An Effective Method to Extract Web Content Information

Pan Suhan, Li Zhiqiang*, Dai Juan
College of Information Engineering, Yangzhou University, Yangzhou, China.

* Corresponding author. Tel.: +8613952751283; email: yzqqLzq@163.com
Manuscript submitted September 10, 2018; accepted November 20, 2018.
doi: 10.17706/jsw.13.11.621-629

Abstract: To simplify the operation of web text content extraction and improve the accuracy of that, a new extraction method based on text-punctuation distribution and tag features (TPDT) is proposed. Combining the distribution of text-punctuation and tag features. Calculating the text-punctuation density in different text blocks and get the maximum continuous sum of density to extracting the best text content from web pages. The method effectively solves the problem of noisy information filtering and text content extraction without the training and manual processing. Experimental results on web pages randomly selected from different portal websites show that the TPDT method has good applicability on various news pages.

Keywords: Content extraction, text density, punctuation distribution, tag tree.

1. Introduction

With the rapid development of web technology, web pages have become the main carrier of information dissemination. According to the latest global digital report of 2018 "We Are Social and Hootsuite", the number of Internet users using the Internet has exceeded 4 billion. One of the main behaviors of Internet users is to read online news. Since web technology develops rapidly and the scale of the Internet is larger in these years, the Internet users have higher requirements for the quality of news media. Various kinds of Web news have replaced the traditional paper news and now become the main way for people to get news hotspots.

In addition to the main content, the web page includes a large amount of noise information that is not related to the subject content, such as navigation information, advertisement text, recommended links, and copyright claims. At present, web development has emerged a lot of new technologies which make different web pages more diverse and personalized. This poses a huge challenge to the extraction of Web content information. And it turns out that the Web page structure has a potential relationship with its corresponding text, punctuation, and tag features. The specific performance is as follows:

- The main text content of the web page is concentrated and mostly in a text block, mostly long text, containing a small amount of hyperlink text, the text block density is large, and the punctuation is densely distributed.

- The noise of the Web page is scattered, mostly short text, contains a large amount of hyperlink text, the text block density is relatively small, and the punctuation is sparsely distributed.
- The content of the text in the web page is mostly distributing in the same type of HTML tag. The main content nodes are always continuous, and all content nodes share the same parent node.

2. Related Work

In recent years, researchers at home and abroad have put forward a lot of methods of web information extraction. The information extraction bases on rule template such as the researches [1]-[3]. For the same web site or a site section, the pages have the same or similar structure, by obtaining the design template and generate the corresponding information extraction wrapper, we can accurately obtain the subject information of the data source. Weninger [4] proposed a Content Extraction via Tag Ratios (CETR) method that calculates the Text-To-Tag Ratio of each line in the web page as a feature, and then clusters texts to content and non-content. Sun F [5] proposed a web information extraction algorithm based on web page text density and feature (CETD). A method based on the observation that the content text is usually lengthy and simply formatted, while noise is usually highly formatted and contain less text with brief sentences. Observing that noise contains many more hyperlinks than meaningful content, they extended Text Density to Composite Text Density by adding statistical information about hyperlinks. Wu G [6] proposed web news extraction via path ratios. The algorithm is based on the text tag path ratio feature and its extended features, and effectively distinguishes the body content and noise content in the webpage through the establishment of the threshold. Shuang Lin, *et a*l. [7] proposed an approach of content extraction based on density. The proposed system is responsible to extract contents from web pages which are used to obtain page contents that are crucial to many web mining applications. Kai S, *et al*. [8] created web content extraction based on maximum continuous sum of text density which is on the basis of a statistical analysis on content features and structure characteristics of News domain web pages to efficiently and effectively extract web content from different web pages. Jian S, *et al.* [9] proposed an efficient method for extracting web news content that using the idea of statistics and web structure. There are also a number of studies based on the similar algorithms. Xin Yu *et al*. [10] proposed a way of web content information extraction based on DOM tree and statistical information. Mingdong Li *et al*. [11] proposed content extraction based on statistic and position relationship between title and content. Manjusha Annam, *et al*. [12] proposed entropy based informative content density approach for efficient web content extraction. The key of the algorithm is to utilize the information entropy for representing the knowledge that correlates to the amount of informative content in a page.

According to statistics, no algorithm has managed to achieve 100% accuracy in extracting all relevant text from websites so far. With the ever changing style and design of modern web pages, different approaches are continually needed to keep up with the changes. Some algorithms just work on certain websites but not on others. There is still much work left to be done in the field of website content extraction [13].

3. Description of Our Proposal

The process of the main contents extraction is shown in Figure 1. The pre-processing step starts with the original website's HTML source extracted from the web page. Then, parsing the web page source code via Beautiful Soup into a DOM structure. The calculation of characteristics section analyses the web page features to determine the best text block. The final text is then extracted from the best t ext block.

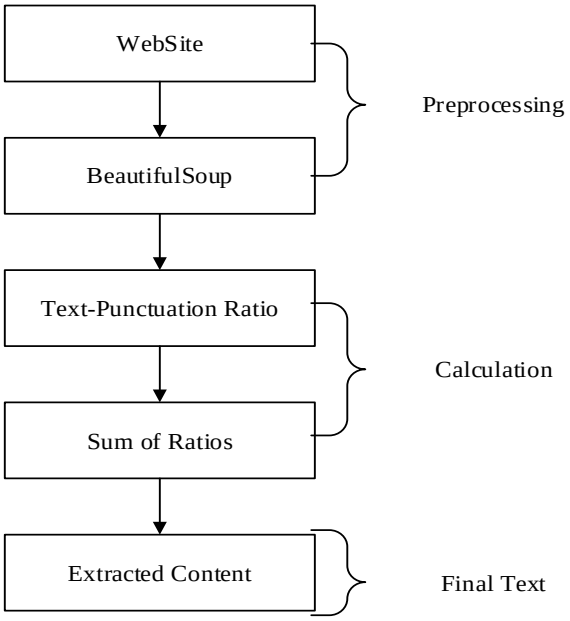

Fig. 1. Flow-diagram of the presented TPDP method.

3.1 Pre-Processing

DOM is a common tool to represent web pages. In DOM, web pages are represented as a set of tags and the hierarchical relationship between these tags. According to the function of each tag, we classify the HTML tags into four categories:

1) Tags for interaction. This kind of tag allows users to create a message and interact with the server or the presupposed functions in the page. The commonly used interaction tags contain <script>, <noscript>, <input>, etc.

2) Tags for style. This kind of tag is used to personalize the style of the content on the page. Typical tags include: <b>, <i>, , etc.

3) Tags for metadata description. This kind of tag is used to describe the metadata and basic attributes for the whole web page. Typical tags are <head>, <title>, <meta>, <link>, etc.

4) Tags as a container. This kind of tag is always used to arrange the content of a web page. In general, the core content is always put in these tags. Frequently-used tags contain , <table>, , <p>, , etc.

In this study, our target is extracting the main content from a web page. So, we will construct the DOM according to the container-kind tags. Other kinds of tags such as tags for interaction, tags for

style and tags for metadata will be blocked by filters at the pre-processing step [14].

3.2 The Calculation Method of the Characteristics

Observing the example webpage of Fig. 2 (the main text content of the webpage is in the rectangular box). According to the observation, we can find that the main text content of the web page is concentrated text, more punctuation distribution and less hyperlinks. However the noise content has short dispersed text with little punctuation and more hyperlinks.

Fig. 2. Sample web page.

Definition 1 text block. TB(v) is a subtree with node v as the root node, where v is a non-text node, and if TB(v) is not empty, then subtree TB(v) is called a text block.

As shown in Figure 3 and Figure 4 are different text block extracted from the web page. Then we can find that the text block has many concentrated text with more punctuation distribution and few hyperlinks. However, the noise block has short text with sparse punctuation distribution and more hyperlinks. Based on the discovered features, we set the following property values:

Property 1. Text length (TL). The number of characters in each HTML tag in the block.

Property 2. Punctuation length (PL). The number of punctuation marks in each HTML tag in the block.

Property 3. Sum of text punctuation density (TPD).

$$\mathrm{TPD} = \frac{PL}{TL + 1}$$

Property 4. Link text length (LTL). The number of characters in each HTML link tag in the block.

Property 5. Link punctuation length (LPL). The number of punctuation marks in each HTML link tag in the block.

Property 6. Sum of link text punctuation density (LTPD).

$$\mathrm{LTPD} = \frac{LPL}{LTL + 1}$$

Property 7. Continuous sum of the block density. (TBD).

$$\mathrm{TBD} = \sum_{tag \in textblock} \frac{PL}{TL+1} - \sum_{linktag \in textblock} \frac{LPL}{LTL+1}$$

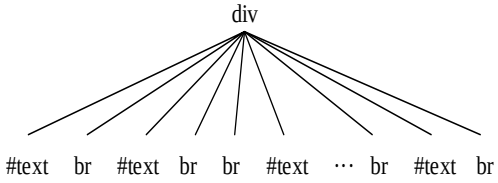

Fig. 3. DOM tree of the text block.

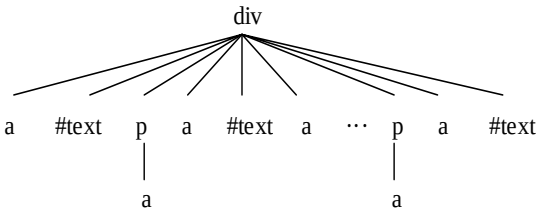

Fig. 4. DOM tree of the noise block.

According to the calculation properties mentioned above, it is easy to find the best text block from the web page. Then, we can extract main content from the best text block more efficiently and quickly.

4. Main Content Extraction

To extract the main content from the web page. We give the main algorithm TPDP proposed in this paper. The pseudocode of the main algorithm is as follows:

Algorithm TPDT:

```
Input: HTML web page(P)
Output: the main content
P = BeautifulSoup(P);
P = filter_tags(P);
for tag in P do:
    blockset = getblock(tag);
for eachblock in blockset do:
    for tag, linktag in eachblock do:
        TL = calculate_TL(tag);
        LCL = calculate_LCL(linktag);
        PL = calculate_PL(tag);
        LPL = calculate_LPL(linktag);
        TPD += PL / (TL + 1);
        LTPD += LPL / (LTL + 1);
    TBD = TPD – LTPD;
maxblock = getmax(TBD);
content = maxblock.text;
```

14: Output content;

The web page text extraction algorithm TPDT based on the text punctuation distribution and the tag feature is given above. The first step uses Python's html parsing library Beautifulsoup to parse the input web page P; The second step of the algorithm is to filter the web page P and filter out the useless tags; step 3-4 loops through the filtered web page and divides the web page into multiple different blocks; step 5-12 loops through tags and link tags in each block, calculate the text length and punctuation length in each tag, and sum of the feature TPD and LTPD, which are values of text punctuation density; step 13 calculates continuous sum of text block density of block in the web page; step 14 obtaining the block corresponding to the largest TBD; step 15 extracts text content from the text block obtained in step 14; the final step is to output the main content extracted from the web page.

5. Experimental Results

In this experiment, we select 100 web pages randomly from different portals as test data sets. Three metrics are used to evaluate the performance of the TPDT, they respectively are precision, recall, and F score. These are standard evaluation techniques used to determine how accurate an extracted information is compared to what it is supposed to be.

Precision (P) is a ratio between the size of the intersection of the extracted content and the actual content and the size of the extracted text. Precision gives a measurement of how relevant the extracted content are to the actual desired content text and is expressed as:

$$\mathrm{P} = \frac{S_E \cap S_M}{S_E}$$

where SE is the extracted content from the algorithm and SM is the actual text extracted by manual annotation.

Recall (R) is a ratio between the size of the intersection of the extracted content and the actual content and the size of the actual text. Recall gives a measurement of how much of the relevant data was accurately extracted and is expressed as:

$$\mathrm{R} = \frac{S_E \cap S_M}{S_M}$$

where SE is the extracted content from the algorithm and SM is the actual text extracted by manual annotation that work as standard comparison for accuracy.

The F score (F) is a combination of precision and recall, allowing a single number to represent the accuracy of the algorithm. It is expressed as:

$$\mathrm{F} = \frac{2 \times P \times R}{P + R}$$

where P is the precision and R is the recall.

These evaluation standards analyse the content extracted from different algorithm and compare them to

the content in the web page itself. If the content extracted from the algorithm matches the actual content closely, the F score will be higher. Using these three scoring standards allow for a numerical comparison between different algorithms. F score is significant because only depending on precision or recall can be misleading. Having a high precision but low recall value means nearly all of the extracted content is accurate but it does not indicate any information on how much relevant information is missing. Having a high recall but a low precision value means nearly all of the content that should have been extracted is, but it also does not indicate any information on how much irrelevant content was also extracted. Therefore, the F score is important as it allows a combination of both precision and recall to determine a balance between the two.

Table 1. Comparison of Results of Different Algorithms

DataSet	CETD			CETR			TPDT		
	AveP	AveR	AveF	AveP	AveR	AveF	AveP	AveR	AveF
sina	81.98%	98.47%	89.47%	59.09%	98.94%	73.99%	96.97%	97.03%	97.00%
people.cn	69.79%	99.91%	82.17%	77.04%	97.89%	86.23%	95.83%	97.88%	96.85%
163	37.75%	89.93%	53.18%	24.40%	88.76%	38.28%	95.31%	96.85%	96.08%
xinhuanet	83.28%	99.25%	90.58%	72.10%	98.68%	83.32%	97.46%	96.33%	96.89%
Average	68.20%	96.89%	78.85%	58.16%	96.07%	70.46%	96.39%	97.02%	96.71%

From the comparison of the experimental results in Table 1, it is found that the TPDT method has good performance in web page extraction. The accuracy of TPDT extraction, recall rate and F value are better than algorithms CETD and CETR.

6. Conclusion

The algorithm proposed in this paper is mainly based on the content distribution and noise distribution characteristics of web pages. Through observation, it can be found that the distribution of the body content in the Web page is relatively concentrated, the format is single, the hyperlink text information is less, and the punctuation symbol distribution is dense. The web page text extraction algorithm TPDT based on the text punctuation distribution feature and the tag feature is realized. The experimental results of multiple data sets show that the extraction result of TPDT algorithm is satisfactory and the algorithm is a simple extraction algorithm without training, no human intervention, and can be used for multi-source, massive and heterogeneous web pages. Complete the extraction task accurately and simply.

However, there is still room for improvement in the web page text extraction algorithm TPDT based on text punctuation features and HTML tag features. In this paper, we just do that based on the sum of continuous tag features in the text block to judge the best text block. Whether it can be combined with other features introduced in the web page and improve the accuracy and applicability of the extraction algorithm is the direction of further research.

Acknowledgment

Many people have offered me valuable help in my thesis writing, including my tutor and my classmates.

Firstly, I would like to give my sincere gratitude to Prof. Li Zhiqiang, my tutor who, with extraordinary patience and consistent encouragement, gave me great help by providing me with necessary materials, advice of great value and inspiration of new ideas. Without his strong support, this thesis could not been the present form.

Finally, I pleased to acknowledge my classmates for their invaluable assistance throughout the preparation of the original manuscript. They graciously make considerable comments and sound suggestions to the outline of this paper.

References

[1] Uzun, E., Agun, H. V., & Yerlikaya, T. (2013). A hybrid approach for extracting information content from web pages. *Information Processing and Management*, *49(4)*, 928-944.

[2] Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2017). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*.

[3] Alarte, J. *et al*. (2014). A benchmark suite for template detection and content extraction. *Eprint Arxiv*.

[4] Weninger, T., & Hsu, W. H. (2008). Text extraction from the web via text-to-tag ratio. *International Workshop on Database and Expert Systems Application IEEE*.

[5] Sun, F., Song, D., & Liao, L. (2011). DOM based content extraction via text density. *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval ACM*.

[6] Wu, G. Q., et al. (2013). Web news extraction via path ratios.

[7] Lin, S., Chen, J., & Niu, Z. (2012). Combining a segmentation-like approach and a density-based approach in content extraction. *Tsinghua Science and Technology*.

[8] Sun, K. *et al*. (2016). Web content extraction based on maximum continuous sum of text density. *Proceedings of the International Conference on Asian Language Processing IEEE*.

[9] Sun, J., *et al*. (2017). An efficient method for extracting web news content. *Proceedings of the International Conference on Engineering and Technology*.

[10] Yu, X. *et al*. (2017). Web content information extraction based on DOM tree and statistical information. *Proceedings of the International Conference on Communication Technology IEEE*.

[11] Li, M. D., *et al*. (2014). Content extraction based on statistic and position relationship between title and content. *Proceedings of the Symposium on Social Networks and Big Data ICCC 2014*.

[12] Annam, M., & Sajeev, G. P. (2016). Entropy based informative content density approach for efficient web content extraction. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics IEEE*.

[13] Carey, H. J., & Manic, M. (2016). HTML web content extraction using paragraph tags. *Proceedings of the IEEE, International Symposium on Industrial Electronics IEEE*.

[14] Liu, Q. T. et al. (2017). Main content extraction from web pages based on node characteristics. *Journal of Computing Science and Engineering.*

Suhan Pan was born in Jiangsu province, China, in 1994. He graduated from the College of Information Engineering of Yangzhou University in 2017. Now, he is a master's student at the same university. His current research is about web data extraction, web data analysis and application.

Zhiqiang Li was born in Jiangsu province, China, in 1974, is a Prof. head of Department of Computer Science and Technology and national senior programmer. He went to the Department of Electronics and Computer Engineering of Portland State University in the United States. He visited the school as a visiting scholar for one year. He studied under the famous Professor Perkowski and achieved gratifying results in academic research. He was highly appraised by the professors. His current research is about web application, quantum circuit synthesis and web data analysis.

Juan Dai was born in Jiangsu province, China, in 1996. She graduated from the College of Information Engineering of Yangzhou University in 2018. Now, she is a master's student at the same university. Her current research is about web data extraction, web data analysis and application.