An Approach of Web Page Information Extraction

Yaohui Li, Lixia Wang, Jianxiong Wang, Jie Yue, Mingzhan Zhao
Department of Computer
Hebei Institute of Architecture and Civil Engineering
Zhangjiakou City, China
e-mail: lyhfirst@126.com

Abstract—**The Web has become the largest information source, but the noise content is an inevitable part in any web pages. The noise content reduces the nicety of search engine and increases the load of server. Information extraction technology has been developed. Information extraction technology is mostly based on page segmentation. Through analyzed the existing method of page segmentation, an approach of web page information extraction is provided. The block node is identified by analyzing attributes of HTML tags. This algorithm is easy to implementation. Experiments prove its good performance.**

Keywords-Information extraction; DOM; page segmentation; HTML tag

I. INTRODUCTION

With the development of Internet, more and more people pay attention to the information on web pages. In order to display information and easy to read, web pages must be regular on layout and structure. Web pages contain not only the main content, but also include many of irrelevant information, such as pop-up ads, unnecessary pictures and irrelevant links. This information is called web page noise, which has a serious impact on the useful information extraction of the Web. According to the size of the content of the noise, the noise of the Web content can be divided into two classes: global noise and local noise. Global noise is refers to the Web has a bigger size of noise content, it usually contains mirror website and approximate Web page. Global noise content not only affects the page collection, indexing and sort quality of retrieval results of information retrieval system on the web, also makes web information storage system waste an enormous amount of disk space to save repeat web pages. Local noise is refers to the content of Web pages that is irrelevant to the theme content of Web page, such as,advertising, navigation and copyright statement, etc. Local noise makes application very difficult to get exact webpage topic content, so it seriously influences the application based on the webpage content. At the same time, the local noise in many cases is accompanied by hyperlink, therefore, the local noise impact on the application based on link between webpage.

Search engine index the entire page content that include irrelevant information. If you do not remove the noise in the webpage content, so the index system must build index for noise content. Because the query words appear in webpage noise content, this webpage is returned as results. The topic of webpage contents could be completely unrelated to the query word.The noise content not only makes the index size become large, which will affect the efficiency, but also led to a decrease in retrieval accuracy.

Web page noise induced topic drift. Accurate identification and removal of noise content of the website, which we call web page purification, is a key technology to improve the accuracy of the results of the search engine. Web page purification can significantly simplify the complexity of the tag structure and reduce the page size, this can save the follow-up processing time and space. Therefore, the web page purification has become an essential part of the search engine. Noisy Information Elimination, an emerging study field, has developed. Some scholars call HTML Pages Purification [1].

II. PAGE SEGMENTATION

In the topic search field, a lot of advertising, navigation bar noise content can lead to topic drift. This shows that the Web Graph based on Web page is not accurate enough to the traditional topic search algorithm.We should study deeply into the Web internal and make the size of processing unit smaller, which improve the accuracy of content analysis.

HTML language is semi-structured. HTML is not semantic-oriented and the actual pages are not fully prepared to follow Web standards. There are difficulties and challenges that the computer analysis and understand web page content. Web page segmentation in accordance with the semantic block is an effective method. HTML pages can not only be regarded as the contents of tags and tag combinations. HTML pages have some independent semantic blocks, generally according to the visual block. Page segmentation acts an important role in the document classification, information extraction, topic information collection and search engine optimization.

According to the current study, the methods of web page segmentation can be classified into four categories: DOM tree-based page segmentation, template-based page segmentation, information analysis method, visual feature analysis method.

A. DOM Tree-Based Page Segmentation

DOM (Document Object Model) is the application program interface (API) for HTML and XML document[2]. Using the Document Object Model, programmers can build documents, add, modify or delete elements and content. DOM is a set of objects and access, interface dealing the document object. As a W3C specification, one important objective for the Document Object Model is to provide a

standard programming interface that can be used in a wide variety of environments and *applications*. The DOM is designed to be used with any programming language. In the DOM, documents have a logical structure which is very much like a tree. Each document contains zero or one doctype nodes, one root element node, and zero or more comments or processing instructions; the root element serves as the root of the element tree for the document.

HTML document include the title, head, paragraph, hyperlinks and other various components. DOM parse the HTML file and generate the internal tree structure of the file, called the document's logical structure or logical structure tree. Tree structure can accurately describe the relative position of elements and it is suitable to describe semi-structured data. DOM-based page segmentation is generally based on the predefined syntactic structure, that is, HTML tags. HTML tags are not independent. There is a certain hiberarchy relationship among them. Each tag effect on the resulting page is different: some only work on the visual features; some only effect the hierarchy; there's both of the two. Reference [3] based on DOM tree use HTML tags to identify the block.

DOM-based page segmentation method to the simple structure page will have better results. However, the present structure of the popular website often complex and not the rule. This method is suitable for used in combination with other methods.

B. Template-Based Page Segmentation

There is a relationship among the location of each module on web page. Reference [4] proposed a page segmentation method using web page layout. First, starting from the root of <body>, the page is divided into a few large pieces, including the page header, page bottom, left or right navigation bar, and get the relative position of each block. Then the contents of a web page is extracted and put into the characteristics template. The page is divided into several blocks.

Reference [5] proposed a method adopted machine learning to generate the template of page assembly. According to the template extraction rules, information on the subject of the page is extracted. However, this method are obvious only the template-based web page. This method does not have universal.

C. Information Analysis Method

This method uses the traditional information retrieval methods. Reference [6] regards the Web document as a character stream, uses some heuristic rules and the PAT algorithm to count the repeated tags sequence. Each repeated tag sequence is a message block. Reference [2] blocks a page according to <TABLE> tags. Words are extracted as features, then calculate the entropy of each word, and then calculate the entropy of each block. If the entropy of the block is greater than the threshold, the block was a important information block, otherwise not. Because these methods do not take into account the structural characteristics of Web documents, the error may be split when the information is not very consistent among the blocks.

D. Visual Feature Analysis Method

The primary strategy behind link building was to create quality links and quality anchor text. The current algorithms view each of the pages on the Web as just one node. Yet, the various blocks on the pages have different semantics. In other words, block on the right side of the page may have text link advertisements while a block on the let side may have an article. A link from the content block could be considered as more likely to be a true recommendation than a link from a text link advertisement block. Search engines may therefore give extra weight to in-content links while devaluing links that appear to be advertisements. Sites that rent links through link networks usually do place them in a block above, below or to the side of the content block.

Visual information of web page includes font size, background color, space, split edge and so on. Reference [7] proposed VIPS(vision-based page segmentation)algorithm that divide the page into visual blocks through the collection and processing of visual features and according to certain rules. This method is more close to the semantic block. VIPS will allow search engines to differentiate between links from the content block and links from other blocks such as text advertisement blocks or footer blocks. As such, algorithms could easily weight links from each block differently.

VIPS algorithm has been used successfully in the current information extraction

VIPS algorithm makes full use of the Web page layout features. First, all of appropriate page block extracted from the DOM tree. With each of the blocks separated, the links in each block can be more accurately analyzed. In other words, the context of the link can also be identified, which helps the search engine make better decisions about its true relevance. Then visual separators are detected among the page blocks, including horizontal and vertical directions. Finally, based on the separators, Web page semantic structures will be rebuilt. For each semantic block, the VIPS algorithm is used continuously to split the block into smaller blocks. The VIPS algorithm is top-down and is very efficient.

Compared with DOM based methods, the segments obtained by VIPS are much more semantically aggregated. Noisy information, such as navigation, advertisement, and decoration can be easily removed because they are often placed in certain positions of a page.

Due to the complexity of visual features, using the heuristic knowledge is often more vague, need to manually adjust the rules. The rules are often very much. Adding a rule will have an impact on the web pages that have parsed. Therefore, how to ensure the consistency of rule set is a major difficulty.

From the above analysis, we can see that the purpose of handling all the content is to exclude the interference of the noise page and get the real subject content.

III. THE IMPROVED INFORMATION EXTRACTION METHOD

Web standards are some standard set that made by the W3C and other standardization organizations to create and interpret web page. Now, more and more web pages comply with W3C standards and use CSS and tags to layout,

not <table> tags. The same HTML code apply different CSS controls. Using CSS achieve separation of content and structure of web pages and improve accessibility and maintainability of web pages.

We first analyze the type of HTML tags. According to W3C specification, HTML tags can be divided into two categories. Structure label determines the layout of web pages, commonly used tags are <table>, <tr>, <td>, , <p> and so on. These tags are line-break. Content label is used to display web content such as pictures, text, links, lists, such as <li>, <b>, <img> and so on. These tags are inline. When we analyse the page structure, we only consider the structure labels. Some pages use non-standard HTML tags, we can use HTMLTidy. HTMLTidy is a widely used label analysis tool, which is characterized by a strong fault tolerance. HTMLTidy can find the error of label in the page, for example, missing end tag, end tag matching error and so on, and make reasonable amendments.

View from the smallest particle size, each DOM node can be viewed as a block. The algorithm starts from the <body>root and attach attributes to each node .The main attributes we have chosen include layout information, content information. Layout information can be directly obtained from the CSS and HTML. Content information can be obtained through the statistics of nodes. Layout information that we have extracted include the coordinate of the node's center and the size of the node. Content information that we have extracted include the text length of the node, the link text length of the node.

Some nodes in the DOM tree are considered the block nodes according to the following rules.

1) The node is line-break type, and <table> tag are generally the layout of the label.

2) The length and width of the node can not less than the threshold. If a node's size is too small, it is certainly not the block node.

3) The node has not hidden attributes

4) The ratio of the link text length to the text length can not exceed the threshold.

5) If the node accords with the above rules and the text length is greater than MAX, the node is the block node. MAX is a predefine value. If there is not such node, the node whose text length is greatest in the DOM tree is block node.

All block node comprise the subject area. The text of such node is the subject information.

To verify the validity of the method, we made a small experiment. From Web, the 100 web pages are selected as the test data. The results of the block extraction are manually judged. In the experiment, the proportion of the pages which block extraction satisfy our requirement is 96%. Experimental results show that the method of subject information extraction is well for most pages. The reason that the result of the block extraction is not well is that some web pages do not comply with Web standards or the default threshold is not the best for some pages.

ACKNOWLEDGMENT

This work was supported by a grant from the Science Technology Research and development Guidance program of Zhangjiakou City (No. 060152) and Technology Research and Development Guidance Program of Hebei Province (No. 052135125)

REFERENCES

[1] Z.G. Zhang, J. Chen, and X.M. Li, "A HTML page purification method," Journal of the China Society for Scientific and Technical Information, 2004, 23 (4) : 387 - 393.

[2] W3C The Document Object Model [EB / OL]. Http://www.w3.org/TR/DOM-Level-2-HTML/

[3] S.h. Lin and J.M. Ho, "Discovering informative content blocks from Web documents," ACM SIGKDD International Conference on Knowledge Discovery&Data Mining, 2002, 588-593

[4] Y Chen, X Xie, W.Y. Ma, and H.J. Zhang, "Adapting Web pages for small-sscreen Devices,".InternetComputing , IEEE, 2005, 9(1):50-56

[5] J.W. OU, S.B. DONG, and B. CAI, "The theme information extraction for Web templating pages,"Journal of Tsinghua University, Vol.45(9), 2005, 1743-1747

[6] D.W. Embley, Y. Jiang, and Y.K. Ng, "Record-boundary discovery in Web documents," SIGMOD international conference on Management of data, Philadelphia, USA, 1999:467-478.

[7] D. Cai, S. Yu, J.R. Wen, and W.Y. Ma, "VIPS.a Vision-based Page Segmentation Algorithm." Microsoft Technical Report.MSR-TR-2003-79,2003