Web Page Performance Enhancement by Removing Noise

Anchal Garg
M.Tech CSE

Bikrampal Kaur
Ph.D

ABSTRACT
Data mining is the procedure of extracting or taking out the information from the huge set of data. Web Mining is an important application of data mining, which is to extract knowledge from Web data including Web documents, hyperlinks, usage logs of web sites, etc. A Web Page contains many blocks such as content blocks, copyrights, privacy notes and advertisements. These blocks like advertisements and copyrights etc. don't come under main content blocks. These blocks are known as noisy blocks or it can be said that these blocks contain noisy information. This noisy information adversely effects web data mining. Eliminating this noisy information will improve web data mining. In this paper, it will be discussed how to identify these noises and how to eliminate them to improve efficiency of web mining. There are many types of algorithms which are used in web mining i.e. Visitor method, Dom Tree. Visitor and Dom Tree both are complex and time consuming methods. We will also discuss removal of noises by using simple LRU algorithm and variants of LRU, which will result into less time consuming algorithm for web mining.

General Terms
Algorithm, Performance

Keywords
Content Extraction, DOM Tree, LRU, Web Mining

1. INTRODUCTION
Data Mining is defined as extorting or taking out the information from the large deposit of data or records. It can also define as mining the information from data [1]. In the field of Information technology, it has massive quantity of data available that required to be transformed into valuable information. This information can be used for various applications like market analysis, production control, fraud detection, science exploration etc [2]. There are different kind of algorithms and techniques available for different types of data to convert into useful or valuable information. Different techniques are like web mining, web content mining, text mining etc. Data mining is studied for different databases like object-relational databases, relational database, data ware houses and multimedia databases etc. The information extraction procedure in data mining consists of some steps from raw data collection to valuable information; which is shown in Figure 1. Data is collected from various sources, then that data is cleansed. In data warehouse, data from various sources is integrated into common source. Then the applicable data is selected to begin the process. Web Content Mining is the process of extracting useful information from the contents of Web. Many techniques from other disciplines are also used in research such as Information Retrieval and Natural Language Processing (NLP). Web Usage Mining is the application of data mining which is used to discover interesting usage patterns from Web data in order to understand and better serve the requires of Web-based applications [5]. The captured data or patterns then help to identify the origin of Web users along with their browsing behavior at a Web site. But a web user sometimes or we can say naturally ignores some parts of the web page which contains additional non- informative contents or which are not of the interest. This also makes it tough to discover main content of document. With the rapid expansion of information on World Wide Web, it becomes a popular place to extract information but also it is really 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, advertisements etc, these noises effects web pages performance and efficiency. So, data mining becomes an interesting feature for discovering valuable information.

In this work, we will focus on to removing the advertisements from the 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, usage logs of web sites, etc.

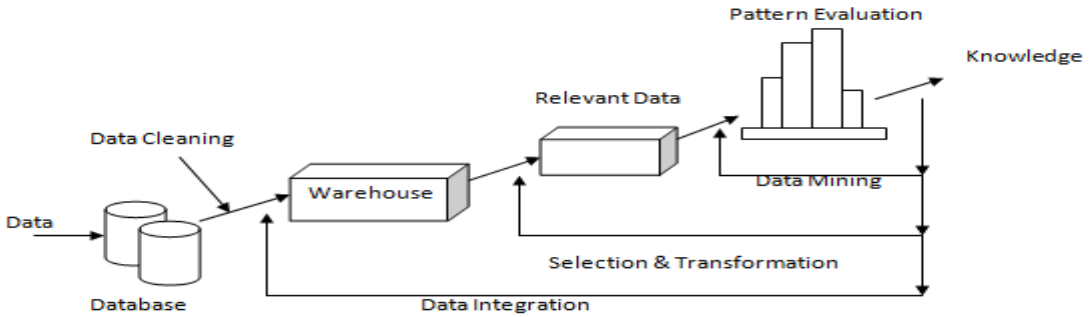

Figure 1: The Base Concept

The remainder of the paper proceeds as follows. We will do literature survey in section 2; section 3 will present related work. Section 4 will gives our approach. Section 5 will show experimental results done to evaluate the effectiveness of our proposed method. And finally, in section 6 we provide conclusions.

2. LITERATURE REVIEW

In this research work **Chaw Su Win, Mie Mie Su Thwin [1],** proposed Effective Visual Block Extractor (EVBE) Algorithm to overcome the problems of DOM-based method and reduced the drawbacks of previous works in Web Page Segmentation. It also proposed Effective Informative Content Extractor (EIFCE) Algorithm to minimize the drawbacks of previous works in Web Informative Content Extraction. Web Page Indexing System, Clustering System and Web Information Extraction System can achieve significant saving and acceptable results by applying the Proposed Algorithms.

In this research work, **Jinbeom Kang [3]** proposed a new technique of Web page segmentation by recognizing repetitive tag patterns which is called key patterns in the DOM tree structure. They reported 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. This idea mainly comes from the observation that Web designers usually build a Web page with structural and repetitive layouts. A number of experiments done for real Web sites showed that REPS greatly contributes to improving the correctness of Web page segmentation..

In this research work **Deng Cai [2],** presented an automatic top-down tag-tree independent method to detect web content structure. It simulated how a user understands web layout structure based on 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.

K.Rajkumar [5] proposed a new method of segmentation (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 area of interest. This paper proposed Dynamic web page segmentation for mobile device. Previous methods for Web page segmentation are not flexible in a dynamic Web environment because they largely relied on heuristic rules.

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 analyze the factors which facilitate the increase in traffic for the page. The empirical validation of the model is performed through prototype implementation.

Jan Zelený [4] provided 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.

Swe Swe Nyein [10], proposed the algorithm based on Content Structure Tree (CST). Firstly the proposed system use HTML Parser to build DOM (Document Object Model) tree from which create Content Structure Tree (CST) which can easily extract the main content blocks from the other blocks. The proposed model then introduced cosine similarity measure to evaluate which parts of the CST tree represent the less important and which parts showed the more important of the page. The proposed system can define the ranking of the documents using similarity values and also extracts the top ranked documents as more relevant to the query. Web page typically contained many information blocks. Apart from the main content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements which are called noisy blocks. These noisy blocks can seriously harm Web data mining.

Thanda Htwe [11], an application of Neural Networks is presented for pattern classification combined with DOM structure to extract content information. Feed forward Neural Network is used to implement the system which used the back propagation learning algorithm. Data is collected from various web sites is used in training and testing. The classification result of back propagation neural network is used for eliminating various noise patterns from Web page. To evaluate proposed system, an experiment is performed on several Web pages. Experiments indicated that method is applicable to extract informative content from Web pages.

3. RELATED WORK

A significant amount of work has been done in information extraction and noise removal that resolves similar problems using different techniques. Web mining can be classified into various categories. Web Content Mining is a process of extracting useful information from the contents of Web. Content data may contain of audio, text, images, video, or structured records such as lists and tables. Text mining and its application has been the most widely researched. Research activities in these fields are also involved using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). Web Structure Mining can be regarded as the process of discovering structure information from the Web. 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 behavior at a Web site. Web usage mining can be categorized further depending on the kind of usage data considered. Several methods have been searched to extract information from web pages using vision based and common layout template. But there is less work done on detecting as well as removing noises from web pages.

Informative Content Extraction is the process of finding the parts of a web page which contain the main textual content of this document. 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.

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. The text-based methods differ from the other two in that they do not at all take the tree structure of the HTML into account. They only look at the text content and analyze certain textual features like e.g. the text-density or the link-density of parts of a page.

4. OUR APPROACH

A web page typically contains many information blocks and also non- informative blocks, which are also called noisy blocks. These noisy blocks can seriously harm the web content mining. In our proposed method, comparison of two different algorithms will be done. In DOM based approach, it splits the HTML document into tree structure. But this approach has many drawbacks like the HTML parser is quite slow and 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.

A new algorithm is used in page replacement that is Least Recent Page Replacement Algorithm. A study has been done on the basics of LRU; some pages have been used many times in the last. On the other hand, some pages have not been used for long time and that pages will probably remain unused for a long time. This idea suggests a feasible algorithm. When a page fault occurs, throw out that pages which has been unused for the longest time. This strategy is called LRU paging. Removal of noisy information or advertisements 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. The concept of web usage will work efficiently here. Web usage mining discovers the usage patterns from the web data. The web usage mining will let us know the recent used advertisement link. On the basis of this information LRU algorithm will work.

To implement LRU algorithm, a linked list should be maintained in memory, with the most recently used page at the front and the least recently used page at the rear. The listing must be updated on every memory reference. But it is time consuming as to move pages to front after finding them in the list. There are another ways also to implement LRU algorithm. Let us consider a method which requires equipping the hardware with a 64-bit counter, C, which is automatically incremented after each instruction. In addition, each page table entry must also have a sufficient field to include the counter. After each memory reference, the current value of C is stored in the page table entry. When a page fault occurs, the operating system checks all the counters in the page table to find the lowest one. That lowest one page is the least recently used.

Variants of LRU are also applied in the proposed method. Different variants of LRU are First in First Out method (FIFO), Optimal Method, LRU-k method.

5. EXPERIMENTAL RESULTS

In this section we will evaluate the performance of proposed algorithms. Our main purpose is to eliminate noise and to increase the efficiency of web pages. We will check the complexity of both algorithms i.e. DOM tree and LRU algorithm. A comparison is done between both algorithms; we work on removing the non- required advertisement. LRU algorithm will work faster than the DOM tree. LRU will give us the least used advertisement. This works by maintaining a linked list in the memory, with the most recently used page at the front and the least recently used page at the rear. Complexity of LRU algorithm is less than the DOM tree algorithm, which makes LRU algorithm better.

The experimental results in the form of graphs are shown below.

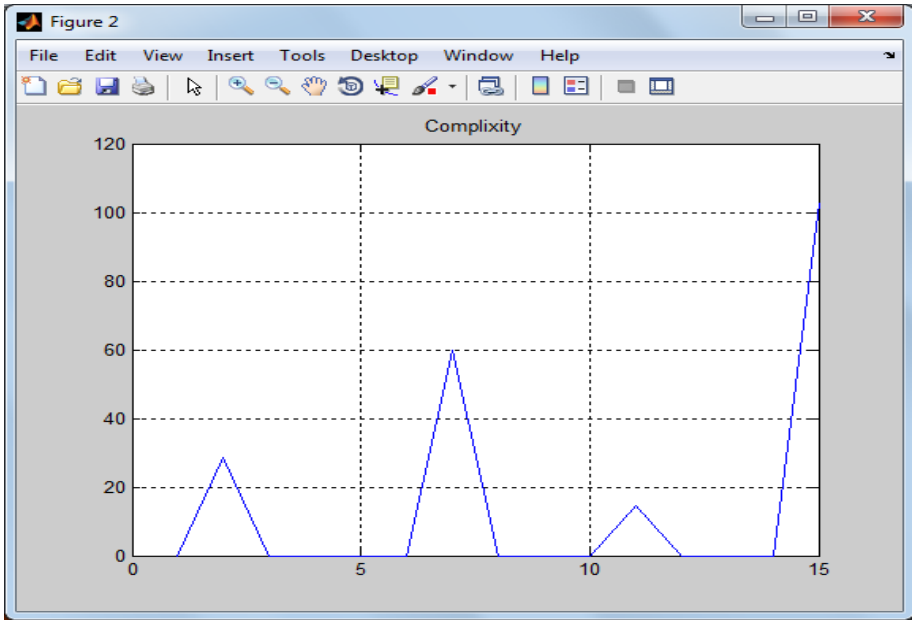

Figure 2: Complexity Using DOM Tree

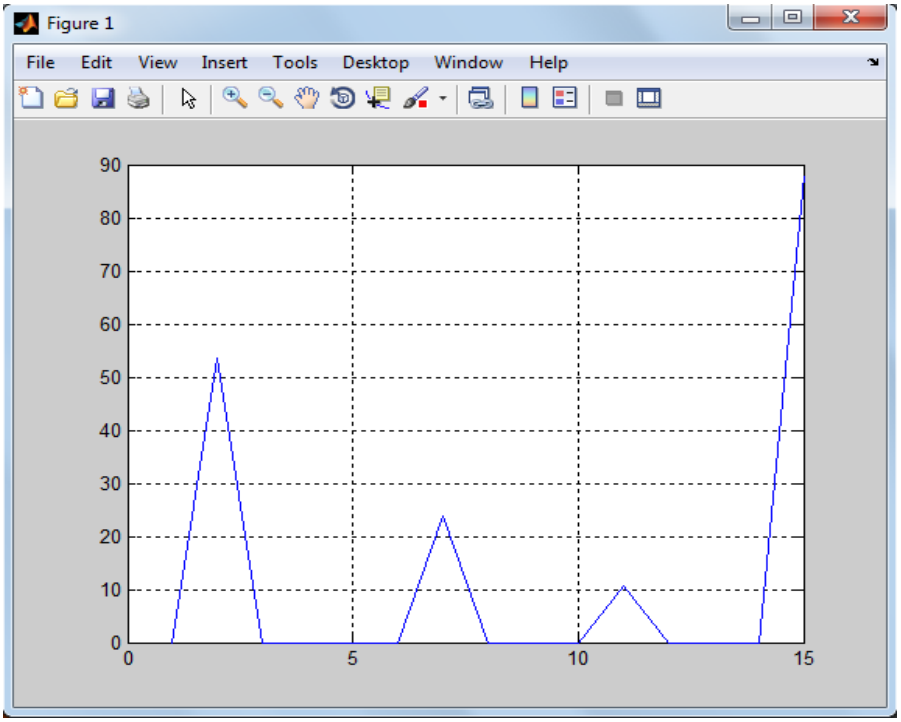

Figure 3: Complexity Using LRU

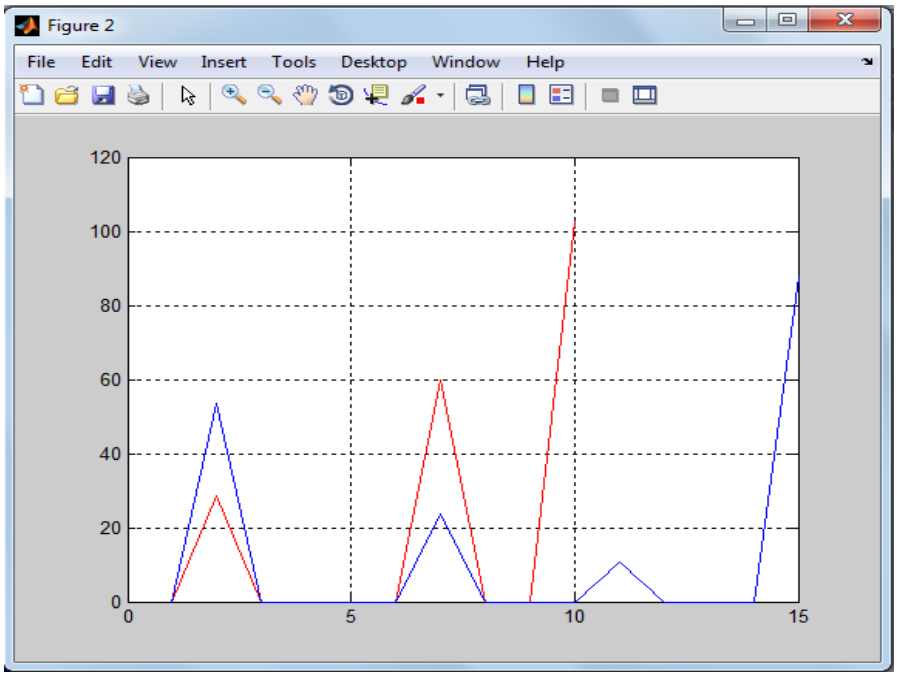

Figure 4: Comparison of Both Algorithms

In Figure 2; complexity of removing advertisement by DOM algorithm is given. In Figure 3; complexity of removing same advertisement by LRU algorithm is given. In Figures 4, comparison of both algorithms is shown in the form of line graph.

6. CONCLUSION

The main objective of this research paper is to discuss various algorithms of web mining. We also focused on discussing various advantages and disadvantages of the algorithms. Algorithms discussed in this paper will give benefit to 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, Least Recently Used and its variants. 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.

7. ACKNOWLEDGEMENT

I am highly grateful to the Principal, Chandigarh Engineering College (CEC), Landran, for providing this opportunity to carry out the present thesis/ work. I would like to express a deep sense of gratitude and thanks to The Head of Department and my mentor Dr. Bikrampal Kaur.

8. REFERENCES

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

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

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

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

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

[6] Kahkashan Tabassum (2010)"A Heuristic-based Cache Replacement Policy forData Caching" IJCSTVo l. 1,Issue 2

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

[8] Gibson D, Punera K, Tomkins A(2005)." The volume and evolution of web page templates" In: Proceedings of WWW'05. New York, NY, USA, 2005: 830-839.

[9] Lei F, Yao M, Hao Y.(2009) "Improve the performance of the webpage content extraction using webpage segmentation algorithm". In: Proceedings of International Forum on Computer Science-Technology and Applications. Chongqing, China, 323-325.

[10] Swe Swe Nyein (2011) "Mining Contents in Web Page Using Cosine Similarity ".

[11] Thanda Htwe, Nan Saing Moon Khan (2011), "Extracting Data Region in Web Page by Removing Noise using DOM Tree and Neural Network" 3rd international Conference on Information and Financial Engineering, IACSIT press, Singapore.

[12] Shuang Lin, Jie Chen, Zhendong Niu(2012) ."Combining a Segmentation-Like Approach And A Density-Based Approach In Content Extraction" TSINGHUA SCIENCE AND Technologyissnll1007-0214ll05/18llpp256-264 Volume 17.