International Journal of Research in Advent Technology, Vol.7, No.3, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Use of Web Page Tag Information for Efficient Information Retrieval

Dr. S. S. Bhamare
School of Computer Sciences Kavayitri Bahinabai Chaudhari North Maharashtra University
Jalgaon (M.S) India.
Email: ssbhamare.nmu@gmail.com

Abstract- The World Wide Web (WWW) contains enormous amounts of web pages which are accessible by users with the intent of searching information. Web pages are formatted using Hyper Text Markup Language (HTML). All the Web pages, pictures, videos and other online content can be accessed via a Web browser. This provides a very useful and helpful 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 to be processed, writing of query and use of searching mechanism to retrieve the relevant information. With the demand of effective page ranking, we have discussed how HTML tag structure information is useful in searching mechanism to improve efficiency of web page information retrieval and provide relevant information.
Keywords- HTML Tags, Index, Ranking, Searching, Information Retrieval.

1. INTRODUCTION

In the colossal network of World Wide Web, web pages contain large amounts of information. Web researchers always require main content of information (e.g., an article text) from the web pages to be gathered, processed and stored quickly and efficiently. Mining the data on the Web has become an important 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. For searching information on WWW search engines are used. Search engines required query as input based on that it provide us relevant information. Different search engines are using different approaches or techniques for the searching user required information from WWW.

Web page is a document commonly written in Hyper Text Markup Language (HTML) that is accessible through the Internet using an Internet browser. HTML is a standard markup language which is commonly used for creating web pages. Every web page is actually an HTML file 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 are linked with each other. Individual web page on the Internet is written using one or another version of HTML code. 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.

The main objective of this paper is to inspect how web page HTML tag structure information are useful for efficient information retrieval from web pages, such as Title: <Title>, Headings: <H1..H1>, Link: <Anchor> etc.

2. RELATED WORK

Michal Cutler & et al (1997) propose Web-based search tool (Webor), it uses HTML tag structure of web page. Webor contains an indexing engine and a search engine. In Webor they grouped HTML tags into six different classes [1].

Kaasinen, E & et al (2000) splitting the web page using tags such as, <UL>, <B>, <TABLE> and <P> for additional changes or summarization [2].

Wong, W & et al (2000) describes 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 [3].

Gupta, S., & et al (2003) suggests a DOM-based content extraction method to assist information access over controlled devices like PDAs. They implemented an advertisement remover by maintaining a list of advertiser hosts, and a link list remover based on the ratio of the number of links and non-linked words [4].

YongZhang & et al (2010) propose a new approach of the web page purification based on improved DOM and statistical learning. They produce a block tree structure that is very helpful for applications such as web page classification, information retrieval and information extraction, it enables to find the main content block of the web page [5].

Malik Agyemang & et al (2005) 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 [6].

3. WORKING OF MAJOR SEARCH ENGINES

Nowadays various search engines are available on WWW for searching of required information but out

of them only few are popular used by users such as Google, Bing and Yahoo etc. These search engines have used common search parameter in their working model. Search engines used spiders or spider bots application for fetching required web pages from web, Indexing method that stores these pages and rank them by their relevancy. The following figure 1. Shows anatomy of typical search engines.

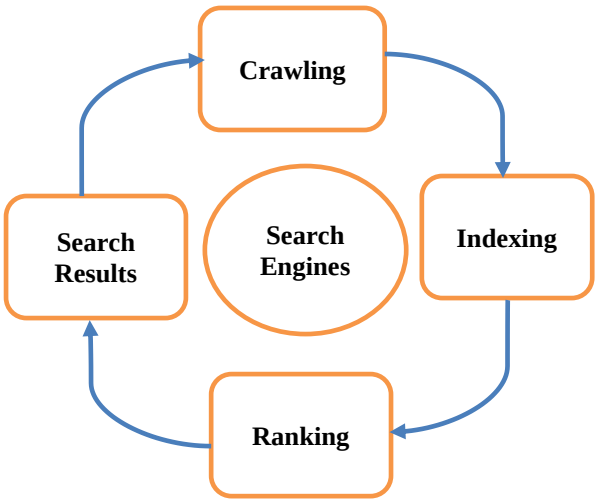

Fig. 1. Anatomy of Search Engine

Generally major search engines can search relevant information on web through three steps process like crawling of web pages, Indexing of web pages and ranking of web pages.

Google is one of the most popular and widely used search engine today. Google also uses crawling, indexing and ranking phases for searching information. Google was the first engine to reflect importance of web structure and uses HTML Tags information to rank the search page. Page Rank method plays a significant role in returning relevant results to simple query. Google also provides other important features which help in improving the search. [9]

Bing search engine was developed by Microsoft in the year 2009 to compete with Google. The searching method of Bing is based on two concepts i.e. Relevancy and Click distance. Relevancy is related with relevant information to provide users, for this it uses frequency with keywords appear in search web pages. The concept click distance is used to rank the search pages. Click distance is nothing but number of clicks is used to access a particular web page information i.e. less number of clicks gives faster accessibility of page information. It also uses URL depth property in that it count number of back slashes in URL have to reach page information. So, longer URL is less important page information.

In 2004 yahoo developed its own web crawler called 'Yahoo Slurp' in order to return relevant information. Yahoo search engine is similarly work like Bing search engine for searching relevant web pages from web based on user search query.

In working of major search engines, Google has gained large popularity than Bing and Yahoo also is the most optimized search engine than others. [9]

4. PROPOSED STUDY

Usually search engine web page ranking algorithm always depends on what type of indexing method is used for searching and storing information.

Our proposed study aims whether the HTML tag structure information is useful to search engine indexing method for ranking of relevant information. Use of web page HTML tags information can help search engines and its mechanism for efficient web page retrieval and ranking and provide us relevant information.

In this study we used three different categories of web pages such as Sports, Technology and Main Pages from three news web sites, i.e. CNNIBN, ABB News and Times of India. We examined and studied HTML tag structure and importance of HTML tags of different 120 web pages of these three category.

We observed that each HTML tag used for constructing the web pages has to play a special role, to identify the main, informative or important content of web pages.

Here we use combination of HTML Tags information and its importance in construction of index for searching tool. For this we study number of HTML tags elements of various web pages for construction of index such as <body>, Title (<title>), Heading (<H1, H2, H3, H4, H5, H6>), Paragraph (<p>) and Anchor text <A> and others.

It is also noted that, as per the importance and information hold by the HTML tags web page HTML tags are further divided into three level of tags as below,

Important tags: Information hold by these tags are relevant with the search query, such as <body>, Title (<title>), Head (<h1, h2, -----, h6>), Paragraph (<p>), Bold (<b>), Strong (), Italic (<i>), Underline <ul>, , <td>, <tr>, <li> etc.

Intermediate tags: Information hold by these tags are mostly related with the search query, such as, Table (<table>) and List (<li>), Anchor text <A> etc. and

Irrelevant tags: Information hold by these tags are not relevant with the search query, such as <script>, <noscript>, <style>, <iframe>, <object>, <img>, , <font> and <form>, (<!.....!> etc. [8].

This three levels of tags information helps to search engines or tools for effective searching and indexing of relevant information. 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.

The webpage typically consists of HTML tags like head, title and body etc. that are the most important tags among the various tags available in the HTML web script. Usually the main content is composed within the DIV and TD sub tags of Body tag. These contents are said to be information on a given set of web pages.

Using these HTML tag information for indexing is surely improve the results of searching techniques and help to search and rank relevant pages. In this study we also consider total computed weights of important tags for indexing method that helps to improve retrieval and ranking of the pages e.g. <title> tag i.e. Title of the web page given a more weight as compared to others tags.

So, in future research with help of this study we propose and implement methodology to develop searching technique mechanism that help to provide efficient and relevant information of user search query.

5. CONCLUSION

In this paper we have discussed how HTML tags structure information are useful in efficient searching. Web page construct with the help of various HTML tags. Each HTML tag can have its own important and hold some values. Web Page HTML tags are divided into list of Important, Intermediate and Irrelevant tags. List of important tags hold useful information i.e. relevant to our search query. Each HTML tags of web page can play special role for searching of information from web documents. Researchers must consider information of HTML tags structure for searching relevant information from web and develop a searching technique mechanism for effective searching of relevant information from web.

REFERENCES

[1] Michal Cutler, Yungming Shih, Weiyi Meng, Using the Structure of HTML Documents to Improve Retrieval, Proceedings of the USENIX Symposium on Internet Technologies and Systems Monterey, California, December 1997

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

[3] 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.

[4] Gupta, S., Kaiser G., Neistadt D. and Grimm P., DOM based Content Extraction of HTML Documents, In the proceedings of the Twelfth World Wide Web conference(WWW 2003), Budapest, Hungary, May 2003

[5] YongZhang, Ke Deng, Algorithm of Web Page Purification Based on Improved DOM and Statistical Learning in 20I0 International Conference On Computer Design And Applications (ICCDA 2010).

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

[7] Simple HTML Guide [Online]: http://www.simplehtmlguide.com /whatishtml.php

[8] List of main html tags. Online; http://www.webseomasters.com/ forum/index.php?showtopic=81.

[9] https://networkedlifeq21.fandom.com/wiki/Networked.life.q21_Wikia

[10] https://www.entrepreneur.com/article/224639