Extracting News Information Based on Webpage Segmentation and Parsing DOM Tree Reversely

Jing Li[1,2(✉)], Yueming Lu[1,2], and Xi Zhang[1,2]

[1] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
{lj_2013, ymlu, zhangx}@bupt.edu.cn
[2] Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing, China

Abstract. A new method of extracting news information based on webpage segmentation and parsing DOM tree reversely is presented and implemented in this paper, which intends to effectively extract news information for data mining. The method is proposed to get webpages' main DOM structure by segmenting webpages, further parse the main DOM structure reversely and finally extract news content, headlines, news agents and publication time. The experimental results show that the proposed method has achieved good performance on accuracy and meets the project demands.

Keywords: Webpage segmentation · Parsing DOM tree reversely · Extracting news information · Main DOM structure

1 Introduction

With increasing development of data mining and search engine technologies, the extraction technologies on webpage information including content, headlines, news agents and publication time become an increasingly hot topic. The information is not only a momentous corpus in the analysis of the net-mediated public sentiment, but also plays a vital role in the area of search engine technologies such as establishing indexes and removing duplicated web pages.

It is a phenomenon that in addition to this valuable information, a plenty of rubbish information consisting of advertisement, hyperlinks, copyright information can also be available at news webpages. Therefore, domestic and foreign researchers have carried much work of investigation and study on the webpage information extraction, and made considerable achievements.

Ying Bin and Yang Huizhi parse webpages into DOM trees, and then estimate webpage content judging by the weight of text characters and the density of link characters of each node in DOM trees in Ref. [1]. Zou Yongqiang and Zhong Zhinong locate webpage content depending on the news webpage features and the statistical regularity of the text blocks in Ref. [2]. However, the methods in the two mentioned Ref. [1, 2] are not suitable for the webpages with short content or continuous

Y. Lu et al. (Eds.): ISCTCS 2014, CCIS 520, pp. 48–55, 2015.
DOI: 10.1007/978-3-662-47401-3_7

hyperlinks. Chen Hansheng, Zeng Jianping and Zhang Shiyong restore each tag's display position of HTML documents in browser window by simulating part of the rendering process that web browser does, and then segment webpages in Ref. [3]. But the complexity is much higher.

But most of the existing information extraction technologies only extract news content without news headlines, news agents and publication time. Taking into account the fact that news agents, publication time and other information also have critical effect on the analysis of the net-mediated public sentiment, we put forward a method mixing webpage segmentation [4] with parsing DOM tree reversely [5] to extract news content, headlines, news agents and publication time from webpages.

2 Extracting News Information Based on Webpage Segmentation and Parsing DOM Tree Reversely

The method is to get webpages' main regions by segmenting webpages and then extract important news corpuses containing news content, headlines, news agents and publication time by parsing DOM tree reversely. It is more efficient and accurate to extract news information on the basis of webpages' main regions.

2.1 Webpage Segmentation

After analyzing news webpages from Sina, Tencent, Phoenix New Media,[1] Sohu and NetEase, it can be found that a news webpage is made up of four parts: the head, the foot, the left and the right. Generally, menus should be put on the head, and copyright information should be put on the foot. There are mostly a set of recommended news links and video links in the right. Ordinarily, webpages put their main parts on the upper-left half position and their related comments, pictures and links on the lower-left half position. Figure 1 demonstrates the layout of a news webpage from Sina.

Moreover, it is common that in webpage source codes, two CSS properties, which are *float: left* and *width: value* in CSS stylesheets, are used to add special effect to news webpages' left parts. And in a news webpage, the left div block is the only one with the two feature properties at the same time. Significantly, the *value* is determined as a pixel value over half of screens' width in the experiment. In this paper, the two CSS properties are defined as the feature properties.

Inspired by this, we design a flow chart showed in Fig. 2 to segment webpages. Taking a news webpage as an example, the detailed steps are as follows:

(1) We use JSOUP to parse an HTML document firstly. JSOUP is a JAVA HTML parser. It can parse HTML from a URL, a file or a string, and provides an extremely efficient API. Its most prominent advantage is that JSOUP is as powerful as the JQuery selector. What's more, JSOUP can correctly deal with non-standard HTML tags such as non-closed tags.

[1] http://www.ifeng.com

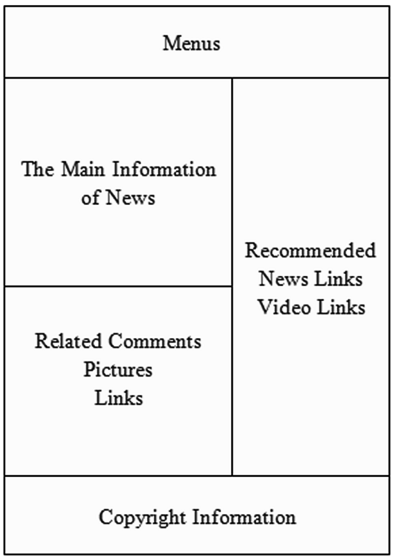

Fig. 1. The layout of a news webpage from Sina

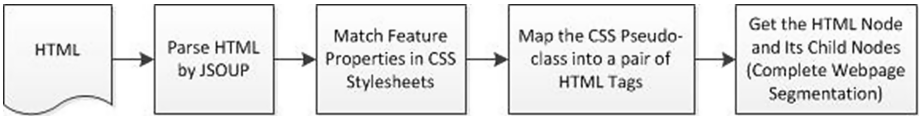

Fig. 2. The process of webpage segmentation

(2) We need to parse CSS stylesheets to search the two feature properties. As long as the two feature properties are matched successfully, we can stop parsing CSS stylesheets, record the CSS pseudo-class which contains the two feature properties and enter into the step (3). Otherwise, if there isn't a CSS pseudo-class with the feature properties to be matched until CSS stylesheets are traversed over, we output the whole DOM tree as the main region regardless of the step (3) and (4).
(3) Allowing for the fact that the webpage segmentation in this paper is aimed to obtain webpages' left parts, we only need to map the CSS pseudo-class recorded in the step (2) into a pair of HTML tags, instead of getting a complete DOM tree by parsing CSS stylesheets and merging all CSS pseudo-classes with corresponding HTML tags [4]. As a result, the time-consuming can be reduced greatly.
(4) We can get the HTML node mapped in the step (3) and its child nodes as the main region to finish segmenting the webpage.

2.2 Extracting News Information Reversely

After obtaining the main regions of news webpages from Sina, Tencent, Phoenix New Media, Sohu and NetEase by means of the above webpage segmentation, we analyze these main regions' DOM structure, and draw a conclusion that the DOM structure has strong similarity. As illustrated in Fig. 3, the main features of the DOM structure are summarized below:

Feature 1. News headlines are in a pair of <h1></h1> tags, represented as the title node (TITLE).

Feature 2. News agents and publication time are usually in a div block, represented as the info node (INFO).

Feature 3. News content is represented as a single div block, named the content node (CONTENT). In its child nodes, each paragraph of news content is a child p-node, and those videos or pictures of news content are represented as other div blocks.

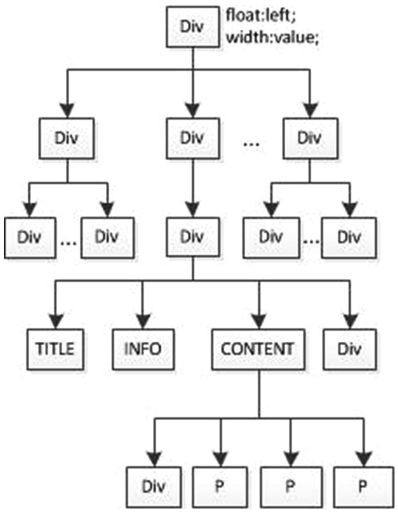

Fig. 3. The DOM structure of the main region of a news webpage

Inspired by this, this paper presents an idea of parsing DOM tree reversely. Firstly, we need to identify the content node (CONTENT) of a news webpage, and then search its headline node (TITLE), news agent and publication time node (INFO) reversely. Finally, the extracted results are text characters of CONTENT, TITLE and INFO. The detail steps are illustrated in Fig. 4 and are described below:

Identify CONTENT. In general, news webpage content is distributed in DOM trees' leaf nodes and wrapped in tags [5]. Given that the feature is not exclusive,

we can't well identify the difference between text nodes and non-text nodes by leaf nodes' tags. In the view of this, it is not practical to decide whether some nodes belong to news text nodes depending on whether they are leaf p-nodes.

As far as the above problem is concerned, we can firstly get rid of leaf p-nodes' tags and their non-p sibling nodes (some div blocks of videos, pictures or layouts) from the main region's DOM structure of a news webpage. In this way, all leaf p-nodes' content is merged and their original parent node is turned into a new leaf node, and then we replace the new leaf node's tags with <t></t>. At this moment, news content is just in one leaf node of the DOM structure. Inspired by this, judging by the length of text characters in <t></t> tags, we can identify that the t-node with the maximum length is CONTENT.

Identify TITLE. Statistical analysis shows, TITLE and INFO are usually CONTENT's sibling nodes or child nodes of CONTENT's uncle nodes. If there is a node with <h1></h1> tags to be found in CONTENT's sibling nodes, we output the sibling node as TITLE. If there isn't, we traverse child nodes of CONTENT's uncle nodes until there is a node with <h1></h1> tags to be found. As a result, the found node is TITLE.

Identify INFO. Analogously, we can use regular expressions to match a time-formats node, while traversing CONTENT's sibling nodes or child nodes of CONTENT's uncle nodes. Finally, the node matched is INFO.

3 Experiment and Evaluation

In order to evaluate the proposed method, we select 2941 news webpages from five typical news websites including Sina, Tencent, Phoenix New Media, Sohu and NetEase in the validation experiment. We use JAVA to implement the proposed method. By manually counting the number of the correctly segmented webpages and the number of the correctly extracted webpages separately, the accuracies of segmenting the 2941 webpages and extracting their content, headlines, news agents and publication time are showed in Fig. 5. It demonstrates that the average accuracy of segmenting those webpages reaches 93.13 %, and the average accuracies of extracting their content, headlines, news agents and publication time are 92.28 %, 92.19 % and 89.56 % separately with the proposed method.

It should be noted that the performance of the webpage segmentation is satisfactory only in the condition that what we pick out for the experiment are news webpages from the five typical news websites. By comparison with segmenting those webpages, the accuracies of extracting news content and headlines are on the low side slightly due to the uncertain distribution of news webpage content and headlines. In addition, there are a series of webpages without news agents and publication time, leading to a little lower accuracy of extracting news agents and publication time.

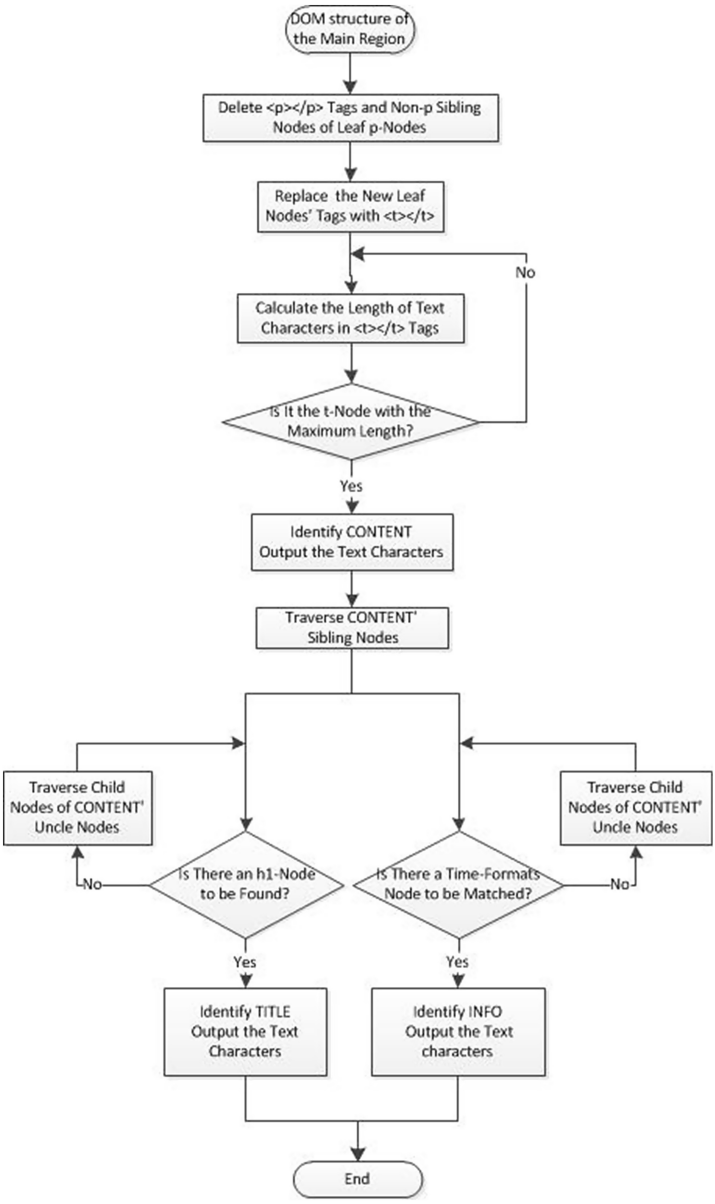

Fig. 4. The process of extracting news information by parsing DOM tree reversely

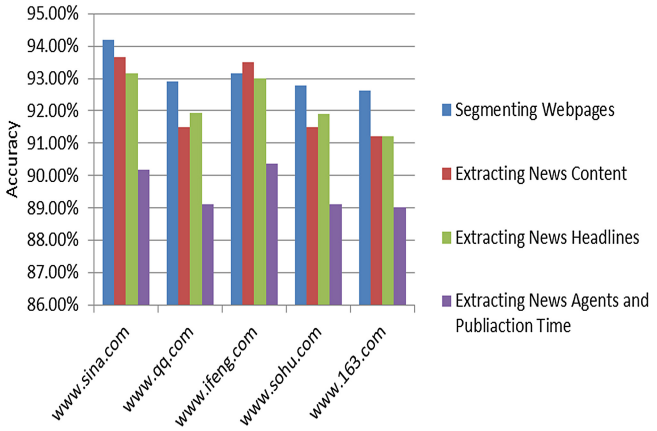

Fig. 5. The accuracies of segmenting webpages and extracting news information for different news websites

4 Conclusion

The method of extracting news information based on webpage segmentation and parsing DOM tree reversely is presented and implemented in this paper. In regard to those webpages from the five typical news websites including Sina, Tencent, Phoenix New Media, Sohu and NetEase, the method has lower complexity and higher accuracy in segmenting news webpages and extracting news information. However, the method is presented in accordance with the project about the analysis of the net-mediated public sentiment, so the universality of the method is insufficient, and it is not recommended to apply the method to extract information from non-news webpages. We will set out to extract news videos and pictures from news webpages [6] in the future work.

Acknowledgement. This work was supported by the Major Research Plan of the National Natural Science Foundation of China [91124002] and the Fundamental Research Funds for the Central Universities [2013RC0301].

References

1. Yin, B., Yang, H.Z.: Content extraction based on unknown structure web. Comput. Technol. Dev. **21**(9), 111-113, 117 (2011)
2. Zou, Y.Q., Zhong, Z.N.: An efficient approach to reduce noise in news webpages. Micro-comput. Appl. **30**(16), 64–67, 71 (2011)
3. Chen, H.S., Zeng, J.P., Zhang, S.Y.: A position information-based web page segmentation method. Comput. Appl. Softw. **26**(7), 155–159 (2009)
4. Zhang, R.X., Song, M.Q., Gong, Y.L.: Parsing DOM tree reversely and extracting web main page information. Comput. Sci. **38**(4), 213–215, 225 (2011)

5. Li, J., Chen, J., Wang, L.F., Ni, H.: Approach to webpage segmentation and information extraction for vertical websites. Appl. Res. Comput. **30**(3), 844–847, 852 (2013)
6. Jia, J., Zhang, S., Meng, F., Wang, Y., Cai, L.: Emotional audio-visual speech synthesis based on PAD. IEEE Trans. Audio, Speech, Lang. Process. **19**(3), 570–582 (2011)