IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.11, November 2022 E-ISSN: 2347-2693

HTML Tag Structure Based Content Retrieval from Web Pages

S.S. Bhamare

School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon (M.S) India

Author's Mail Id: ssbhamare.nmu@gmail.com

DOI: https://doi.org/10.26438/ijcse/v10i11.3539 | Available online at: www.ijcseonline.org

Received: 15/Oct/2022, Accepted: 02/Nov/2022, Published: 30/Nov/2022

Abstract— With the immense quantity of information in the World Wide Web, the World Wide Web (WWW) contains enormous amounts of web pages which are accessible by users. Web pages formatted in HTML (i.e. Hyper Text Markup Language) are found on this network of computers. All the Web pages, pictures, videos and other online content can be accessed via a Web browser. This provides a very useful and helpful means of collecting information. Information retrieval systems can help to retrieving the relevant information from web documents. This process of information retrieval involves three stages such as identifying the documents want to be processed, writing of query and use of searching mechanism to retrieve the relevant web document information. This paper discuss how HTML Tags structure of web page are useful for retrieval of main or informative content from web pages for efficient web mining operations.

Keywords— WWW, Web Page, HTML Tags, Text Density**.**

I. INTRODUCTION

In the massive network of World Wide Web (WWW), web pages have large amounts of information. Web researcher always requires main or informative content (e.g., an article or informative text) from the web pages. Mining the data on the Web has become a major task for locating useful information from the Web.

In this age of information, there exist a huge amount of electronic data and information worldwide. Exploiting the information resources and turning them into useful knowledge available to concerned people is a great challenge. In construction of web pages HTML tags are used and plays important role in retrieval of web page information.

HTML tags makes entire structure of web page. This structure provides content information of web document and can be useful for efficient retrieval. This HTML tag structure of web pages is useful for identification of noise content and informative content from web pages. It is essential to remove noise content such as copyright notices, advertisement, link etc. for effective web mining.

In the rest of the paper we discuss about web page structure, HTML tags as hidden keywords within a web page that define how the web browser must format and display the web page content, usefulness of tag attributes and HTML tags, HTML tags based splitting operation for web page main content retrieval.

II. RELATED WORK

Many researchers have worked in this area for efficient web page information retrieval. Most of them concentrate on Web page HTML tag structure for information retrieval. Malik et. al. taking the advantage of the HTML structure of web and n-gram technique for partial matching of strings, an n-gram based algorithm for mining web content outliers is proposed. To save time, the optimized algorithm uses only data captured in <Meta> and <Title> tags. Experimental results indicate that the proposed n-gram-based algorithm is capable of finding web content outliers. In addition, using texts captured in <Meta> and <Title> tags give the same results as using text embedded in <Meta>, <Title>, and <Body> tags [1].

Gupta et. al. developed a framework that employs an easily extensible set of techniques. It incorporates advantages of previous work on content extraction. They work with DOM trees and implement an approach in a freely available web proxy to extract content from HTML web pages. This proxy can be used both centrally, administered for groups of users, as well as by individuals for personal browsers [2].

Pan Ei San et. al. Proposes a Content Extraction algorithm describes how to get high performance without parsing DOM trees. After observation the HTML tags, one line may not contain a piece of complete information and long texts are distributed in close lines, this system employs Line-Block concept to assess the distance of any two neighbor lines with text and Feature Extraction such as text-to-tag ratio (TTR), anchor text-to-text ratio (ATTR) and new content feature like Title Keywords Density (TKD) to identify noise from content. After extracting the features, the system uses these features as parameters in threshold method to classify the block as content or non-content [3].

Kaasinen et. al. Splits the web page using tags such as <P>, <TABLE> and <UL> for further changes or summarization [6].
Wong W et. al. defines tag types for page segmentation by giving a label to each part of the web page for classification. Apart from the tag tree, some other algorithms utilize the content or link information [7].

Deng Cai et. al. proposes visual feature of web page-based method. It uses to apply such visual information as font size, layouts, background color etc. to divide web page into corresponding visual blocks. The method simulates how people observe web pages. In the web page center main content blocks of information would be present and first caught by user's eyes. But due to complexity of vision feature, it is hard to find a universal rule set [8].

Zhao Xinxin et. al. proposes tag window-based web content information extraction method, which could deal with some special circumstances that all the web content information is put into one <td> or morel <td> tag. But it requires semantic analysis and similarity judgment, which enhance complexity [10].

III. WEB PAGE STRUCTURE

Generally Hyper Text Markup Language (HTML) is used to written web page document that is available through the internet using an internet browser. HTML is Standard markup language which is mostly used for construct web pages. Each web page on the Internet is written using one version of HTML code or another. HTML code ensures the proper formatting of text and images, so that the internet browser can display them as they are meant to appear. Without HTML, a web browser would not know how to display text as elements or load images or other elements. HTML also provides a basic structure of the page, upon which Cascading Style Sheets are overlaid to change its feel and appearance. One could think of HTML as the backbone (structure) of a web page, and Cascading Style Sheets (CSS) as its skin (appearance).

A. HTML File

Every web page is an HTML file. Each HTML file is just a plain-text file, but with an .html file extension instead of .txt, and is made up of many HTML tags as well as the content for a web page. A web site always contains many html files that link with each other.

B. HTML Tag

HTML tags are the hidden keywords within a web page that define how the web browser must format and display the web page content. Most of the tags must have two parts, an opening, and a closing part. For example, <body> is the opening tag and </body> is the closing tag. The closing tag has the same name of the tag as the opening tag, but has an additional forward-slash (/) character and this interprets as the "end" or "close" character, for example

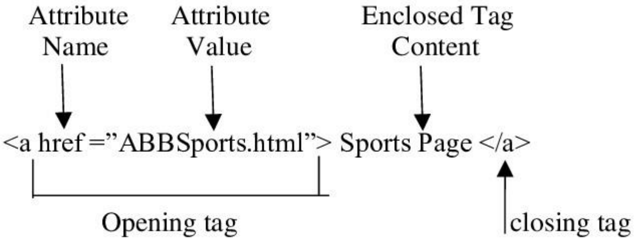


Fig. 1

There are some HTML tags that are exceptions to this rule, and where a closing tag is not compulsory. The <img> tag used for showing images is one example of this. Each HTML file should have the essential tags for it to be valid, in order that internet browsers will understand it and display it correctly [11].

C. Tag Attributes

HTML tag attributes allow us to customize a tag, and are defined within the opening tag, for example:
<img src="imageflower.jpg"> or <p align="centre">…</p>
Tag attributes are often allocated a value using the equals sign, such as border="0" or width="40%", but there are some that only need to be declared in the tag like this: <hr noshade>.

Most of the attributes are not necessary for most of the tags and are only used when you want to change something about the default way a tag is displayed by the browser. However, some HTML tags such as the <img> tag has important attributes such as src (source) and alt which are needed for the web browser to display the web page correctly [11].

Distinct HTML types of tags are used to design or construct web pages. As mentioned earlier, data of web page is covered within a pair of open and close matching HTML tags. Some HTML tags can be used to design a web page layout and hold the main content of web pages and some tags cannot have any use, they are only used to make segment on layout and to display links and pictures on web pages. Generally main or informative content is present within a pair of <BODY> and its corresponding tags.

Typically, each HTML tag used for constructing the web pages must play a special role. To identify the main or informative content of web pages, a list of important HTML tags is used. It is noted that, web page HTML tags may have important tags such as <body>, Title (<title>), Head(<H1, H2, H3, H4, H5, H6>), Paragraph (<p>), Bold (<b>), Strong (<strong>),Italic (<i>) etc., intermediate tags such as, Table (<table>) and List (<li>) etc. and not important or irrelevant tags such as <script>, <style>, <iframe>, <object>, and <form> etc. [12]. It is essential to mention that if no tag is found between <body> and

</body>, then one single block is considered as the whole page. Thus, web page HTML tags are divided into two categories namely, important tags called as contain holder or relevant tags and non-important tags called as decoration or irrelevant tags based on the content of the web page as shown in table 4.1.

IV. METHODLOGY

In this proposed method Web page HTML Tag Structure is used for content retrieval. It helps to determine main content tags and irreverent content tags of web pages.

A. Dataset Used

The experimental assessment is done on different categories of Web Pages such as Sports, Technology and Main Pages from three news web sites, CNNIBN, ABB News and Times of India.

These Web pages (or Web documents) are extracted through search engines by giving query. Here, web scrapping method used to extract Web page HTML tag content information (or HTML source code) of each web page is used. This method is implemented using open source python library.

B. Preprocessing the Web Page Tags

Initially the HTML syntax of web document is checked because most of the HTML Web pages are not well-formed, to make it correct through an HTML parser [16].

In preprocessing on web page tags, those tags that do not contain any text, as well as invalid tags such as <script>, <style>, <marquee> <meta>, <anchor> etc., which are not related to the main content of web page are filtered out or removed first to improve the determination of content and irreverent tags from web pages. Here HTML tag based filtering technique is implemented by using regular expression. Web page tags that do not contain any text as well as invalid tags are found and removed.

C. Html tags splitting operation for content retrieval

In HTML tags splitting operation, HTML tags are split based on the content composed in the HTML web script. Webpage usually consists of HTML tags, head, title, and body that are the valuable tags amongst the various tags available in the HTML web script. Usually, the content is arranged inside the DIV and TD sub tags of Body tag. These contents are said to be information. On set of web pages, splitting operation is conducted on HTML tags of HTML Tag structure, generally HTML tags can be divided into two categories (shown in table 4.1),

Contain holder tag (Main/Informative Content Tags): Used to plan the layout of the web page, these tags visually divide the web page into several content blocks.

Description tag (Irrelevant Content Tags): Used to describe a segment of content in the web page, these tags have no use of the layout of the web page but just a picture or a hyperlink.

Table 4.1: List of common contain holder and decoration Tags

Sample List of Common Holder/ Relevant Tag (Main Content tag)	<body>…. </body>, <table>…. </table>, <tr>…. </tr>, <td>…. </td>, <ul>…. </ul>, <div>…. </div> and <form>…. </form>, <title>…. </title>, <h1, h2, h3, h4, h5, h6>… < /h1, /h2, /h3, /h4, /h5, /h6>, <p>…. </p>, <b>…. </b>, <strong>…. </strong>, <i>…. </i> and <li>…. </li> etc.
Sample List of Common Description or Designer Tags (Irrelevant Content Tags)	<a>….</a>,<img>….</img>,<font>….</font>, <span>….</span>,<style>….</style>, <link>….</link>, <script>….</script>, (<!.........!>), <noscript>…. </noscript>, <hr>….</hr>, ….</br>, and <form>…. </form> etc.

These tags can split into two types of files i.e. Contain holder tag file & Description or Designer tag file as shown in below Figure.4.1

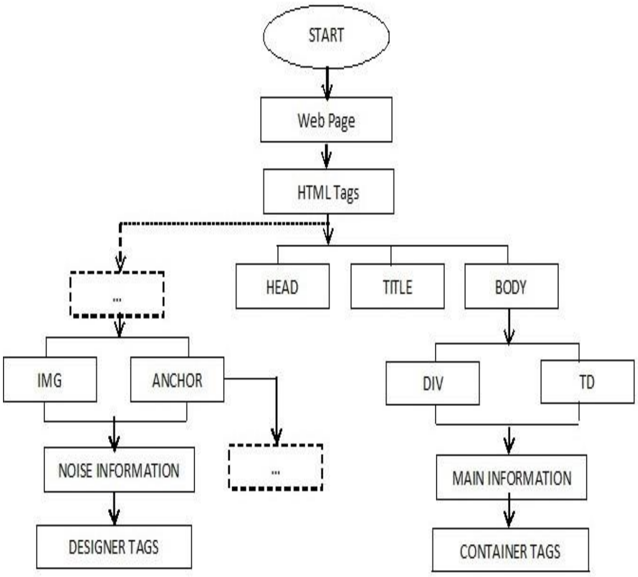


Figure 4.1. HTML tag splitting operations

Mostly useful content - informative content - is found in <BODY> tag and its corresponding tags. Body tags are used to identify the content of a web page. Focus is given to extract the information of <BODY> and there corresponding tags. While performing HTML Tag Splitting Operation, must have to Computing Text density or Tag Weight of each tag including corresponding tag of web page and finally based on computed text density, identifying two sets of tags i.e. High Density Tag (HDT) and Low Density Tag (LDT) for Content Extraction.

Text Density: with the help of Text density to find the irrelevant noise content which is more formatted and contains a small text and main content usually is lengthy and less formatted. Once an HTML Page has been parsed, the number of characters per tag and tags that each node contains can be figured out.

Text density or Tag weight can be defined as,

$$TDi = TCi / Ti \qquad \textbf{(Eq. 4.1)}$$

Where,

TDi = Text Density or Tag Weight of each tag (node)
TCi = Count No. of Characters (including corresponding tags)
Ti = Count No. of tags (including corresponding tags)

Text density or Tag weight computed by Eq 4.1,

The following example 4.1 shows sample brief segment of HTML code and its corresponding computed text density of tags.

Segment of HTML code	Text density for the four tags computed as follows

<div class="main">
<div class="article">
<div class="heading">
PM visit to USA</div>
<div class="article - body">
Indian PM is schedule to visit USA
</div></div></div>
	1. <div "main">:
	Chars=48, Tags=3, Density=16
	2. <div "article">:
	Chars=48, Tags=2, Density=24
	3. <div "heading">:
	Chars=14, Tags=1, Density=14
	4. <div "article-body">:
	Chars=34, Tags=1, Density=34

Example 4.1: HTML code and corresponding text density of tags.

Example 4.1 shows the text density results for sample HTML code of a web page. There are tag nodes with a comparatively high text density for example tag number 2 and 4 having more text density such as 24 and 34 as compared to others two. Certainly, the high text density tags portion can be taken as the web page's content. Similar way to apply this method of computing text density on each web page for identifying main or informative content of web pages. [14]. The high-density tags hold the main informative contents of web pages that generally required for different web mining tasks. The low-density tags of web pages generally considered as a noise content tag contain irrelevant information and that should be eliminate before web mining.

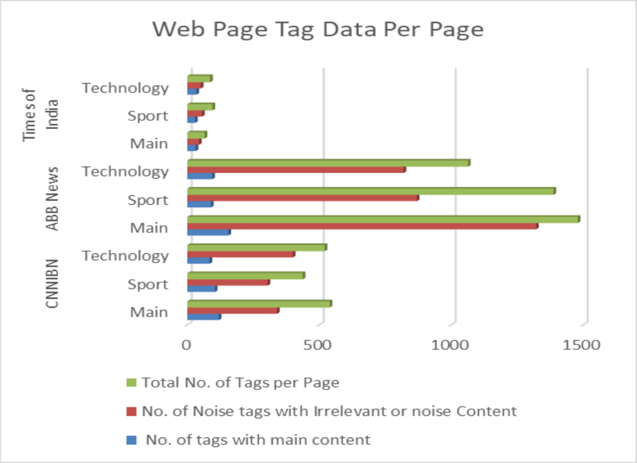
V. Result and Discussion

This experiment provides statistical information of web pages for performance evaluations, while working on web page tag information the text density of each tag including internal tags are computed using Eq. 4.1 and through best threshold value (Tr) to identify Less density tags or LDT (usually more formatted and contain a small text) and High density tags or HDT (usually lengthy and less formatted) to determine main and noise or irrelevant content tags of web pages. Finally, extracting useful content of web pages.

The following table 5.1 and graph 5.1 shows the No. of tags with main content and No. of tags with Irrelevant or noise Content information out of total number of tags of nine sample web pages.

Table 5.1: Web Page Tag information

Web Site	Source Category	No. of tags with main or informative content (HDT)	No. of noise tags with irrelevant or noise Content (LDT)	Total No. of Tags per Page
CNNIBN	Main	120	340	540
	Sport	105	305	438
	Technology	85	400	521
ABB News	Main	156	1322	1480
	Sport	90	870	1388
	Technology	95	820	1064
Times of India	Main	32	45	66
	Sport	30	56	96
	Technology	35	52	87



Graph 5.1: Web Page Tag Information

In the experimental results of HTML Tag structure based method it is observed that, In different category of Web pages comparatively main content tags are found less than irrelevant or noise content tags out of total number tags of web pages, such as in source category of Sports, Technology and Main web Pages from web sites CNNIBN found 120, 105 and 85 tags of main content 340, 305 and 400 irrelevant tags out of 540, 438 and 521 total number of tags of web pages and similar for other category of web pages. Overall on an average found 60 to 65 percent of irrelevant or noise information of tags and 35 to 40 percent main or informative contents of tags on web pages. It is essential to retrieve only main content tags information and eliminate irrelevant tags information for effective web mining operation.

VI. Conclusion

In web page tag information, observed that most of the part of web pages occupy by irrelevant or noise content information as compared to main content information. it is also found that generally web page main content tags can hold large amounts of text or characters as compared to others. The text density of each tag along with their corresponding tag of each page is computed and through the best threshold value the main content tag i.e., High density tag and Less density tag of web pages are identified and only the information of high-density tags i.e. main or informative content tag of web pages are extracted for effective information retrieval and used in further web mining tasks.

REFERENCES

[1] Malik Agyemang, Ken Barker, Rada S. Alhajj, Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams ACM Symposium on Applied Computing, pp.**482-487, 2005**.

[2] Gupta Et. Al Automating Content Extraction of HTML Documents World Wide Web: Internet and Web Information Systems, Online version published in **2004.**

[3] Pan Ei San, Boilerplate Removal and Content Extraction From Dynamic Web Pages, International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.**4**, No.**6**, **2014.**

[4] Li Xiaoli, Shi Zhongzhi Innovative Web Page Classification through Reducing Noise Journal of Computer Science and Technology, Vol.**17,** No**. 1**., **2002**

[5] A.K. Tripathy, A.K. Singh An Efficient Method Of Eliminating Noisy Information In Web Pages for Data mining, in Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04), **2004.**

[6] Kaasinen, E., Aaltonen, M., Kolari, J., Melakoski, S., and Laakko, T., Two Approaches to Bringing Internet Services to WAP Devices, In Proceedings of 9th International World-Wide Web Conference, **pp. 231-246, 2000.**

[7] Wong, W. and Fu, A. W., Finding Structure and Characteristics of Web Documents for Classification, In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Dallas, TX., **USA, 2000**.

[8] Deng Cai, Yu Shipeng and Wen Jirong, "VIPS: a vision-based page segmentation algorithm", Microsoft Technical Report, **MSR-TR-2003-79, 406-417, 2003.**

[9] Sun Chengjie and Guan Yi, "A Statistical Approach for Content Extraction from Web Page", Journal of Information Processing, Vol.**18**, Issue.**5**, pp.**17-22, 2004.**

[10] Zhao Xinxin,Suo Hongguang and Liu Yushu, "Web Content Information Extraction Method Based on Tag Window. Application Research of Computers, Vol.**24,** Issue.**3**, pp.**144-145, 2007.**

[11] Simple HTML Guide, **2014.**

[12] List of main html tags. Online; accessed 25 march, **2014.**

[13] S S Bhmare, B.V. Pawar "An Efficient Method of Web Page Noise Cleaning for Effective Web Mining", International Journal of Computer Applications (0975 – 8887) Vol.**146** – No**.3**, pp.**18-22**, **2016.**

[14] Dandan Song, Fei Sun, Lejian Liao.‖ A hybrid approach for content extraction with text density and visual importance of DOM nodes‖. In the proceedings of Springer Knowl Inf Syst, Verlag London. Vol.**42**, pp.**75-96, 2015.**

[15] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, New York, **1983.**

[16] Soma Chatterjee, Kamal Sarkar "A Comparative Study of Three IR models for Bengali Document Retrieval" International Journal of Computer Sciences and Engineering E-ISSN 2347-2693 Vol.**07**, Issue.**1**, pp.**220-225**, **2019.**

AUTHORS PROFILE

S. S. Bhamare is working as an Assistant Professor in School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University (formerly Known as North Maharashtra University), Jalgaon. His total teaching experience of 12 years and published more than 12 papers in reputed peer reviewed national and international journals & conferences. His research area includes Web Mining, Information Retrieval and IOT.