's Experimental Result

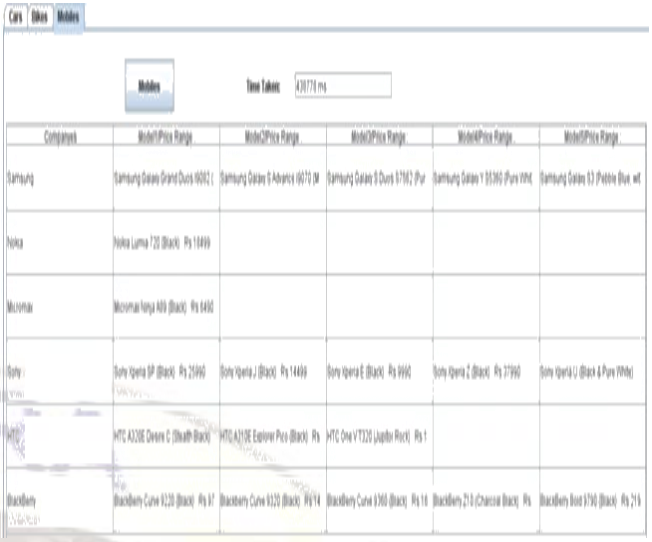

Figure 4: Mobile's Experimental Result

The graph represents the throughput of algorithm in mili-seconds. In our proposed work there are 3 attributes. These results are dynamic, because the speed and quality of internet effects on throughput of algorithm. So the result sets vary to the conditions and system performance.

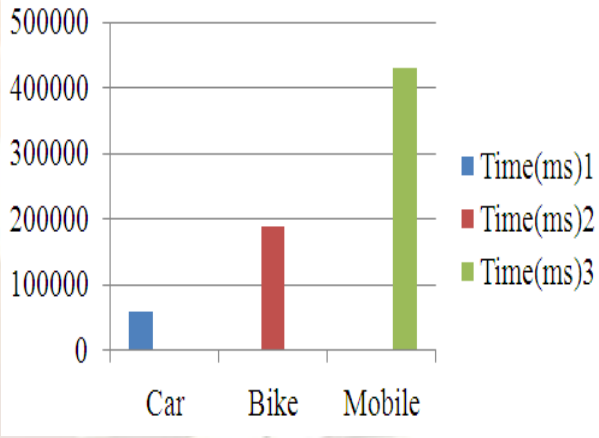

Figure 4.2: Result Based Graph

This is the major advantage of this algorithm that the data doesn't loss its reliability and correctness. Any individual organization can retrieve the updated data time by time. Another advantage of this work is that the information is stored in tabular form, as it is much simple and easy to understand.

V. CONCLUSION

We present the methodology of Document Object Model Parser Tree Algorithm works upon data content mining and data clustering. As we represented, the algorithm removes the unnecessary data like add-ons, pop-ups, unwanted material, photographs etc. With the help of this algorithm, any kind of individual organization can easily access its useful data time by time dynamically. This algorithm can be used as base for large scaled project for a

company or firm. It is pretty useful especially where there is a rich media data. It clusters the data which is related to particular organization and isolates it from the data which is not useful for that specific task.

VI. FUTURE WORK

- Test the proposed algorithm can work on large scaled data and privileges or need to improve factors.
- It can help to improve the efficiency of related algorithms.
- The proposed algorithm works on high speed internet. When we run it on slow speed network then its results are not as accurate as these should be. So, in future work can be done on the efficiency of algorithm to work on slow speed networks to give the better results.

Acknowledgements

As a part of my course I have taken the problem as **"Retrieve Information using Improved Document Object Model Parser Tree Algorithm"** as my Thesis Topic. I am very thankful to Mrs. Navjot Kaur, Assistant Professor, Punjabi University, Patiala for giving me such a valuable support in doing my work. She provided all the relevant material that was sufficient for me to complete my thesis work. She provided help and time whenever asked for. Last but not least, a word of thanks for the authors of all those books and papers which I have consulted during my thesis work as well as for preparing the report. At the end thanks to the Almighty for not letting me down at the time of crisis and showing me the silver lining in the dark clouds.

REFERENCES

Journal Papers:

[1] Dr. M S Shashidhara2, Dr. M. Giri, An Efficient Web Content Extraction Using Mining Techniques, *International Journal of Computer Science and Management Research Vol 1 Issue 4 November, 2012.*

[2] Kamlesh Patidar, Preetesh Purohit, Kapil Sharma,Web Content Mining Using Database Approach and Multilevel Data Tracking Methodology for Digital Library, *IJCST Vol. 2, Issue 1, March 2011.*

[3] G. N. Shinde,Inamdar S.A,Web Data Mining Using An Intelligent Information System Design, *Int. J. Comp. Tech. Appl., Vol 2 (2), 280-283*

[4] Niki R.Kapadia, Kanu Patel, Mehul C.Parikh, Partitioning Based Web Content Mining, *International Journal of Engineering Research & Technology* (IJERT), ISSN: 2278-0181, Vol. 1 Issue 3, May-2012.

[5] Niki R.Kapadia, Kinjal Patel, Web Content Mining Techniques– A Comprehensive Survey, *IJREAS*, Volume 2, Issue 2(February 2012), ISSN: 2249-3905.

Books:

[6] Jiawei Han, Micheline Kamber,"*Data Mining Concepts and Techniques"2nd Edition,*

[7] Jiawei Han, Micheline Kamber,"*Data Mining Concepts and Techniques"2nd Edition,*

[8] Jiawei Han, Micheline Kamber,"*Data Mining Concepts and Techniques"2nd Edition,*

[9] Jonathan Robie, Texcel Research*,"What is the Document Object Model?"*

[10] Jonathan Robie, Texcel Research*,"What is the Document Object Model?"*

[11] Shekhar Palta, Eliminating Noisy Information from News Websites and Extraction of the News article, *International Master on Information Technology IV Edition, Pisa, Italy*.

Thesis:

[12] 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.*

[13] R.Cooley, B.mobasher, J.Srivastva, *Web mining:* Information and Pattern Discovery on the World Wide Web, *Department of Computer Science and Engineering, university of Minnesota, MN 55455, USA*

Proceedings Papers:

[14] Huiping Peng, Discovery of Interesting Association Rules Based On Web Usage Mining, *IEEE Coference, pp.272-275, 2010.*

[15] G.Poonkujhali, K.Thiagarajan, K.Sarukesi, g.V.Uma,Signed Approach for Mining Web Content Outliers,*World Academy of Science and Technology 32 2009*

[16] Qing Lu, Lise Getoor, Link-based Classification using Labeled to Unlabeled Data, *Proceedings of the ICML-2003, Washington DC,2003*