Research on Advertisement Replacement in Webpage using DOM and LRU

Sheenam Garg[1], Prof. Hiteishi Diwanji[2]
[1, 2] Department of Information Technology
[1, 2] L.D.College of Engineering, Ahmedabad, Gujarat, India

Abstract- *Web Mining is used to capture relevant data about consumer, individual user and several others. In web pages only some information is useful in real world applications. Web Page has some additional contents like hyperlinks, header footer, navigational panel, advertisements that may cause the content extraction to be complex. These additional contents in Web Pages are treated as noisy contents. Advertisements that may not be of interest for user are also noise elements. We have proposed a technique which extracts the advertisements from the webpage and replaces them with the new advertisements. The DOM Based page Segmentation is used to extract the noisy and informative content block. Blocks containing Urls are extracted from the webpage. Then using the regular expressions urls having domain name different from the webpage and those Urls having the same domain name as webpage but redirecting to another website will be considered as advertisements. After parsing the webpage, based on the number of user clicks and the position of the advertisements Least Recently Used Advertisement is extracted and replaced from the new advertisement and the webpage with the replaced advertisement is displayed. This provides advertisements of user interest and in turn, results in increase of E-commerce transactions ending in large profits.*

Keywords- Content Extraction, DOM Tree Generation, Repetition tags, Statistical relation, Pattern Tree, Multilevel Pages, Redundant Data, Noisy data

I. INTRODUCTION

Web Mining is a Data Mining technique to automatically discover and extract information from World Wide Web. Web Mining is used to capture relevant data about consumer, individual user and several others. The contents of Web pages are the primary focus of Web mining applications [1].Web Mining decomposed into Resource Discovery, Information Selection & Pre-processing, Generalization and Analysis. We can classify web mining in 3 types according to its mining techniques that is web structure mining, web content mining, web usage mining.

A user is mainly interested in the original content of web page so, the process of identifying and fetching main content blocks from a web page is called content extraction. The term content extraction was found by Rahman[2]. The content extraction is very useful for pre-processing the data in many fields such as web mining, recommendation system, decision making, expert system, knowledge discovery and so on. It is also useful to special tasks such as false advertisement detection, demand forecasting, and comment extraction on product reviews [3]. The DOM Based page Segmentation is used to discard the noisy content block and extract the informative content block from Web Pages. Initially a XML or HTML Web Page is converted into DOM tree and noise is removed using DOM Based Page Segmentation which converts the page into blocks and regions. Performance of Web Content extraction is analysed based on complexity and efficiency of the method. For content extraction firstly DOM tree is generated. HTML attributes, Tag pattern generation, Subject detection, Node density, Visual information, text density etc. are used for precise content extraction and removing noisy data. In this survey paper we are discussing above techniques in detail.

II. DOM TREE GENERATION

Document Object Model (DOM) [4] is a standardized, platform-independent and language-independent interface for accessing and updating content, structure and style of any web documents. We can generate DOM tree for each HTML page where tags are internal nodes and the detailed text and images are leaf nodes. For example,

```
<HTML>
<HEAD>
<TITLE> text </TITLE>
</HEAD>
<BODY>
<P> p text</P>
<IMG SRC= “1.jpg”></IMG>
</BODY>
</HTML>
```

Dom tree for above HTML code is given below:

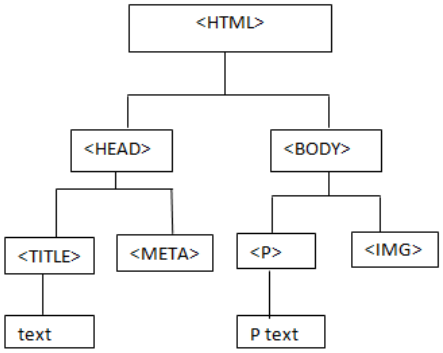


Figure 1 DOM tree

The growth of web pages on internet continues and the web Page organization is very essential. The Web Pages can be categorised into Navigation page and content page. A DOM based block text identification method proposed which detects the Navigation Page. This approach used to extracting the text segment block from a Web Page.

III. LITERATURE SURVEY

There are various techniques used for content extraction and noise removal. Each method has different percentage of content extraction and noise removal. According to the type of website i.e. government website, multilevel website, shopping website different content extraction techniques are applied for efficient and precise non-redundant content extraction and noise removal.

1. Effectual Web Content Mining using Noise Removal from Web Pages [5]

In this technique the following noises are removed step by step: (1) Primary noises such as Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and other Uninteresting Data such as audio, video, multiple links. (2) Redundant Contents and (3) Noise Contents having low block importance. These noises are removed by performing three operations. First, using the Block Splitting operation, primary noises are removed and only the useful text contents are partitioned into blocks. Second, using Simhash algorithm, the duplicate blocks are removed to obtain the distinct blocks. For each block, three parameters namely Keyword Redundancy (KR), Linkword Percentage (LP) and Titleword Relevancy (TR) are calculated. Using these three parameters block importance value (BI) is calculated. Based on a threshold value the important blocks are selected using sketching algorithm and the keywords are extracted from those important blocks.

2. Improving Web Data Extraction By Noise Removal [6]

It is generally observed that Web Pages of a single web site often follow similar layout pattern, and the noise elements are repeated in almost all web pages. In this technique, an algorithm is developed to extract the Visual Blocks of a Web page [7] of a web site using DOM and Visual Characteristics, and then it is converted to the Pattern Tree. The Pattern Tree of different web pages of a single web site is mapped to find the similarity pattern among the web pages of the website. For each node the Node Importance Measure is calculated, which is used to discriminate noise and main element of the web page.

3. Content Extraction Based on Statistic and Position Relationship Between Title and Content [7]

In previous technique consistency in the design of webpages of a website is considered. In this technique a web information extraction model based on statistical and positional relationship between the title and content is proposed. First the HTML file of the webpage is parsed and converted into DOM tree. Then each node is indexed starting from root node as zero. Next step is to calculate the attributes of each node. These attributes are: c which is content length of each node, a which is anchor length of each node, c_to_c which is ratio of single node content length to the total content length, a_to_a which is ratio of single node anchor length to the total anchor length, a_to_c which is ratio of anchor length to content length within a single node, need_ which is a Boolean type variable marking whether the node has effective texts. After the attributes are calculated the tag of <title> will be extracted as titlePage. Then traverse the DOM tree from top to bottom to compare each text node with titlePage.

Now, skip text node whose attribute of c is too short, filter out space and double quotation, and duplicate removal.

4. Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices [8]

This approach is used for Web page segmentation by recognizing repetitive tag patterns called key patterns in the DOM tree structure of a page. Repetition-based Page Segmentation (REPS) algorithm, which detects key patterns in a page and generates virtual nodes to correctly segment nested blocks.

In REPS, Web page segmentation by using the repetition detection algorithm proceeds in 4 phases as follows.

1. Less meaningful tags such as <a>, <b>, <script>, <span>, and “#comment” in the HTML source of the page are

removed. After this pre-processing step, a Web page is represented as a DOM tree structure.

2. A sequence is taken from a DOM tree of a Web page using the tags in the child nodes of the root node. This approach of considering only one-depth child nodes and ignoring all other “deep” descendant nodes, gives the advantage of reducing computational costs while still preserving some hierarchical features of the DOM tree.
3. Generate candidate Web page blocks by using the key patterns. A key pattern is a repetitive pattern in a sequence that is longest and most frequent.
4. Finally key pattern-based Web page segmentation recognizes blocks in a page by modifying the 1-depth DOM tree into a more hierarchical structure by building virtual nodes.

Although this technique builds virtual nodes for nested blocks, it has difficulty in finding deeply nested blocks for some pages since the block in REPS is determined by the number of repetitions. Also, it could not control the number of blocks expected from the segmentation.

5. The Research and Implementation of Web Information Extraction Technology Based on Multi-level Pages [9]

This technique has two methods of web information extraction. The first method is width priority analysis method based on regular expressions. The second method is depth priority analysis method based on DOM tree. For width priority analysis first with the label positioning information is located at the innermost label and using regular expressions target information is extracted. The information fragments extracted is then saved in local file, next the related links in the same page is extracted and stored in the URL queue. Now pages in URL queues are downloaded and fragments of information and URL are extracted on the new page in the same way as described above. The above steps are repeated until all the information fragments are extracted. Finally take out fragments of information stored locally and integrate into structured data.

For depth priority analysis Build the original DOM tree with the home page. The useful information on the current page can be directly extracted, and the information on the next level pages can be downloaded according to URL queue, after that connect the page to relevant URL node as a DOM sub-tree. In a similar way, all of pages’ information located in different levels of the website will be expanded to the DOM tree as a child node. Of course, if all the pages are crawled down and added to a DOM tree, this tree will be very large, it takes a very large storage space. So we can use dynamic pruning method. When finish crawling all the information of each level pages, put the following sub-tree node cut. And generate a new sub-tree extracting the information in the next level page. This method will greatly reduce the storage space of the tree.

The advantage of width priority analysis is that it is more flexible and drawback is that the process is too complex. The advantage of depth priority analysis is that there is no need to write a regular expression, the operation is relatively simple; Drawback is that it only apply to extract the contents of the label information for the Web and need to build a DOM tree.

IV. PROPOSED METHODOLOGY

In this work, we work on to replace the less popular advertisements. This can be done with the use of least recent used algorithm. This algorithm is the 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.

Steps involved :

1: Input the web content to the parser
2: Apply Dom tree algorithm
3: Extract all the links in the webpage.
4: Compare the Domain name of the link with the Domain name of the Website using Regular Expression.
5: Links having the same Domain name will be considered as Website links and others as External links.
6: If the External links contains Keywords such as Name of the ad providers than those links are considered as Advertisement Links.
7: According to the position of the advertisement on the webpage, no. of clicks are counted from the log file.
8: The Advertisement having the least number of clicks for that position within a specified period of time than that advertisement is replaced with the new highest paying advertisement for that position.

Algorithm : Advertisement replacement by LRU.
Input: Web page.
Output: Advertisements, Updated Webpage.

```
//Initial Parsing of page

pages[] = Get_links_of_pages()
forEachpages as page
```

```
{
    document        = get_content_of_this_page( page );
    a_links[]       = document.find_a_tags();

    forEacha_links as link
    {
        matchResult =
        ink.match_regular_expression_from_db();
        if( matchResult.success() )
        {
            new_href        = generate_dynamic_link();
            new_iframe      = generate_dynamic_iframe();

            link.href       = new_href;
            link.replace_first_image_with( iframe );
            ads_new_entry_in_db();
            document.save();
        }
        else
        {
            continue_for_next_anchor_tag();
        }
    }
}

// by clicking on ad (the dynamic link)

ad_id   = get_refrence_id_of_ad ();
ad      = find_or_fail_ad_from_db ();

increment_click_for( ad );
url     = ad.get_ad_url();
redirect_to( url );

// cron job fires on particular time interval

pages[] = Get_links_of_pages();
forEachpages as page
{

    ad = fetch_least_clicked_ad_in_a_page()
    forEach ads as ad
    {
        if ( ad.clicks<minimim_clicks_threshold )
        {
            ad = find_new_best_bead();
            update_in_ad_live_table( ad, page, position );
        };
} }
```

V. EXPERIMENTS AND RESULT

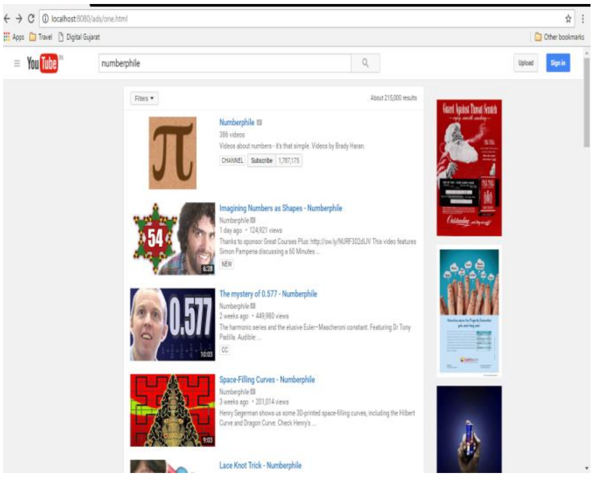
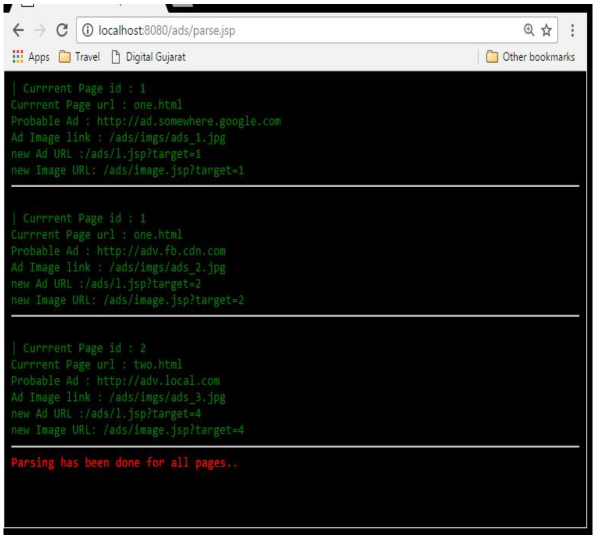
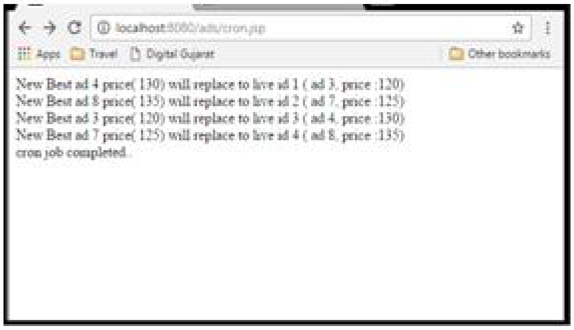

localhost:8080/ads/one.html
Apps
Travel
Digital Gujarat
Other bookmarks
YouTube
numberphile
Upload
Sign in
Filters
About 215,000 results
Numberphile
Imagining Numbers as Shapes - Numberphile
The mystery of 0.577 - Numberphile
Space-Filling Curves - Numberphile
Lace Knot Trick - Numberphile

Fig: 2. Before initial parsing, simple html page with some ad



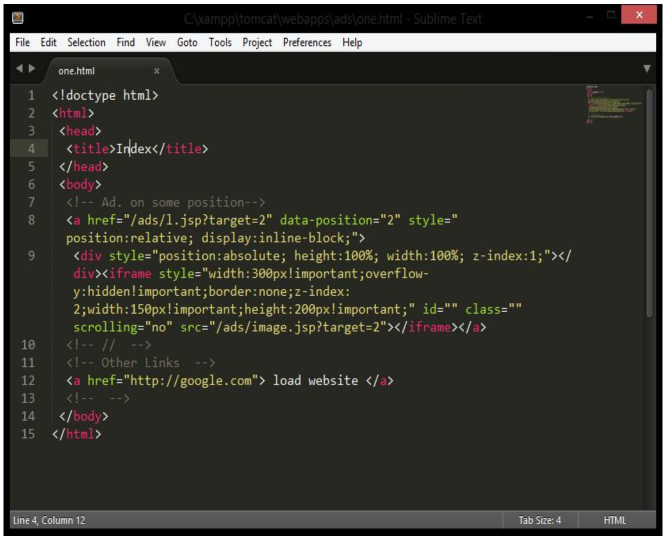
```
| Currrent Page id : 1
Currrent Page url : one.html
Probable Ad : http://ad.somewhere.google.com
Ad Image link : /ads/imgs/ads_1.jpg
new Ad URL :/ads/1.jsp?target=1
new Image URL: /ads/image.jsp?target=1

| Currrent Page id : 1
Currrent Page url : one.html
Probable Ad : http://adv.fb.cdn.com
Ad Image link : /ads/imgs/ads_2.jpg
new Ad URL :/ads/1.jsp?target=2
new Image URL: /ads/image.jsp?target=2

| Currrent Page id : 2
Currrent Page url : two.html
Probable Ad : http://adv.local.com
Ad Image link : /ads/imgs/ads_3.jpg
new Ad URL :/ads/1.jsp?target=4
new Image URL: /ads/image.jsp?target=4

Parsing has been done for all pages..
```

Fig: 3. Applying initial parsing



```
New Best ad 4 price( 130) will replace to live ad 1 ( ad 3, price :120)
New Best ad 8 price( 135) will replace to live ad 2 ( ad 7, price :125)
New Best ad 3 price( 120) will replace to live ad 3 ( ad 4, price :130)
New Best ad 7 price( 125) will replace to live ad 4 ( ad 8, price :135)
cron job completed..
```

Fig: 4. Cron fired on particular interval


C:\xampp\tomcat\webapps\ads\one.html - Sublime Text
File Edit Selection Find View Goto Tools Project Preferences Help
one.html
Line 4, Column 12
Tab Size: 4
HTML

```html
<!doctype html>
<html>
 <head>
  <title>Index</title>
 </head>
 <body>
  <!-- Ad. on some position-->
  <a href="/ads/1.jsp?target=2" data-position="2" style="position:relative; display:inline-block;">
   <div style="position:absolute; height:100%; width:100%; z-index:1;"></div><iframe style="width:300px!important;overflow-y:hidden!important;border:none;z-index:2;width:150px!important;height:200px!important;" id="" class="" scrolling="no" src="/ads/image.jsp?target=2"></iframe></a>
  <!-- //  -->
  <!-- Other Links  -->
  <a href="http://google.com"> load website </a>
  <!--  -->
 </body>
</html>
```

Fig: 5. After Parsing static links have been updated to dynamic links inside iframe.

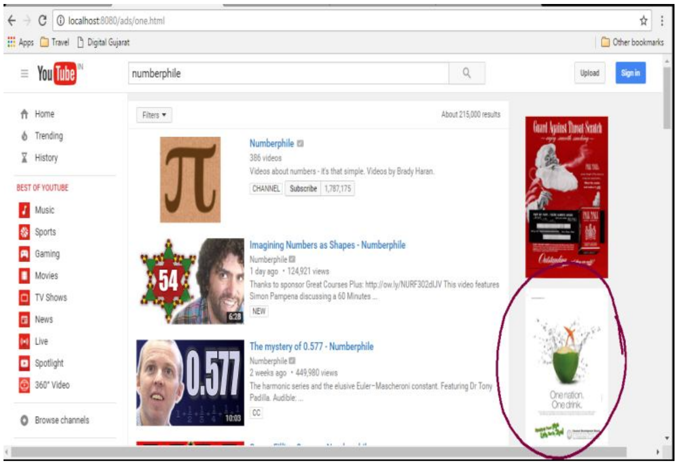


Fig: 6. Ad had clicks less then threshould so it has been updated after cron fiired

VI. CONCLUSION

In this paper, we have reviewed different techniques for informative content extraction and noise removal. Here, designing of the webpage has been given more emphasis to recognize the informative noise content. For precise and efficient content extraction and noise removal we can work on visual importance and can improve the efficiency of the DOM tree algorithm.

REFERENCES

[1] Shuang Lin, Jie Chen, ZhendongNiu, “Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction” ,TSINGHUA SCIENCE AND TECHNOLOGY, ISSNll1007-0214ll05/18llpp256-264 Volume 17, Number 3, June 2012

[2] A.F.R.Rahman, H.Alam and R.Hartono, “Content extraction from HTML documents”, International workshop on Web Document Analysis, pp. 7-10, 2001.

[3] WaridPetprasit and SaichonJaiyen, “Web Content Extraction Based on Subject Detection and Node Density”, 978-1-4799-6049-1/15/$31.00 ©2015 IEEE

[4] W3C Document Object Model (2009) Website. http://www.w3.org/DOM

[5] P. Sivakumar, “Effectual Web Content Mining using Noise Removal from Web Pages,” Springer Science+Business Media New York 2015

[6] Neetu Narwal, “IMPROVING WEB DATA EXTRACTION BY NOISE REMOVAL”, IEEE 2013

[7] Mingdong Li, Pingping Xu, Chencheng Yang “Content Extraction Based on Statistic and Position Relationship Between Title and Content”, IEEE/CIC ICCC 2014 Symposium on Social Networks and Big Data

[8] Jinbeom Kang, Jaeyoung Yang, Nonmember and Joongmin Choi, Member, “Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices IEEE Transactions on Consumer Electronics, Vol. 56, No. 2, May 2010

[9] Hengyu Lai, Yifei Wei, Yali Wang, Mei Song, Xiaojun Wang, “The Research and Implementation of Web Information Extraction Technology Based on Multi-level Pages”, ISSC 2014 / CIICT 2014, Limerick, June 26-27

[10] Neetu Narwal, Mayank Singh, “Web Content Extraction A Heuristic Approach”, (IJCSIS) International Journal of Computer Science and Information Security,Jan 2013.

[11] “Implementation” http://en.wikipedia.org/wiki/Implementation.

[12] Kurt Thomas, Elie Bursztein, Chris Grier_, Grant Ho, Nav Jagpal, Alexandros apravelos, et al. “Ad Injection at Scale: Assessing Deceptive Advertisement Modifications”, 2015 IEEE Symposium on Security and Privacy.

[13] Mohammad Mehdi Yadollahi, Masoud Asadpour, “AWS: Automatic Webpage Segmentation”, 2016 Second International Conference on Web Research (ICWR).