A STUDY OF CONTENT EXTRACTION FROM WEB PAGES BASED ON LINKS

R.Gunasundari [1] and Dr.S.Karthikeyan [2]

[1] Research Scholar, Karpagam Univarsity, Coimbatore, India
gunasoundar@rediff.com
[2] Director, School of Computer Science, Karpagam Univarsity, Coimbatore, India

ABSTRACT

Extracting main content from web page is the preprocessing of web information system. The content extraction approach based on wrapper is limited to one specific information source, and greatly depends on web page structure. It is seldom employed in practice. A new content extraction method is thus proposed in this paper, which can discover web page content according to the number of punctuations and the ratio of non-hyperlink character number to character number that hyperlinks contain. It can eliminate noise and extract main content blocks from web page effectively. Experimental results show that this approach is accurate and suitable for most web sites.

KEYWORDS
content extraction; wrapper; HTML tree; web pagenoise

1. INTRODUCTION

The Internet booming brings prosperity of many applications such as information retrieval, knowledge sharing, etc, and makes information overload at the same time. So it is an intractable problem of obtaining accurate information and knowledge from the Internet. Web page information extraction transforms the content of semi-structured web pages to structured text, which can be queried easily by the users. Nowadays, content extraction from web page is a key step for knowledge acquisition and preprocessing and lays a good foundation for the future text processing.

Traditional information extraction methods can be categorized into three types: free text extraction, semi-structured and structured web page extraction. Note that most of the existing web pages are semi-structured, in this paper we focus on only semi-structured text. Wrapper, which is a widely used method, suffers from the limitation of specific information source. However, it is impossible to establish so many wrappers for various web pages. Another type of semi structured extraction methods are based on the knowledge model, which also has defects. For example, the method based on ontology description requires experts' support and bring very necessary to bring forward an universal approach for the content extraction from web page.

In this work, we investigate a new approach of web page content extraction. Search engines offer an easy access to web pages, which makes information extraction an easier task. The preprocessing work like filtering and parsing has to be done before content extraction. The extracted content, which can be used for knowledge classification and knowledge processing, provides information sources for enterprises or other organizations.

DOI : 10.5121/ijdkp.2012.2303

In the following section, we give some related works and the origin of this algorithm. Then, in the main text, the extraction algorithm is described in detail.

2. RELATED WORK

The methods of web page content extraction vary due to different emphasis on the problem.

1) The extraction approach based on web page structure [1], maps web page to HTML tree [2] and then retrieves table nodes of the tree to detect the content. But how to identify the text blocks is unclear and it is still difficult to determine the threshold value.

2) Deng and Yu [3] put forward an extraction method based on visual feature of web page. It applies such visual information as font size, layouts, background color etc. to divide web page into visual blocks. The method simulates how people observe web pages. As in the center of web page, main content blocks would be first caught by people's eyes. But because of the complexity of vision feature, it is hard to find a universal rule set.

3) Sun and Guan [4] proposed an extraction approach based on statistic. It recognizes main content by character number on the assumption that character occurs more in content than in other part of web page. But it proved to be a unsatisfied method with low accuracy in practical experiment.

4) Zhao [5] put forward a web content information extraction method based on tag window, which could deal with some special circumstances that all the web content information is put into one <td> or several <td>. But it requires semantic analysis and similarity judgment, which enhance complexity.

5) The method of STU-DOM tree of web pages theme automatic extraction[6] is based on its own information to construct semantic tree, and prunes irrelevance nodes in accordance with local and contextual relevance. Several indicators are proposed about relevance. It is a beneficial exploration on the web pages information extraction method.

Through analysis above, we can conclude that machine learning or wrapper need a large number of artificial participation. The approaches based on DOM [7] are suitable for the vast majority of pages in <table> <td> tags and helps to determine the theme and extract content by statistics. The method of identifying content by the ratio of number of character in web page to that of hyperlink[8], [9] provides a reference for this paper. After summarizing others' methods about information extraction, a content extraction method based on link density has been put forward after analyzing the interference of link nodes in content.

Large numbers of news sites and web pages statistics show that the content of news web page is very focused and tends to be concentrated under a handful of nodes, and the remaining nodes mainly are noise nodes such as advertising, pictures, etc. Several important threshold values are acquired by a lot of tests and statistic, and the correct rate and the accuracy have greatly improved. A new way of thinking as well as a method for dealing with such issues is proposed.

3. THE APPROACH BASED ON LINK DENSITY

3.1 Related Conceptions

According to the statistics of a large number of websites, for the news web pages of semi-structure, the main page is filled with a large number of unrelated images and advertising links. The location of the main content is very centralized and has a good hierarchical structure. From

the statistical results we can find that the threshold values of the node containing content are obviously different from that of other nodes in the same level. According to this statistical information, we put forward an algorithm that judges the content by several parameters in the nodes.

1) Link Text Density (LTD): The ratio of all link text length to all content text length under a given node. This value is an important indicator to judge advertising block. The greater the value, the greater the possibility of advertising block.

2) Link Amount(LA): The number of link nodes of all child nodes under a given node. It is an important judgment index for the accurate content. The great value indicates that there are more child nodes, thus corresponding work has to be carried out to check its child nodes.

3) Link Amount Density (LAD): The ratio of the number of all child links to the number of all child nodes under a node. The value measures the number of link nodes in a node. It judges whether the content contains a number of links.

4) Node Text Length(NTL): The length of all text page with labeled nodes removed. LAD and NTL should be combined to perform the functions of accurate content judgment.

The four parameters are widely used in the algorithm, of which LTD and NTL are very important parameters for content location judgement, and LA and LAD are indicators for accurate content judgment.

3.2. The Steps of Content Extraction

Step 1 Standardizing the web page tags [10].

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

By comparing LTD, LA, NTL and LAD with corresponding neighbor nodes on the same level, the node including basic content is selected and processed in next step. The aim of this process is to select the optimum node containing content. According to the experimental results and statistics, we select
LTD = 0.4, LA = 10, NTL = 100, LAD = 0.5, as the basic content node judgment parameters. If a node is not satisfied with this condition, the text under this node is not identified. As the news web page is a tree structure, the content must be under a general node.

Step 4 Extracting the content

If the optimal nodes meet the following conditions:
LTD = 0.1, LA = 10, NTL = 50, LAD = 0.1, the text under them is the accurate content. The content is extracted by tools such as htmlparser. 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.

3.3 Algorithm Description

1) input the page to be processed
2) standardize the html tags
3) remove the tags unrelated to the content
4) construct the tree
5) n = 1, tag = false
6) Traverse the nth hierarchy nodes, m is the total number of the nodes in the n hiberarchy
7) FOR i = 0 to m
calculate those value: LTD [i], LA [i], NTL[i],
LAD [i] of the Node[i];
IF (LTD[i], LA [i], NTL[i], LAD [i] are within the scope of basic conditions of content) THEN
Tag = true;
END IF
END FOR
8) IF tag = true then
9) FOR i = 0 to m
10) IF (LTD [i], LA [i], NTL[i], LAD [i] are within the scope of basic conditions of content) THEN
11) IF (LTD[i], LA [i], NTL[i], LAD [i] are satisfied the requirements of precision content extration) THEN
exact text from the Node [i];
12) ELSE
Traverse Node [i] of nodes, n = n +1, return to
Step 6, the first dealing with n +1 layer Node [i];
END IF
13) ELSE
Continue;
END IF
END FOR
14) ELSE
This web page is not suitable for the algorithm
END IF

4. EXPERIMENTS AND APPLICATION VALIDATION

4.1 A Sample for The Algorithm

An actual example is showed to illustrate the information extraction process of the algorithm. The original web page is shown in Fig. 1. After preprocessing, the structured tree is shown in Fig. 2. The exact content location in the web page is found finally by the algorithm. The correct node <div class = ”article” id = ”karnataka bankcontent”> contains the whole content indeed.

Fig. 1. The original source page

As shown in Fig. 2, at first, the second level, which contains two nodes <head>and <body>, will be traversed. By comparing their threshold values, we find that the node <body> contains basic content. Then in the step 4 of the algorithm , the <body> node does not meet the requirements of accurate content node. Therefore, we need to analyze its child nodes (level 3): <div class = ”logo nav”>, and . By reciprocating the node <div class = ”article” id = ” karnataka bankcontent ”> meets the requirements of the accurate content node. Finally after the adjustment in step 5, the nodes and do not meet the requirements of accurate content node. So the whole extraction process is complete. The node <div class = “article” id = ” karnataka bankcontent”> is the proper node containing content.

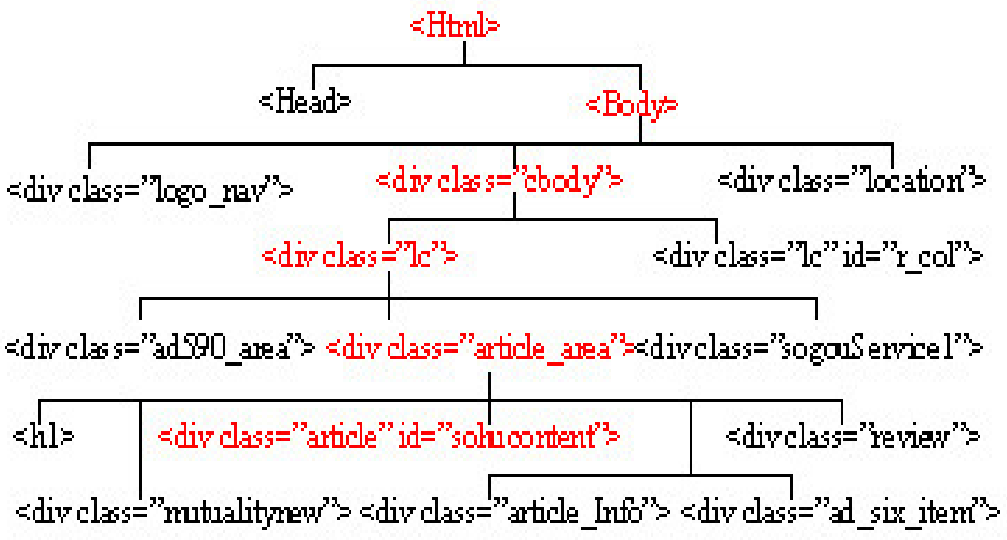

Fig. 2. The structure tree of the sample web page

Fig. 3. The result after extracting the sample web page

The result of content extraction is shown in Fig. 3. The size of original source page file is 92.4 k Bytes, which contains many pictures and advertising links. There is only less than 1 k Bytes after content extraction. This extraction is extremely necessary for enterprises and organizations to obtain information and knowledge from the Internet which contains a huge treasure of knowledge.

4.2 Test Data and Analysis

Experiments are performed using java platform –eclipse 3.2.1 and structure analysis tools - HtmlParser 2.0. By using baidu news search engine, we enter keywords and obtain pages from baidu news list.

The extraction is correct if the location node is correctly identified and the content is completely extracted (the content may contains a few links in the head or tail). While the incorrect results are mainly the following types:

1) The content is extracted incompletely.
2) Extraction is completely but not precise, i.e. including a large number of links.
3) The content is not extracted.

Table 1. The Content Extraction results of some central web sites

Web sites	Page Num	Correct Num	Wrong Num	Accuracy
Sohu	50	49	1	98 %
Sina	50	49	1	98 %
163	50	47	3	94 %
Yahoo	50	47	3	94 %
Xinhuanet	50	46	4	92 %
Peopledaily	50	47	3	94 %
Tom	50	46	4	92 %
China	50	47	3	94 %
Chinanews	50	46	4	92 %
21cn	50	48	2	96 %
Sum up	100	472	28	94.4 %

Table 1 shows the extraction results for some mainstream news websites. After randomly selecting different sites from baidu news search, the accuracy of our approach is nearly 95%, which meets the requirement of content extraction. However, some contents are not extracted correctly. This mainly dues to the fact that web page information is rather scattered, with all advertising and text in a node. There are too many nodes on the same level, and some web pages contain too many links. The threshold values set in the paper are not satisfied. In this case we can adjust the threshold values in terms of the specific websites.

The experimental system provides a useful way of collecting information for knowledge workers. Using different search engines, real-time news and information on the Internet will be obtained and the corresponding contents will be extracted. The extraction results could be used for content distribution and storage, as well as for the decisions of organizations.

V. CONCLUSIONS AND FUTURE WORKS

Web page information extraction technology is a current hot research issue. Many methods and techniques have been proposed, however, there is still not a method that meets the requirements of information collection in all kinds of fields. The algorithm based on the link density and statistic provides a solution for web page content stored in any multi-node, and resolves the problem of content extraction from the semistructured web page. Web page source files are filtered by regular expressions, and contents are extracted from structure trees directly. The above

process can reduce quantity of data transmission and complexity. These experiments show that the approach is universal and applicable for most of the existing news websites nowadays. The algorithm is suitable for data collection workers and other professionals. A lot of work need to be done before the use of search engine interface, such as concept retrieval and the expansion of semantic and synonyms, which are necessary for further investigations.

REFERENCES

[1] Shian Hua and Lin Jan Ming Ho, (2002) "Discovering Informative Content Blocks from Web Documents",In the proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD'02), Alberta, Canada, 588-593.

[2] Chang Yuhong, Jiang Zhe and Zhu Xiaoyan, (2004) " Web Page Structure Analysis Based on Tag Tree Method", Computer Engineering and Applications, 129-132.

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

[4] Sun Chengjie and Guan Yi, (2004) " A Statistical Approach for Content Extraction from Web Page", Journal of Information Processing, 18(5):17-22.

[5] Zhao Xinxin,Suo Hongguang and Liu Yushu, (2007) "Web Content Information Extraction Method Based on Tag Window. Application Research of Computers 2007 24(3):144-145.

[6] WangQi,Tang Shiwei and YangDongqing, (2004) "*DOM-Based Automatic Extraction of Topical Information from Web Pages",* Journal of Computer Research and Development,41(10): 1786-1792.

[7] Cui Jixin,ZhangPeng and Yang Wenzhu, (2005) "*DOM based Web information extraction",* Journal of Agricultural University of HeBei,28(3):90- 93.

[8] Gupta S, Kaiser GE, Neistadt D Grimm P, (2003), "*DOM based Content Extraction of HTML Documents",* in the proceeding of the 12th World Wide Web conference(www 2003),Budapest, Hungary, 207-214.

[9] Song Ruihua and Ma Shaoping,(2003) "*A HTML Parser to Improve Chinese Search Engine",* Journal of Chinese Information Processing, 17(4): 19-26.

Authors

R. Gunasundari is presently working as a Assistant Professor in the Departmnet of MCA, Karpagam University, Coimbatore. She has seven years teaching experience. She has participated and presented ten papers in national and three papers in International conferences.Currently she is working in the area of data mining and its application to web mining.

Dr. S. Karthikeyan received the Ph.D. Degree in Computer Science and Engineering from Alagappa University, Karaikudi in 2008. He is working as a Professor and Director in School of Computer Science and Applications, Karpagam University, Coimbatore. At present he is in deputation and working as Assistant Professor in Information Technology, College of Applied Sciences, Sohar, Sulatanate of Oman. He has published more than 14 papers in Natrional/International Journals. His research interests include Cryptography and Network Security.