International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

eDEW: Effective Data Extraction from Web

Shalaka Patil

Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune, India

Abstract: ***Internet has become most popular place for accessing World Wide Web (WWW). With the enormous growing amount of information over Internet, accurate and efficient web data extraction has become necessary. Nevertheless, there are various kind of web pages which are having structured, semi-structured and unstructured data. A web page is a formation of many information blocks. Besides an informative block, web pages often consist of the distracting elements such as advertisements, copyrights, navigational panel, etc which are called as "Noise". Useful content or Information Extraction from the web pages becomes a critical issue for web users and web miners. The user can be misguided by the noise of the web page. So an effective web data extraction for users to conceive the useful information from the noisy information is urgently required. The main feature of web pages is that Web data extraction mainly deals with unstructured and semi structured form of data.***

Keywords: DOM Tree, Information Extraction, Pattern Tree, Web Mining, Web Data Extraction

1. Introduction

Web mining is the phenomenon of retrieving useful information or knowledge by adopting the data mining techniques from web data. Web is growing along with its strengths and its weaknesses. The strength is that web user can access information on single click over search engine, even if the quality varies. The weakness is that there is the problem of abundance and type of information [1]. The data mining techniques may be applied for mining information on the Web, but data mining deals with structured form of data while web mining deals with unstructured form of data. So mining of web data is one of the most challenging tasks for the data mining.

Web mining is categorized into three areas:

- Web Structure Mining
- Web Usage Mining
- Web Content mining

Web Structure Mining focuses on discovering and creating a model of the data organization which can be used to classify web pages. Web Usage Mining performs an observation over user's behavior in interacting with a particular web site using a web log files, user profiles, user session, cookie, etc for extracting the usage patterns of web user.

Web Content Mining is the process of identifying the content of Web pages as well as results of Web searching [2]. The web content data consist of structured data such as data in the tabular format, unstructured data as free text, and semi-structured data such as HTML documents.

Nowadays, many of a web pages are automatically generated by templates which do not only contain an actual content but also some noise such as advertisements, banners, company logos, copyrights, links, navigational panels, privacy announcements/notices, scrolls, service channels, etc. Figure 1 shows a sample of a web page displaying many advertisements (noise).

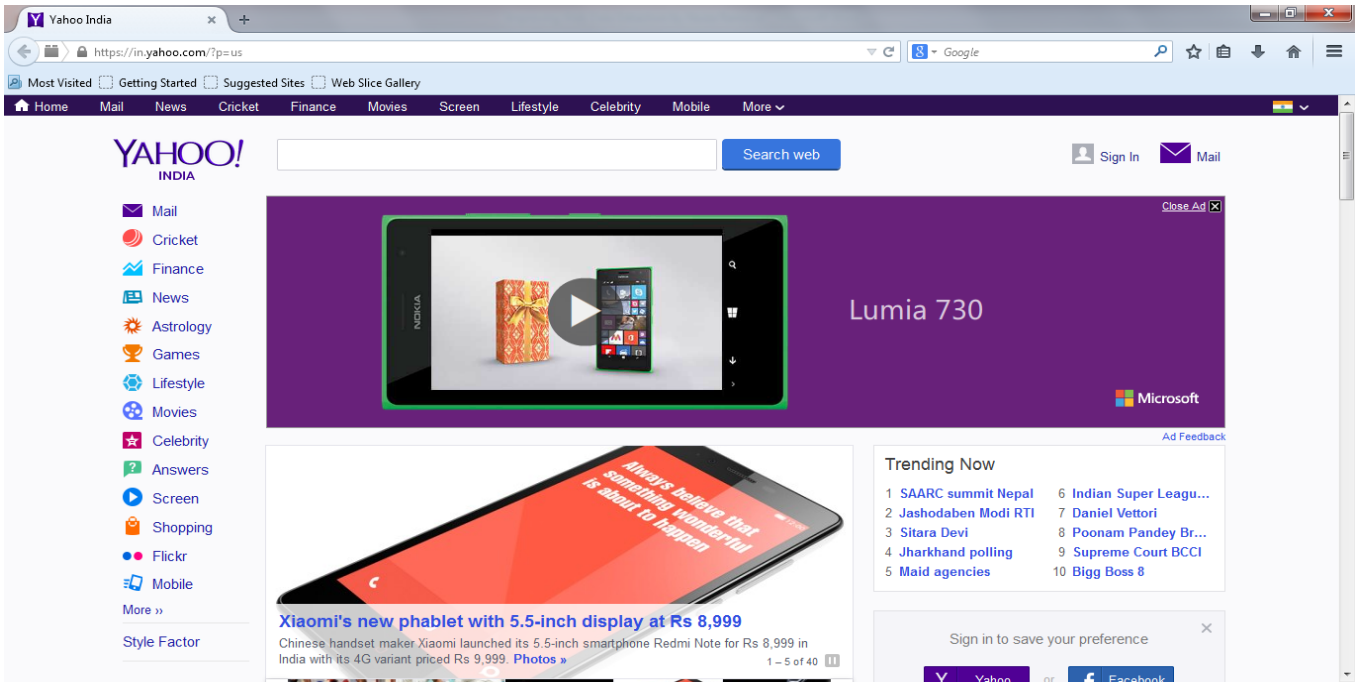

Figure 1: A yahoo mail web page displaying advertisements (noise)

Web sites can be classified based on the information represented by the web pages of that web site. As some web pages display only static information, whereas other web pages fetches the information from the backend database dynamically during runtime, even some web pages run complex scripts to generate data at the time of display. At last web browser displays web page which can be the merger of different types of visual blocks. The content blocks displayed on a web page can be viewed as a combination of Informative Blocks which functionally work as information representative i.e. main content and non-informative Blocks i.e. noise such as advertisements, copyrights, links, navigational panels, etc [3].

Effective extraction of high-quality contents from web page is crucial for many web applications such as information retrieval, abstract summary, automatic text categorization, machine translation, topic tracking, and helping end users to access the Web more easily over constrained devices like Personal Digital Assistants (PDAs) and cellular phones for providing better access to the Web [4].

2. Page Representation

Although XML is more structurally well representative for describing the contents of a page than HTML, majority web pages are written in HTML and count of such web pages is increasing rapidly. Most of the current Web pages are still in HTML rather than in XML on the Web. The large number of HTML pages on the Web is not likely to be transformed to XML pages in the near future. And these kinds of web pages contains huge amount of non-informative blocks. As per the studies non-informative blocks represent between 40% and 50% of the Web and they are growing every year. Hence, elimination of such noise from HTML pages is to be focus. According to counts, about 70% of all Web sites use HTML tag <TABLE> to develop their web pages [5]**.** Document Object Model (DOM) defines HTML and XML documents as a tree structure, in which tags are internal nodes of the tree, and texts or hyperlinks to other trees, are leaf nodes. Apparently, <TABLE> is not the only way to develop web pages. If a web page contains enormous number of content which may includes too many texts, based on the specification of W3C DOM, it can be represent using several tags, such as the title, headings, <P>, or <TR> and <TD> embedded in <TABLE>. Besides tabular tags, the content enclosed by the <TITLE> tag considered as special information as many web sites assign the same title to their web pages, such as the company's name or the default name generated by authoring tools.

3. DOM Tree

Even though XML was proposed for several years, large amount of information is present in HTML format over the Internet. The Document Object Model (DOM) specification provides an application programming interface (API) and number of classes for accessing valid HTML and well-formed XML documents. Each HTML page corresponds to a DOM tree [5][6]. Figure 2 shows a segment of HTML codes and its corresponding DOM tree.

HTML DOM tree extracts the structural information from the web pages and presents an HTML document as a tree structure comprising of different types of nodes i.e. Element Nodes and Text Nodes. An Element Node is represented as a Container node with opening and closing tag or Empty node, which does not have a closing tag. Each Element node has a name and a set of attributes associated with it and the content displayed within the tag. A Text Node represents the data that may be a sequence of characters, image, multi-media file residing between tags in the source file [3]. DOM provides each web page with a fine-grained structure, which illustrates not only the content but also the presentation of the page. Perhaps it is sufficient to gain information about web pages from the DOM tree, but the DOM tree also filled with lots of unrequired information which is need to filter [6].

Although a DOM tree is used to represent the layout and representation style of a web page but it is difficult to find the noise from a single web page and eliminate them based on individual DOM trees. Thus, DOM trees are not enough to remove the noise which considers both layout and representation style of a web pages. Hence a new tree structure, called *Style Tree* (ST) is to employ which enables for compression of the common presentation styles of a set of related Web pages [7].

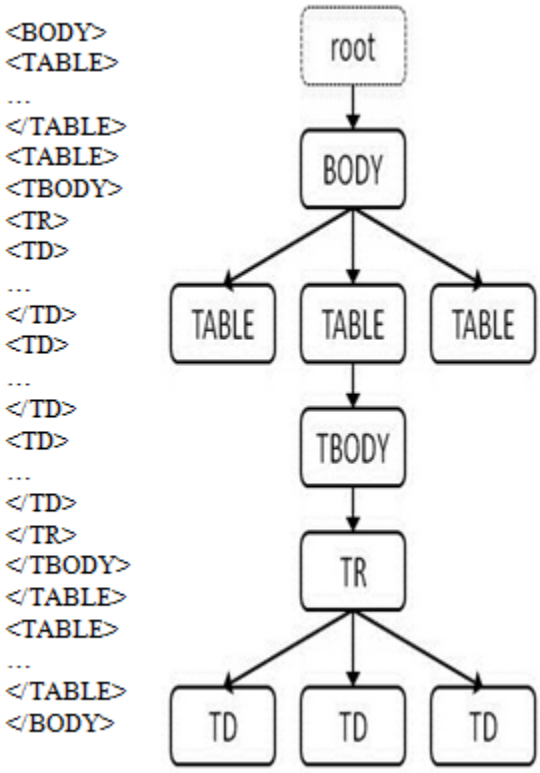

Figure 2: Graphical Representation of DOM

4. Style Tree

An example of Style Tree is given in Figure 3 which is a combination of DOM trees D_{T1} and D_{T2}. As well Figure 3 shows two DOM trees D_{T1} and D_{T2} which represents all tags in D_{T1} have their corresponding tags in D_{T2} except for the four tags P-IMG-P-A at the bottom level. Thus according style tree feature DOM tree D_{T1} and D_{T2} can be compressed. Count is used to indicate that how many web pages have an

identical style at a specific level of the style tree. In Figure 3, both web pages start with BODY, and thus BODY has a count 2 [7]. Below BODY, both web pages also have the same presentation style till level 4 of TD-TD-TD. The whole sequence of tags (TD-TD-TD) is known as *style node*, which is enclosed in a dash-lined rectangle in Figure 3. A style node represents a sequence of tag nodes at a given in a DOM tree. In the style tree, these tag nodes are called as the element nodes so as to distinguish them from tag nodes in the DOM tree. In Figure 3, we can see that below the right most TD tag, D_{T1} and D_{T2} diverge, which is reflected by two different style nodes in the style tree i.e. P-IMG-P-A and P-BR-P respectively. This means below the right most TD node, two different presentation styles are present. The page count of these two style nodes are both 1. Thus, it is clear that the style tree is a compressed representation of the two DOM trees. So now it enables to observe which parts of the DOM trees are common and which parts are different.

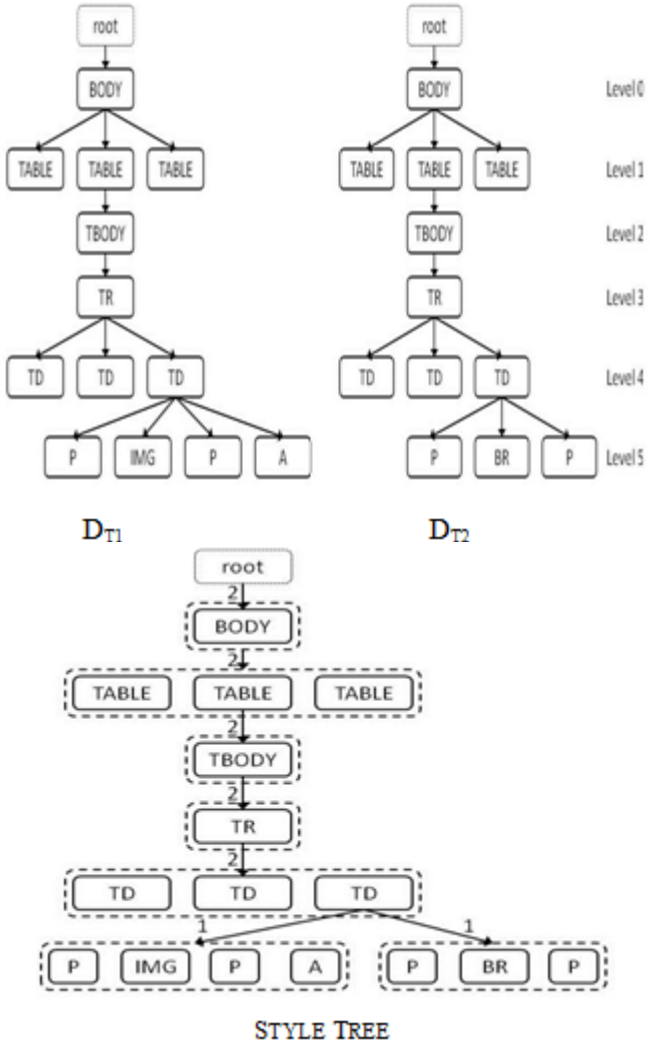

Figure 3: DOM trees and the style tree

5. Web Page Segmentation

Web page segmentation is performed over web pages to differentiate informative block from noise blocks which identifies the informative block based on both layouts and contents of the web pages. Web page creators tend to organize their content in a reasonable way, with proper features such as color, image, link, position, size, word, etc. Contents of different topics are arranged in separate information blocks. Informative blocks always stand in conspicuous position and the non-informative blocks are inconspicuous [8].

A Style Tree is used to represent both layouts and contents of a web page. Node count is calculated based on the Style Tree for each style node. It is observed that node count of the noise element usually more as it follows the similar style across web pages, where as the main content shows variation. Usually the noise element shows repeatedness, where as the main content shows uniqueness. But both of the presentation layouts are used to determine the actual noise over web page [3].

6. Conclusion

In this paper, the study is proposed to deal with the removal of noisy data from the web pages. The non-informative data considered as primary noise have to be removed, to improve the performance of web content mining. There are many web pages are present over the Internet who shares identical elements with similar style layout but different contents which are non-informative blocks. Although DOM tree can provide basic information about web pages, A Style tree is an effective technique to find actual content of Web site apart from noise but the construction of style tree is a complex task.

7. Acknowledgement

This work is an ongoing part of post-graduation at Savitribai Phule Pune University for a Master of Engineering in Computer at Pune. An author takes this opportunity to thank Dr. P. K. Deshmukh of Computer Engineering Department for his encouragement, support and untiring cooperation.

References

[1] Neeraj Raheja, Dr. V.K.Katiyar, A Survey on Data Extraction in Web Based Environment, International Journal of Software and Web Sciences, 5(2), June-August, 2013, pp. 135-139.
[2] Margaret H. Dunham, Data Mining: Introductory and Advanced Topics, Pearson Education, 2003, pp. 193-218.
[3] Neetu Narwal, Improving Web Data Extraction by Noise Removal, Communication and Computing (ARTCom 2013), Fifth International Conference on Advances in Recent Technologies,20-21 Sept. 2013, Page(s): 388 – 395.
[4] Yan Guo, Huifeng Tang, Linhai Song, Yu Wang, Guodong Ding, ECON: An Approach to Extract Content from Web News Page, Web Conference (APWEB), 2010 12th International Asia-Pacific, 6-8 April 2010, Page(s): 314 – 320.
[5] Shian-Hua Lin, Jan-Ming Ho, Discovering Informative Content Blocks from Web Documents, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, July 23-26, 2002, Page(s): 588-593.

[6] Yuancheng Li, Jie Yang, A Novel Method to Extract Informative Blocks from Web Pages, Artificial Intelligence, 2009. JCAI '09. International Joint Conference on 25-26 April 2009, Page(s): 536 – 539.
[7] Lan Yi, Bing Liu, Xiaoli Li, Eliminating Noisy Information in Web Pages for Data Mining, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, August 24-27, 2003, Page(s): 296-305.
[8] YuJuan Cao, ZhenDong Niu, LiuLing Dai,YuMing Zhao, Extraction of Informative Blocks from Web Pages, Advanced Language Processing and Web Information Technology, 2008. ALPIT '08. International Conference on 23-25 July 2008, Page(s): 544 – 549.

Author Profile

Shalaka B. Patil have completed Bachelor of Engineering in Computer Science and Engineering from Savitribai Phule Women's Engineering College, Aurangabad in 2013. Currently, She is pursuing Masters of Engineering in Computer Science from JSPM`s Rajarshi Shahu College Of Engineering, Tathawade, Pune, Maharashtra, India