ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

Main Content Extraction From Web Page Using Dom

Ms. Pranjali G. Gondse[1], Professor Anjali B. Raut[2]

Abstract: Today internet has made the life of human dependent on it. Almost everything and anything can be searched on net. The rapid growth of World Wide Web has been tremendous in recent years. With the large amount of information on the Internet, web pages have been the potential source of information retrieval and data mining technology such as commercial search engines, web mining applications. Internet web pages contain several items that cannot be classified as the informative content, e.g., search and filtering panel, navigation links, advertisements, and so on called as noisy parts. Most clients and end-users search for the informative content, and largely do not seek the non-informative content. A tool that assists an end-user or application to search and process information from Web pages automatically, must separate the "primary or informative content sections" from the other content sections. These sections are known as "Web page blocks" or just "blocks." First, a tool must segment the Web pages into Web page blocks and, second, the tool must separate the primary content blocks from the non informative content block .Main focus is on review and evaluation of algorithm , capable of extracting main content from web page. Proposed algorithms outperform several existing algorithms with respect to runtime and/or accuracy. Furthermore, a Web cache system that applies proposed algorithms to remove non informative content blocks and to identify similar blocks across Web pages can achieve significant storage savings will be shown.

Keywords: DOM Tree, information extraction, web mining

I. INTRODUCTION

Nowadays, World Wide Web has become one of the most significant information resources. It delivers the information mainly in the form of the Web pages. The rapid expansion of the Internet has made the WWW a popular place for disseminating and collecting information. Extracting useful information from Web pages thus becomes an important task. Usually, apart from the main content blocks, web pages usually have such blocks as navigation bars, copyright and privacy notices, relevant hyperlinks, and advertisements, which are called Non-informative blocks. These blocks are not relevant to the main content of the page. These items are required for web site owners but they will hamper the web data mining and decrease performance of the search engines. These blocks are very common in web pages. Major efforts have been made in order to provide efficient access to relevant information within the web pages. Today's Web pages are commonly made up of more than merely one cohesive block of information. So extracting exact information content becomes difficult. Efficiently extracting high-quality content from Web page is crucial for many Web applications such as information retrieval, automatic text categorization, topic tracking, machine translation, abstract summary, helping end users to access the Web easily over constrained devices like PDAs and cellular phones. The extracted results will be the basic data for the further analysis. So content extraction from Web page has attracted many researchers recently. The advantage of identifying non-content blocks from web pages is that if user does not want non-content blocks these can be deleted. These non-content blocks are normally large part of the web pages so eliminating them will be a saving in storage and indexing.

II. RELATED WORK

To identify Informative content from web page is relatively easy task for human being because he can easily identify important content by visual inspection but it is difficult task for computer.

Finn et al. discuss methods for content extraction from "single-article" sources, where content is presumed to be in a single body. The algorithm tokenizes a page into either words or tags; the page is then sectioned into 3 contiguous regions, placing boundaries to partition the document such that most tags are placed into outside regions and word tokens into the center region. This approach works well for single-body documents, but destroys the structure of the HTML and doesn't produce good results for multi-body documents, i.e., where content is segmented into multiple smaller pieces, common on Web logs ("blogs").

Extracting content from HTML documents has been well studied and numerous methods have been developed. Perhaps the most simplistic approaches are seen in handcrafted web scrapers which specifically know how to extract article text by looking for known HTML-cues with regular expressions written in Java or Perl or with specialized tools designed for content extraction such as NoDoSE or XWRAP . An obvious disadvantage of this approach is that different rule expressions need to be manually created for each website. Furthermore, an individual website may also change its structure or layout over time making this approach in need of continuous maintenance. McKeown et al. in the NLP group at Columbia University detects the largest body of text on a webpage (by counting the number of words) and classifies that as content. This method works well with simple pages. However, this algorithm produces noisy or inaccurate results handling multi-body documents, especially with random advertisement and image placement. Rahman et al. propose another technique that uses structural analysis, contextual analysis, and

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 3, March 2014

summarization. The structure of an HTML document is first analyzed and then properly decomposed into smaller subsections. The content of the individual sections is then extracted and summarized. However, this proposal has yet to be implemented.

Lin and Ho proposed an extraction method based on information entropy, the web page is divided into content block according table tag, each block has entropy, and then information blocks are obtained by comparing with threshold value. But this method just applies to web pages which contain table tags, while increases the complexity of the algorithm. Yi and Liu put forward an extraction approach based template, this method assumes that the same part of two web pages having same format, so it is simple and effective to remove noises by comparison of two web pages coming from one source. But it is difficult to identify so many templates for a variety of web pages. The extraction approach based on framework web pages and rules supposes it is reasonable the noise blocks generally locate in the secondary positions in the page. This method compares the ratio of width and height attributes of every table tag, and removes the tags of bigger ratio. It is hard to work well on table tags with less height and width attributes. The table tag is the only Processing Objects to this approach. Web content information extraction method based on tag window which could cope with some special circumstances that web pages content text locate in table and div tags, all page content information is put into one td or several tds, and the length of body text is short as that of the other information such as navigation bars, and the copyright, etc. But during the process of judging body text, it involves word segmentation and computing similarity of string, which has enhanced the complexity of information content extraction. Pan and Qiu put forward a web page content extraction method based on link density and statistic, which recognizes main content according to the different properties between content nodes and non content nodes of web page represented as a tree. But the threshold values in this paper do not always adapt to some news pages such as stock news, so it is still hard to find a set of universal parameters.

These methods, based on removing noise is suitable to delete a lot of nodes without any web content information, they can contribute to filter the unrelated part using the layout properties of noises in the web pages. The most current ways lack of enough considering on removing noise in the preprocessing. Furthermore, large numbers of financial news have so many hyperlinks in the information text that those methods which over-reliance on links have poor results. After analyzing these methods about removing noise and information content extraction, and moreover, the relation among punctuation mark density, length of information text and anchor text is considered enough in the extraction stage. We put forward a method based on removing noise and characteristics of body text.

III. ANALYSIS OF PROBLEM:

Typically, a modern web document comprises of different kinds of content. Besides the article posting as the main content it also contains other noisy contents such as user comments, navigational menus, headers, footers, links to other web pages, advertisements, copyright notices which scatter over the page. Considering the fact that a web document contains various forms of contents, it influences the way Human browses the documents. When browsing a particular web documents, most of the time users typically focuses on the main content and ignore the additional contents. The presence of noisy contents may degrades the performance of such Information Retrieval application for example the quality of the search, accuracy of information extraction, and the size of the index

IV. PROPOSED WORK:

Proposed approach concentrates on web pages where the underlying information is unstructured text. The technique used for information extraction is applied on entire web pages, whereas they actually seek information only from primary content blocks of the web pages. The user specifies his required information to the system. Web crawlers download web pages by starting from one or more seed URLs, downloading each of the associated pages, extracting the hyperlink URLs contained therein, and recursively downloading those pages. Therefore, any web crawler needs to keep track both of the URLs that are to be downloaded, as well as those that have already been downloaded. DOM analyzer defines the concept of blocks in web pages. Most web pages on the internet are still written in HTML. Even dynamically generated pages are mostly written with HTML tags, complying with the SGML format. The layouts of these SGML documents follow the Document Object Model tree structure of the World Wide Web Consortium.2. The relevant pages given out by the web crawler are represented in a form of DOM tree HTML DOM is in a tree structure, usually called an HTML DOM tree. Following Figure illustrates a simple HTML document and its corresponding DOM tree. We are interested only in the <BODY> node and its offspring. In this example, <BODY> node has three children: element nodes <B> and <I>, and text node #and. Element node <B> has a text node child #bold, and element node <I> has a text node #italic. Following the DOM convention, we use <> to indicate element node, and use # to indicate text node.

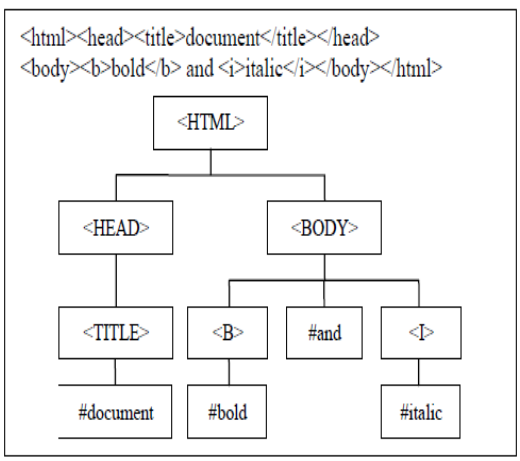

Figure : Simple HTML document and its corresponding DOM tree.

Page segmentation algorithm partition the web page based on the first tag in the list to identify the blocks, and then sub partitions the identified blocks based on the second tag and so on. It continues to partition until there is any tag left in a block in the block set which is a part of the list of tags. This ensures that the blocks are atomic in nature and no further division is possible on them.
Steps for content Extraction:-

Step 1: Cleaning the HTML page:

a. Symbols, "<" and ">", should only contain html tags. When used in other place, they should be replaced by "&lt" and "&gt" respectively.
b. All tags must be matched, i.e. every starting tag has a corresponding ending tag.
c. Attributes of all tags must be encircled by quotation marks.
d. All tags must be nested correctly. For example, <a>. . .. . is a correct nest, while <a>. . .. . is incorrect.

Step 2 : Preprocessing the web page tags.

All tags on the page form a tree structure. Those nodes that do not contain any text should be removed, as well as invalid tags such as<script> <style> <form> <marquee> <meta> etc, which are unrelated to the content. Then the structure tree is built.

Step 3: Judging the location of content

The aim of this process is to select the optimum node containing content. If a node is not satisfied with this condition, the text under this node is not identified.

Step 4 :Extracting the content

The content is extracted by tools such as html parser. If the node is not satisfied with the conditions, return the step 3 in order to find the optimal nodes of the next level nodes (the child nodes of the node).

Step 5 :Adjusting the extraction results from step 4

In step 3, only the node that most likely contains the content is selected. But if the structure of a web page is relatively decentralized, it is very prone to extract a section or a paragraph of the whole content. As the adjacent nodes on the same level are free of judge, in this step, we must adjust the above result. The text also should be extracted from the adjacent nodes that meet the conditions of the precise content extraction. So all text will be extracted from the qualified nodes on the same level.

V. CONCLUSION

In this paper I have proposed a method which gives the informative content to the user. Using DOM tree approach contents of the web pages are extracted by filtering out non informative content. With the Document Object Model, programmers can build documents, navigate their structure, and add, modify, or delete elements and content. With this features it becomes easier to extract the useful content from a large number of web pages.
In future this approach will be used in information retrieval, automatic text categorization topic tracking, machine translation, abstract summary. It can provide conceptual views of document collections and has important applications in the real world.

REFERENCE

[1] Jae-Woo LEE "A Model for Information Retrieval Agent System Based on Keywords JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN COMPUTER ENGINEERING ISSN: 0975 – 6760| NOV 12 TO OCT 13 | VOLUME – 02, ISSUE – 02 Page 297
Distribution" International Conference on Multimedia and Ubiquitous Engineering(MUE'07), IEEE 2007 0-7695-2777-9/07
[2] Marti A. Hearst andXerox PARC "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages" Association for Computational Linguistics,
2007
[3] A. F. R. Rahman, H. Alam and R. Hartono "Content Extraction from HTML Documents"
[4] [4] Wolfgang Reichl, Bob Carpenter, Jennifer Chu-Carroll, Wu Chou "Language Modeling for Content Extraction in Human-Computer Dialogues".
[5] S.-H. Lin and J.-M. Ho, "Discovering informative content blocks from web documents," in KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2002, pp. 588–593.
[6] Jinbeom Kang, Joongmin Choi, "Detecting Informative Web Page Blocks for Efficient Information Extraction Using Visual Block Segmentation", International Symposium on Information Technology Convergence, pp 306-310, November 2007.
[7] Shian-Hua Lin, Jan-Ming Ho, "Discovering informative content blocks from Web documents", Proceedings of ACM SIGKDD'02, July 2002.
[8] Suhit Gupta, Gail Kaiser, David Neistadt, Peter Grimm"DOM – based Content Extraction Of HTML Documents".
[9] Yan Guo, Huifeng Tang, Linahi Song, Yu Wang, Guodong Ding "ECON: An Approach to extract Content from Web News Page",12th International Asia-Pasific Web Conference.