Web Content Information Extraction Based on DOM Tree and Statistical Information

Xin Yu
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: 1090771320@qq.com

Zhengping Jin
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: zhpjin@bupt.edu.cn

Abstract—**Booming web pages contain a lot of information, while they contain little content and much unrelated noise information, such as script code, links, advertising and so on. These unrelated noise information occupies a lot of space, which is not suitable for the transition to small mobile devices, data mining and information retrieval. Therefore, web information extraction technology becomes more and more important. However, most extraction methods cannot adapt various and heterogeneous web structure and have poor generality and extracting efficiency. In this paper, we propose a method which can adapt to the heterogeneity and variability of web pages and gets high precision and recall. Our method is based on DOM structure to divide one web page into several blocks, and extract content blocks with statistical information instead of machine learning repeating training and manual labeling, which gets a good performance in Precision, Recall and F1.**

Keywords-content information extraction; DOM structure; content blocks; statistical information

I. INTRODUCTION

With the development of the network, a large number of information has occupied the web page in a very fast speed. As we can see, the network has become one of the world's largest data sources, which contains lots of valuable information. Meanwhile, irrelevant information such as advertising and web site navigation is becoming more and more and even occupied half of the web page, which makes it difficult to position and extract content information, and the format of the web page differs in thousands ways, and changes constantly, how to guarantee the versatility and adaptability of extraction technology is a huge challenge.

Web information extraction technology mainly includes the extraction technique based on the wrapper, the DOM tree structure, the page segmentation, and the statistical information. The technique based on the wrapper needs a large number of training samples and manual annotations, and when the web page structure changes, it should be labeled new training sets and manual annotations, which make automatic degree is not high. The content extraction technology based on the page segmentation aims to divide web page into several blocks [1-6]. Now page segmentation technology is mainly based on the visual information [1, 2] and specific tags [3-5]. The method based on visual information needs to download and parse a lot of style information, which is time consuming, and the use of visual heuristic rules is relatively fuzzy, people need to continue to summarize rules. The method based on specific tags cannot be used on all webpage, which makes its generality is poor. The extraction based on the wrapper is mainly to extract specific or similar web pages [7-9]. They need a large number of training samples and manual annotations, and when the web page structure changes, it should be labeled new training sets and manual annotations, which make automatic degree is not high. The content extraction technology based on statistical information is to count the information contained in nodes listed in the web page to extract the content information. Debnath et al. [10] selected the content blocks of text, images, and script codes with certain characteristics to sum up the FE (Feature Extractor) algorithm. Gottron et al. [11] proposed the CCB (Content Code Blurring) algorithm which is to select the same format of the source code characters through the content block. Weninger et al. [12] proposed the CETR (Content Extraction via Tag Ratios) algorithm, which took the text-label ratio of the HTML document line and the rate of change of the neighboring rows. There are also some methods which took text density as the basis for measuring the content characteristics [13-16]. Song et al [15] proposed text density method CETD (Extraction with Text Density), took the number of text and the number of tags as the text Density, but only text and link text information as measure, which would ignore the hierarchy structure that contained thematic characteristics.

In order to adapt to the web page heterogeneity and variability, and to improve the accuracy and generality of extraction, in this paper, we put forward a method named CEDS, which is based on DOM structure and statistical information that can ensure the generality and effectiveness of extraction. First we divide the webpage into some blocks according to the DOM structural relations about the correlation and similarity between the elements, and then we distinguish the information blocks according to the thematic feature statistics, removing the noise information block and leaving the content information block. Last but not least, we remove the noise information in the content information block based on some structural information. Our methods get 97.90% for Recall, 98.54% for Precision and 98.21% for F1, which is higher than existing methods.

978-1-5090-3944-9/17/$31.00 ©2017 IEEE

II. Methods

Nowadays, the news webpage is divided into two types: thematic news pages and picture news pages. Thematic news pages have a large number of text with only a small number of pictures for the auxiliary role, on the contrary, in the picture news pages, the pictures play as the main display, the text only explains the description of the pictures. In this thesis, our object is the thematic news pages.

A. DOM

The Document Object Model (DOM) is a platform and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents, recommended by the W3C organization [17].

B. Page Segmentation

We will refer to the DOM structure rules of the webpage to divide pages into several information blocks. Before the start of the segmentation operation, we have to do some pre-processing work.

1) Add hierarchical attribute for each node, which we call level. The hierarchical information could enrich node structure information, which has many relationships with content information of news webpage. We can see the hierarchical information in the Figure 1. The deeper the DOM structure is, the larger the node level is.

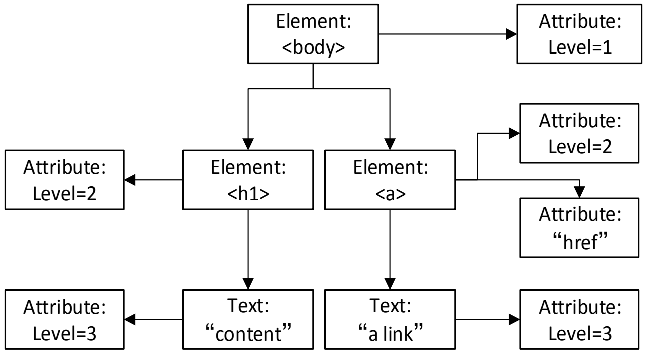

Figure 1. Example of a DOM tree with level.

2) Traverse the DOM tree to get the leaf node as the basic object set. According to the DOM tree structure in Figure 1, the leaf node is the display atomic object of the page, and the text information is all stored on the leaf node.

3) Get the real node for extraction. The simple leaf nodes cannot fully reflect the DOM structure sometimes. For example, a leaf node is wrapped in some layers of decorations, and its parent node only has one child node, it means that these decorations are used to decorate only a leaf node. In a result, this parent node can represent this leaf node. As is shown in Figure 2, <li> node can represent the text leaf node. We call the method as GetRealNode(node), and we can see it in the Algorithm 1.

```
<ul level="17">
    <li level="18"><a level="19">Bagehot’s notebook</a></li>
    <li level="18"><a level="19">Buttonwood’s notebook</a></li>
    <! --omit the nodes with the same structure of above-->
    <li level="18"><a level="19">Speakers’ Corner</a></li>
    <li level="18"><a level="19">The Economist explains</a></li>
</ul>
```

Figure 2. A brief segment of HTML code

4) Fusion based on least common ancestor. We find that the content has similarity in the DOM structure, so nodes with structural similarity should be assigned to the same block. It is shown in Figure 2, the <li> nodes are structurally similar and linked with each other in content level, so they should be assigned to a block.

```
Algorithm 1 fuseByCommonAncestor (leafNodeList, root)
/*isCommonLevel (N1, N2) is to determine if N1 and N2 has the same level, if so return true, else return false*/
/* isBrother(N1, N2) is to determine if N1 and N2 is slibing nodes, if so return true, else return false*/
/* getLCA (N1, N2) is to get the least common ancestors node for N1 and N2*/
INPUT: LeafNodeList and Root Node
OUTPUT: Node Block
begin
    for every adjacent leaf node N1 N2 in LeafNodeList do
        N1 ← GetRealNode(N1)
        N2 ← GetRealNode(N2)
        isCommonLevel ← isCommonLevel (N1, N2)
        isBrother ← isBrother(N1, N2)
        parent ←getLCA (N1, N2)
        if isCommonLevel && isBrother && parent != root
            addIntoBlock(parent)
        end
    end
end
```

The leaf nodes on the DOM structure have the visual continuity in the page display, and the nodes with content correlation have similarity in the DOM structure, such as the same node level, sibling node relationship and non-rooted least common ancestor node, and if the adjacent leaf nodes meet the above three conditions, they are fused into a block. See Algorithm 1 for details.

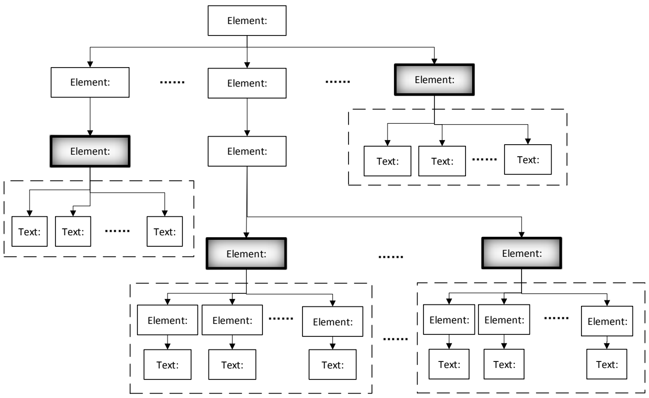

Figure 3. DOM tree after page segmentation

After the above operation, the leaf node in the DOM structure is divided into some information blocks, which is divided by dashed lines in Figure 3. The study object has changed from many leaf nodes into a smaller number of intermediate nodes.

C. Statistical Information

We select an article from Sina news network (http://news.sina.com.cn/) as an example, which has been divided into blocks used the above method. After a large number of case studies, we found that the content information and noise information has some differences. The content has more plain texts, less hyperlinks and less modification, the noise has less plain texts, more hyperlinks and more modification.

$$\text{ContentImportance}_N = \sum_{i=0}^{n} \frac{textLength_i - linkTextLength_i}{textLength_i} * \frac{1}{(textLinkCount_N + imgLinkCount_N)} * \frac{1}{typicalLevel_N} \quad (1)$$

- i: the order of each node in the block
- $textLength_i$: the length of text in the node i
- $linkTextLength_i$: the length of hyperlink text in the node i.
- $imgLinkCount_N$: the number of hyperlink images in the node i
- $typicalLevel_N$: the level which can represents the level structure information of Node N

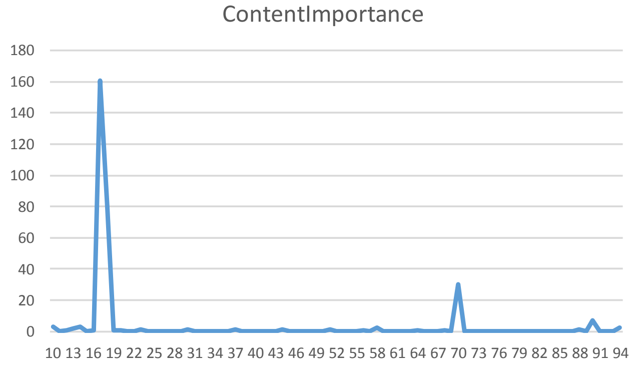

Figure 4. ContentImportance of every block

Each node has the level attribute which has been marked in the above section. For each information block, it has several nodes, we need to obtain a most representative level of the block. After lots of analysis, we found that: the largest number of node level is the most common level, and we select the maximum level when there are several such levels.

Through the Figure 4, it is found that the ContentImportance can effectively expand the difference between the content information and the noise information.

D. Threshold

The purpose of the threshold in this thesis is to divide the information block into two classes. The variance is an indicator to describe the difference between the classes, the larger the variance is, the more differences between the classes. After the above analysis, we take the variance as the basic principle.

$$\mu = \sum_{i=1}^{N} x_i \ * \frac{1}{N} \quad (2)$$

$$s^2 = \sum_{i=1}^{N} (x_i - \mu)^2 \quad (3)$$

$$\tau = \varepsilon\mu \quad (4)$$

$$\tau = arg \max s^2 \quad (5)$$

The purpose of the threshold in this thesis is to divide the information block into two classes. The variance is an indicator to describe the difference between the classes, the larger the variance is, the more differences between the classes. After the above analysis, we take the variance as the basic principle.

We find the τ in an iterative way, making the largest variance between the two classes. Equations 2-4 represent the average, variance and threshold. N is the number of classes, x_i is the ContentImportance value for each lass, μ is the average value, s^2 is the variance, τ is the threshold, ε is the threshold coefficient, and after analysis, the range is in [1, 3], the iterative step is 0.1, and the threshold of the standard deviation is maximized by the iterative method. The formula is as follows.

After the threshold is obtained, it is determined that the information blocks which is larger than the threshold value is the content block, the others are the noise block.

E. De-noising from Content Block

The subject information block after threshold selection is also likely to contain some noise information, such as some comments will be caught in news content information block. After lots of analysis, the de-noising algorithm is obtained. If the level distance between the adjacent nodes is larger than four, the ancestor node of the adjacent leaf nodes is not the root of the content information block and the level of these leaf nodes is not the common level, these nodes are noise node existing in the content information block.

III. Evaluation

The data set used in this thesis is derived from the recent collection of thematic news pages that are randomly obtained through reptiles. The basis of comparison and the evaluation is obtained by manually.

$$Precision = \frac{LCS(A,B).length}{A.length} \quad (6)$$

$$Recall = \frac{LCS(A,\ B\ ).length}{B.length} \quad (7)$$

$$\text{F1} = \frac{2PR}{P+R} \quad (8)$$

The performance of our content extraction system is based on three evaluation indicators: Recall and Precision and F1. Recall is equal to the ratio of the correct results of the system to all possible correct results. Precision is equal to the ratio of the results of the system's correct extraction to the total of the extracted results. F1 is used to average Recall and Precision performance.

In the Equations 6-8, A is the sequence of the content information obtained by the method we used in this thesis, B is the content information obtained by manual means, LCS (A, B) is the longest common subsequences in the two information sets, we can see the experiment result in the Table I.

TABLE I. EXPERIMENT RESULT (P FOR PRECISION, R FOR RECALL)

Data Source	Website	P	R	F1
The Economist	http://www.economist.com/	98.12%	96.34%	96.98%
Sina News	http://news.sina.com.cn/	99.29%	98.16%	98.49%
Global News	http://www.huanqiu.com/	97.33%	98.41%	97.49%
Sohu News	http://www.sohu.com/	96.58%	99.12%	97.47%
Firefox News	http://www.firefoxchina.cn/	97.17%	99.82%	97.98%

In this paper, we compare the existing methods with our method named CEDS in the Table II.

TABLE II. EXPERIMENT RESULT COMPARED WITH OTHER METHODS

Method	Precision	Recall	F1
FE	74.19%	60.21%	66.47%
CCB	81.27%	90.34%	85.57%
CETR	86.98%	93.70%	90.22%
CETD	94.49%	94.58%	94.53%
CEDS	97.90%	98.54%	98.21%

Through the comparison, we can see that our method gets 97.90% for Recall, 98.54% for Precision and 98.21% for F1, which are higher than other methods, proving its effectiveness.

IV. CONCLUTION

This paper aimed at adapting the heterogeneity and variability of the web page, and improving the accuracy and versatility of the extraction. We based on the DOM structure to divide one original web page into multiple information blocks, and extracted the content information blocks by the statistical principle. Finally, we de-noised from content blocks to complete the extraction work. Our methods do not need manual labeling, and without machine learning repeating training, and our methods proved that it has also been a good effect in Precision, Recall, and F1 value.

ACKNOWLEDGMENT

This work is supported by NSFC (Grant No. 61502044), and the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

REFERENCES

[1] Cai D, Yu S, Wen J, et al., VIPS: a vision based page segmentation algorithm. Technical report, Microsoft Technical Report,MSR-TR-2003-79, 2003.

[2] Xu Z, Miller J. Identifying semantic blocks in Web pages using Gestalt laws of grouping[J]. World Wide Web, 2015: 1-22.

[3] Lin S, Ho J. Discovering informative content blocks from web documents. In: Proceedings of SIGKDD'02. New York, NY, USA, 2002: 588-593.

[4] Carey H J, Manic M. HTML web content extraction using paragraph tags[C]//Industrial Electronics (ISIE), 2016 IEEE 25th International Symposium on. IEEE, 2016: 1099-1105.

[5] Shen W, Zou X. An Algorithm on Web Article Automatic Extraction Based on DOM Structure[J]. International Journal of Hybrid Information Technology, 2015, 8(3): 243-254.

[6] Bing L, Guo R, Lam W, et al. Web page segmentation with structured prediction and its application in web page classification[C]//Proceedings of the 37th international ACM SIGIR conference on Research & Development inInformation Retrieval. ACM, 2014: 767-776.

[7] Annam M, Sajeev G P. Entropy based informative content density approach for efficient web content extraction[C]//Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. IEEE, 2016: 118-124.

[8] Gibson J, Wellner B, Lubar S. Adaptive Web-page Content Identificatio// Proc of the 9th ACM International Workshop on Web Information and Data Management. Lisbon, Portugal, 2007: 105-112.

[9] Ziegler C N, Skubacz M. Content Extraction from News Pages Using Particle Swarm Optimization on Linguistic and Structural Features//Proc of the IEEE / WIC / ACM International Conference on Web Intelligence. Fremont, USA, 2007: 242-249.

[10] Debnath S, Mitra P, Giles C L. Automatic Extraction of Informative Blocks from Webpages / / Proc of the ACM Symposium on Applied Computing. Santa Fe, USA, 2005: 1722-1726.

[11] Gottron T. Content Code Blurring: A New Approach to Content Extraction / / Proc of the 19th International Conference on Database and Expert Systems Applications. Turin, Italy, 2008: 29-33.

[12] Weninger T, Hsu W H, Han Jiawei. CETR-Content Extraction via Tag Ratios / / Proc of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 971-980.

[13] T. Weninger and W. H. Hsu, "Text extraction from the web via text-to-tag ratio," in 2008 19th International Workshop on Database and Expert Systems Applications. IEEE, 2008, pp. 23–28.

[14] Sun F, Song D, Liao L. DOM based content extraction via text density.In: Proceedings of the 34th International ACM SIGIR Conference. Beijing, China, 2011: 245-254.

[15] Song D, Sun F, Liao L. A hybrid approach for content extraction with text density and visual importance of DOM nodes[J]. Knowledge and Information Systems, 2015, 42(1): 75-96.

[16] Annam M, Sajeev G P. Entropy based informative content density approach for efficient web content extraction[C]//Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. IEEE, 2016: 118-124.

[17] W3C Document Object Model (2012) Website. https://www.w3.org/DOM/.