ISSN:0975-9646

Eliminating the Noise from Web Pages using Page Replacement Algorithm

Rajni Sharma, Max Bhatia

Department of Computer Science and Engineering
Lovely Professional University, Phagwara

Abstract- Data mining is the process of mining information from the large set of data. It further has many categories like text mining web usage mining and web content mining. There are many types of algorithm which are used in web mining i.e. Visitor method, Dom tree and least recent used algorithm. Visitor and Dom tree is the complex and time consuming method. Least Recent Used algorithm is less time consuming and less complex algorithm for web mining.

***Keywords*: Web Mining, DOM Tree and LRU**

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 documents. Content data corresponds to the collection 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 [4]. 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 behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered [6]. Informative Content Extraction is the process of finding the parts of a web page which contain the main textual content of this document [7]. A human user nearly naturally performs some kind of Informative Content Extraction when reading a web page by ignoring the parts with additional non-informative contents, such as navigation, functional and design elements or commercial banners at least as long as they are not of interest [8]. Though it is a relatively intuitive task for a human user, it turns out to be difficult to determine the main content of a document in an automatic way. 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. These approaches are grounded in results from quantitative linguistics which indicate that statistically text blocks with similar features are likely to belong together and can thus be fused together [9]. The optimal similarity threshold depends on the wanted granularity and needs to be determined experimentally. In 2nd section we will do literature survey and in 3rd section detail of least recently used algorithm will be discuss.

II. REVIEW OF LITERATURE

In paper **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 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 analyzes key patterns in a page and creates virtual nodes to segment nested block. In this paper **Swe Swe Nyein[11]** an algorithm is proposed that extract the main content from the web documents. 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. In paper **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 this paper [10] **Shuang Lin et.al (2012)** they proposed a content extraction approach that join a segmentation-like scheme and a density-based scheme. In their approach they designed structures called BLE&IE blocks to gather related contents or noises. Next they used this density-based technique and redundancy removal to obtain the final content. Based on their scheme a tool called Block Extractor was developed. Experimental results on their data set showed that this novel technique was effective and robust compared to three other density-based method. Density-based method in content extraction whose task was to extract contents from Web pages are commonly used to obtain page contents that are critical to many Web mining applications. But traditional density-based method cannot effectively manage pages that consist of short contents and long noises. To overcome this problem they proposed a content extraction approach. In this paper **Yan Gu [13]** gives a simple but effective method named ECON to fully-automatically extract content from Web news page. ECON uses a DOM tree to show the Web news page and leverages the substantial features of the DOM tree. ECON finds a snippet-node by which a part of the content of news is wrapped firstly and then backtracks from the snippet-node until a summary-node is found and the whole content of news is wrapped by the summary-node. During the process of backtracking ECON removes noise. Experimental results showed that ECON can achieve high accuracy and fully satisfy the needs for scalable extraction. Moreover ECON can be applied to Web news page written in many popular languages such as Japanese, Portuguese, Russian, Chinese, English, French, German, Italian, Spanish, Arabic. ECON can be implemented much easily. 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 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. In this paper **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.

III. LEAST RECENTLY USED PAGING ALGORITHM

DOM tree has some disadvantages like it has high complexity and time consuming process. To overcome this problem a new algorithm is used in page replacement that is least recently used (LRU) [12]. A good approximation to the best possible algorithm is based on the surveillance that pages that has been greatly used in the last. A small number of instructions will most likely be a lot used all over again in the next the minority. 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 [13].To fully apply LRU, it is essential to keep a linked list of all pages in memory, with the most recently used page at the front and the least recently used page at the rear.

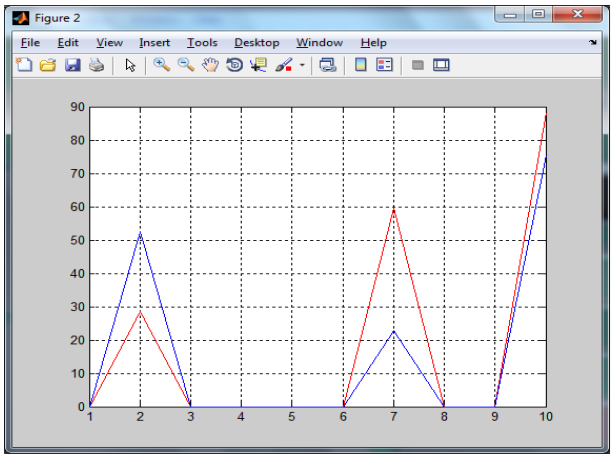
The complexity is that the listing must be updated on every memory reference. Finding a page in the list, deleting it, and then moving it to the front is a very time consuming operation, even in hardware also [9]. There are further many ways to implement LRU with special hardware. Let us consider the 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 [10]. 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.

IV. PROPOSED METHODOLOGY

The web pages are managed in the certain fixed manner. The users are required to extract the useful information from the web page. The user's useful information is the useful data for the users and other information is noisy one. The user extracts the useful information from the web page on the basis of web page template. Data mining on the Web thus becomes an important task for discovering useful knowledge or information from the Web. However, useful information on the Web is often accompanied by a large amount of noise such as banner advertisements, navigation bars, copyright notices, etc. Although such information items are functionally useful for human viewers and necessary for the

Web site owners, they often hamper automated information gathering and Web data mining, e.g., Web page clustering, classification, information retrieval and information extraction. Web noises can be grouped into two categories according to their granularities: The user extracts the data from the web page using the content page holder tag. The concept of DOM trees has been used to extract the useful information. The DOM tree technique comes under the technique of web segmentation. In the DOM-based segmentation approach, an HTML document is represented as a DOM tree. Another intuitive way of page segmentation is based on the layout of webpage. In this way, a web page is generally separated into 5 regions: top, down, left, right and center. Another application of block importance is on web page classification. The main problem in the HTML parser is that HTML parser is quite slow. The second problem in DOM trees is that it requires a lot of time to construct the DOM trees which will reduce the efficiency. In this thesis, we work on to remove the noisy data from the web page .To enhance the efficiency and to increase the response novel algorithm will be proposed.

RESULTS



Graph1 Advertisement Scheme Interface

As illustrated in the graph 1, DOM tree and LRU complexity is been shown and graph is plotted to compare the results graphically. Red line shows Dom tree complexity and Blue line shows LRU complexity. It indicates that DOM tree is more complex than LRU.

CONCLUSION

The main objective of this research paper is to discuss various algorithms of web mining. We also focused LRU algorithm advantages and disadvantages of the same. We believe that algorithms discussed in this paper will gave benefit for various research scholars. Its experiential result shows that LRU is better than DOM tree.

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"* IJCST Vol. 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] 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.

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

[12] T. Gottron, (2009) *"Evaluating Content Extraction on HTML Documents,"* Institut für Informatik, Johannes Gutenberg-Universität, Mainz, Germany.

[13] Yan Gu (2010), *"ECON: An Approach to Extract Content from Web News Page"* 12th International Asia-Pacific Web Conference.