ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

Enhancement in DOM Tree to Reduce Time and Complexity – A Proposal

Punam Bajaj [1], Payal Joshi [2], Anchal Garg [3]

Assistant Professor, Department of Computer Engineering, Chandigarh Engineering College, Landran, Mohali, India[1]

Assistant Professor, Department of Computer Engineering, Chandigarh Engineering College, Landran, Mohali, India[2]

P.G. Student, Department of Computer Engineering, Chandigarh Engineering College, Landran, Mohali, India[3]

Abstract: Data mining is the process of mining information from the large set of data. A Web Page contains many blocks such as content blocks. Other than content blocks, there are such blocks like copyright, privacy notices and advertisements. These blocks don't come under main content blocks, but these are known as noisy blocks or noisy information. Eliminating these noises will improve web data mining. In this paper, we will discuss how to identify these noises to improve efficiency of web mining. And also removal of noises using simple LRU algorithm. Least Recent Used algorithm is less time consuming and less complex algorithm for web mining.

Keywords: Content Extraction, DOM Tree, LRU, Web Mining.

I. INTRODUCTION

Data Mining is define as extracting the information from the large set of data. It can also define as data mining is mining the information from data [1]. In the field of Information technology, it has enormous amount of data available that require being bitter into useful information. This information further can be used for various applications like market analysis, customer retention, production control, fraud detection, science exploration etc [2]. Web Content Mining is the process of extracting useful information from the contents of Web. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities in these fields are also involved using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision the application of these techniques to Web content mining has not been very rapid. Web Usage Mining is the application of data mining method to discover interesting usage patterns from Web data in order to understand and better serve the requires of Web-based applications [5]. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. With the explosive growth of information on World Wide Web, it becomes difficult to identify the correct or relevant information because there are many distracting features available around the actual content of web pages. Useful information is surrounded by noises such as banners, privacy notices, advertisements etc, these noises effects web pages performance and efficiency. Web page noises can be categorized into two parts-

Global Noises: These are noises which are no smaller than a single page itself which includes mirror sites, duplicated web pages etc.

Local Noises: Noises within the web pages are local which includes banners, navigational panels, links, advertisements etc.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

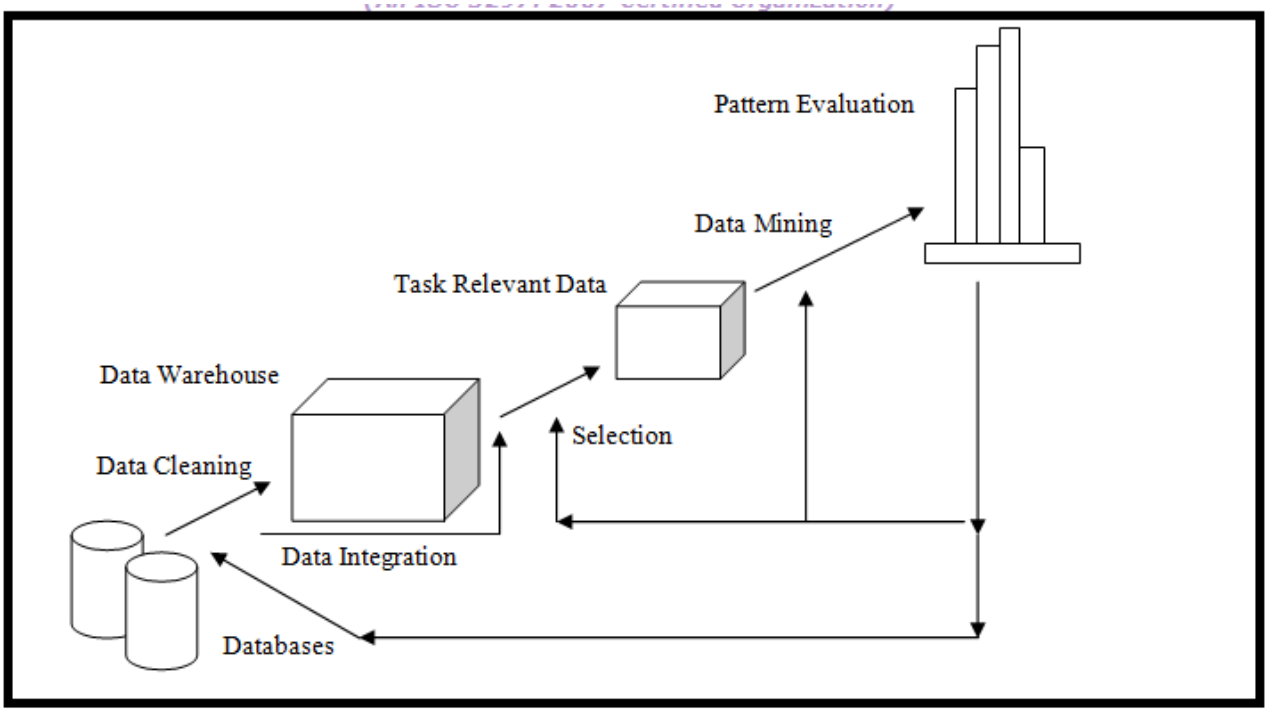

Fig.1. The Base Concept

In this work, we will work on detecting and eliminating noises from web pages to improve the performance of web mining which is the application of data mining techniques to extract knowledge from Web data including Web documents, hyperlinks between documents, usage logs of web sites, etc.

II. LITERATURE REVIEW

In paper **Jinbeom Kang [3]** proposed a new tech technique of Web page segmentation by recognizing repetitive tag patterns which is called key patterns in the DOM tree structure of a page. They report that on the Repetition-based Page Segmentation (REPS) algorithm which identify key patterns in a page and create virtual nodes to correctly segment nested blocks. A number of experiments done for real Web sites showed that REPS greatly contributes to improving the correctness of Web page segmentation. The REPS algorithm analyses key patterns in a page and creates virtual nodes to segment nested block.

K.Rajkumar [5] a new method of segmentation is introduced (DWS) which segments web pages based on either reappearance based technique by analyzing reappearance tag patterns from the DOM tree structure of a web page. Based on the analysis of tag patterns it gave implicit nodes to segment the nested block correctly nor it will segment pages based on web layout data like <TABLE>, <DIV> and <FRAME> tags based on key pattern in the web page. If it consist of reappearance tag in tag pattern means it will segment based on reappearance based segmentation. Else it will segment based on web layout data. From that segmented block hyperlink is displayed on the mobile first and after that user select hyperlinks based on his area of interest. The interested data information alone is displayed to the user. Based on the detection of tag patterns it build implicit nodes to segment the nested block correctly. From that segmented block hyperlink is displayed on the mobile device first and then user select hyperlinks based on his area of interest. In

In this paper **Chaw Su Win, Mie Mie Su Thwin [1]** proposed Effective Visual Block Extractor (EVBE) Algorithm to overcome the problems of DOM-based method and reduce the drawbacks of previous works in Web Page Segmentation. It also proposed Effective Informative Content Extractor (EIFCE) Algorithm to minimize the drawbacks

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

of previous works in Web Informative Content Extraction. Web Page Indexing System Clustering System and Web Page Classification, Web Information Extraction System can achieve significant saving and satisfactory results by applying the Proposed Algorithms.

Jan Zelený [4] provides an overview of distinct approach which can be used for finding a relevant content on the web page. Each technique has its advantages and disadvantages and their usage should be considered according to a particular task which required to be solved. Many of presented algorithms were originally targeted at a analysis of content on news servers. But if they consider how modern web pages are designed the same method can be applied to blogs, CMS-based sites and also most of company web pages.

In this paper **K.S.Kuppusamy [7]** proposed a model for micromanaging the tracking activities by fine-tuning the mining from the page level to the segment level. The proposed system enables the web-master to find the segments which receives more focus from users comparing with others. The segment level analytics of user actions offers an important metric to analyse the factors which facilitate the increase in traffic for the page. The empirical validation of the model is performed through prototype implementation.

In this **Deng Cai [2]** presents an automatic top-down tag-tree independent method to detect web content structure. It simulated how a user understands web layout structure based on his visual perception. Comparing to other existing method their approach is independent to underlying documentation representation like HTML and works well even when the HTML structure is far different from layout structure. Experiments showed satisfactory results.

III. RELATED WORK

There is a much related work in information extraction and noise removal that resolves similar problems using various techniques. Web Content Mining is the process of extracting useful information from the contents of Web. Content data corresponds to the group of facts, a Web page was designed to convey to the users. It may contain of audio, text, images, video, or structured records such as lists and tables. Text mining and its application has been the most extensible researched. Research activities in these fields are also involved using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision the application of these techniques to Web content mining has not been very rapid. Web Usage Mining is the application of data mining method to discover interesting usage patterns from Web data in order to understand and better serve the requires of Web-based applications. Usage data confines the individuality or source of Web users along with their browsing behaviour at a Web site. Web usage mining itself can be categorized further depending on the kind of usage data considered. Informative Content Extraction is the process of finding the parts of a web page which contain the main textual content of this document. Several methods have been explored to extract information from web pages using vision based and common layout template. But there is less work on detecting as well as removing noises from web pages.

Content Extraction Techniques

Content extraction systems try to extract useful information from structured or semi structured documents. A tree structure is used to see presentation style of the page.

Web Page Segmentation Techniques

There are many approaches used to segment web pages into regions and blocks. One of them is DOM (Document Object Model) based segmentation approach, in this scheme an HTML document is showed as a tree. Document object model specification builds XML and HTML documents into tree like structure. Another approach is based on layout of web pages. The drawback is that layout based approaches doesn't fit for all pages.

IV. PROPOSED WORK

A web page typically contains many information blocks. Besides the content blocks, these also contain noisy blocks. These noisy blocks can seriously harm the web content mining. In DOM based segmentation approach, it splits the

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2014

HTML document into tree structure. But this approach has many drawbacks like the HTML parser is quite slow. Another problem is that it takes a lot time to build DOM tree structure which effects performance. Now, we will require an efficient algorithm which will be overcome all these problems. To overcome this problem a new algorithm is used in page replacement that is Least Recent Page Replacement Algorithm. A good approximation to the best possible algorithm is based on the observation that pages that has been greatly used in the last. On the other hand, pages that have not been used for long time will probably remain unused for a long time. This idea suggests a practicable algorithm. When a page fault occurs, throw out the page that has been unused for the longest time. This strategy is called LRU paging. Removal of noises will be done with Least Recent Used algorithm. This algorithm is a page replacement algorithm in which the least recently used pages are removed from the web page. In this work, we use the concept of web usage. The web usage mining will tell us the recent used advertisement link. On the basis of this information LRU algorithm will work.

The complexity is that the listing must be updated on every memory reference. Finding a page in the list, deleting it, and then moving it to the front is a very time consuming operation, even in hardware also. There are another ways to implement LRU. Let us consider the way first which is very simple. This method requires equipping the hardware with a 64-bit counter, C, that is automatically incremented after each instruction. In addition, each page table entry must also have a field large sufficient to include the counter. After each memory reference, the current value of C is stored in the page table entry for the page just referenced. When a page fault occurs, the operating system examines all the counters in the page table to find the lowest one. That page is the least recently used.

V. EXPERIMENTAL RESULTS

In this section we will evaluate the performance of proposed algorithm. Our main purpose is to eliminate noise and to increase the efficiency of web pages. We perform a classification technique to evaluate the algorithm. We will check data set before and after removal of noises. We show that how noises badly effect the web content mining.

VI. CONCLUSION

The main objective of this research paper is to discuss various algorithms of web mining. We also focused on to discuss various advantages and disadvantages of the same. We believe that algorithms discussed in this paper will gave benefit for various research scholars. This paper helps in detecting and removing noises from web pages. This proposed algorithm aims to improve performance based on a new technique, LRU. By the use of this algorithm, we find the least used links which are affecting the performance of web pages. We evaluate our algorithm which leads us to improved results.

REFERENCES

[1] Chaw Su Win, Mie Mie Su Thwin, "*Informative Content Extraction By Using Eifce"* International Journal Of Scientific & Technology Research Volume 2, Issue6, 2013

[2] Deng Cai, "*VIPS: a Vision-based Page Segmentation Algorithm"* Microsoft Research Microsoft Corporation One Microsoft Way Redmond, WA 98052, 2003

[3] Jinbeom Kang, Jaeyoung Yang, Nonmemberand Joongmin Choi, " *Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices"* , IEEE Transactions on Consumer Electronics, Vol. 56, No. 2., 2010

[4] Jan Zelený, "Web Page Segmentation And Classification" Journal of Data and Knowledge Engineering, 2010

[5] K.Rajkumar, "Dynamic Web Page Segmentation Based on Detecting Reappearance and Layout of Tag Patterns for Small Screen Devices", 2011

[6] Kahkashan Tabassum, "*A Heuristic-based Cache Replacement Policy for Data Caching",* IJCST, Vo l. 1, Issue 2, 2010

[7] K.S.Kuppusamy, "*A Model for Web Page Usage Mining Based on Segmentation",* International Journal of Computer Science and Information Technologies, Vol. 2 (3), 2011