ISSN 2616-5775 Vol. 5, Issue 4: 18-21, DOI: 10.25236/AJCIS.2022.050403

Webpage Intelligent Parsing Algorithm Based on Text and Symbol Density

Junyu Xie

School of Information Management, Beijing Information Science & Technology University, Beijing, China

Abstract: *Web page intelligent parsing is an inevitable part of data collection. News web pages contain a lot of information with little relevance to the topic, which makes it difficult to locate the text content directly and quickly during the data collection process. This paper proposes a web page intelligent parsing algorithm based on text and symbol density. Through empirical research on the pages of mainstream news websites in China, the algorithm can quickly and accurately extract the text of news web pages.*

Keywords: *Web page extraction; intelligent page parsing; text density; symbol density*

1. Introduction

News web pages are cluttered and often contain a lot of content with little relevance to the title or body text. For example, inserting headers, footers and other content to maintain the integrity of the web page, adding content such as pictures and tables to improve the aesthetics of the page, adding program scripts to enhance the fluency and interactivity of the web page, adding content such as navigation and menus for convenience user browsing, and join advertising promotion to increase revenue. This kind of content is scattered in different positions on the news web page, even close to the content of the news body and attached to the body content of the web page. The existence of these irrelevant contents makes the crawler unable to quickly locate the required text information during the news data collection process. Therefore, how to quickly and accurately extract the text content from a large number of semi-structured news web pages in the process of data collection is a topic worthy of further study.

2. Research Status

At present, the more mature information extraction methods mainly include template-based methods and statistics-based methods.

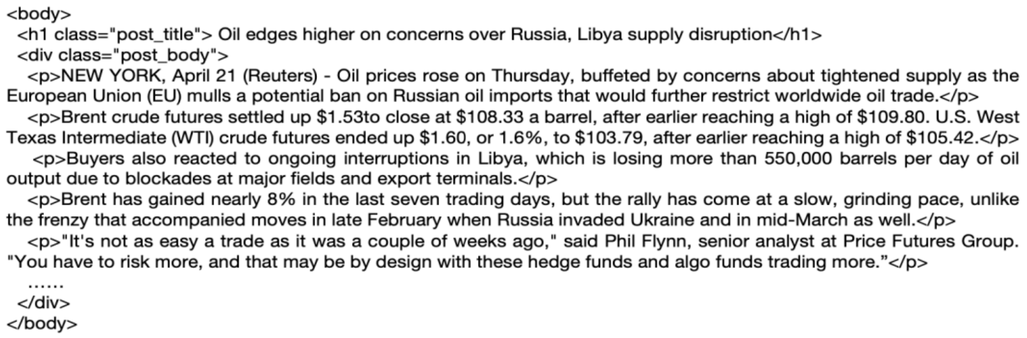
The template-based extraction method is an information extraction method that focuses more on the accuracy of the extraction results. This extraction method requires the user to write a template for web page information extraction in advance, and match the web page information according to the pre-written template.Xue Mei et al.[1]Using the structural and hierarchical characteristics of web page design templates, using web page link classification algorithm and web page structure separation algorithm, each information unit in the web page is extracted, and the corresponding Wrapper is output. Using Wrapper can automatically extract information from similar web pages.Wenchao Yang et al.[2] proposes an adaptive multi-information block Web information extraction based on DOM tree. The method first parses the web page into DOM tree through NekoHtml, and then determines the information blocks containing keyword groups, so as to realize Web information extraction.Marek Kowalkiewicz et al.[3]present an empirical evaluation and comparison of two content extraction methods in HTML: absolute XPath expressions and relative XPath expressions. We argue that the relative XPath expressions, although not widely used, should be used in preference to absolute XPath expressions in extracting content from human-created Web documents. Evaluation of robustness covers four thousand queries executed on several hundred webpages.First of all, templates are rules written in a fixed language format. Users first need to be familiar with how to write templates. Second, there are various web pages in the network, and the label formats in different It is difficult to formulate a widely applicable template; finally, it is more troublesome to expand the template that has been formulated, which usually affects the extraction of the content information of the entire web page.

The extraction method based on statistics is to extract the text information of the webpage according to the weight of the characters contained in different webpage tags. At present, the methods based on statistics mainly include statistical text density, label density and line block distribution. Mehta et al. [4] proposed the concepts of threshold and data filter on the basis of DOM tree, which are used to detect and delete irrelevant and redundant data in web pages, so as to dynamically eliminate the noise content of different structured web pages to extract the key points of web pages. content. K. Nethra et al.[5] proposes a hybrid approach to extract main content from Web pages. A HTML Web page is converted to DOM tree and features are extracted and with the extracted features, rules are generated. Decision tree classification and Naïve Bayes classification are machine learning methods used for rules generation.Chengjie Sun et al.[6]proposes a method to extract the text content from Chinese news web pages by relying on statistical information. The method firstly represents the web page as a tree according to the HTML tags in the web page, and then uses the number of Chinese characters contained in each node in the tree to select the node containing the text information. This method overcomes the shortcomings of traditional web content extraction methods that need to construct different wrappers for different data sources, and has the characteristics of simplicity and accuracy.Compared with the template-based extraction method, this method has better universality and scalability, but it also has certain limitations.

In summary, the template-based extraction method is more accurate and efficient, but it also has great limitations due to the need for users to define their own templates. The extraction method based on statistics is more suitable for the form of the webpage with concentrated content of the webpage body. For the situation that the webpage contains a lot of useless information, the extraction accuracy is not high. This paper proposes an automatic extraction method of news web pages based on text and symbol density. By dividing the DOM structure of news web pages into blocks, the scope of extracting news text is narrowed, and then the text density and symbol density in the blocks are combined to calculate the final density score. Density scores filter out body content.

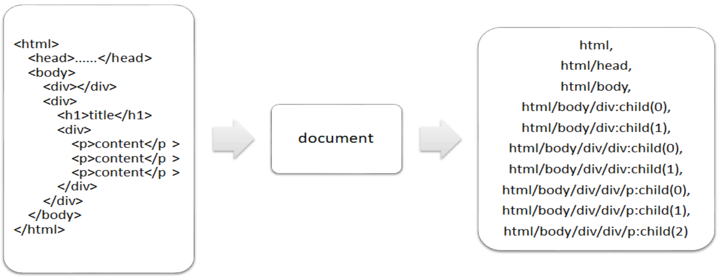
3. Web Page Extraction Algorithm

In the body page of news web pages, there are many obvious features. For example, the number of words in the text is large, the number of words in the <a> tag is small, the number of punctuation marks is more, and the number of paragraphs is more. The title of the text is usually enclosed in the tag <h*></h>, and the text is usually enclosed in the tag <p></p>. <a>Links or <span> tags may be enclosed within <p></p> tags, but only need to find the <p></p> that contains the body content. No matter what the tag contains, it can be considered the main content of the news. An example of html code in the news body page area, as shown in Figure 1:



```
<body>
  <h1 class="post_title"> Oil edges higher on concerns over Russia, Libya supply disruption</h1>
  <div class="post_body">
    <p>NEW YORK, April 21 (Reuters) - Oil prices rose on Thursday, buffeted by concerns about tightened supply as the
European Union (EU) mulls a potential ban on Russian oil imports that would further restrict worldwide oil trade.</p>
    <p>Brent crude futures settled up $1.53to close at $108.33 a barrel, after earlier reaching a high of $109.80. U.S. West
Texas Intermediate (WTI) crude futures ended up $1.60, or 1.6%, to $103.79, after earlier reaching a high of $105.42.</p>
     <p>Buyers also reacted to ongoing interruptions in Libya, which is losing more than 550,000 barrels per day of oil
output due to blockades at major fields and export terminals.</p>
    <p>Brent has gained nearly 8% in the last seven trading days, but the rally has come at a slow, grinding pace, unlike
the frenzy that accompanied moves in late February when Russia invaded Ukraine and in mid-March as well.</p>
    <p>"It's not as easy a trade as it was a couple of weeks ago," said Phil Flynn, senior analyst at Price Futures Group.
"You have to risk more, and that may be by design with these hedge funds and algo funds trading more."</p>
    ……
  </div>
</body>
```

Figure 1: News body page html code example



```
<html>
  <head>……</head>
  <body>
    <div></div>
    <div>
      <h1>title</h1>
      <div>
        <p>content</p >
        <p>content</p >
        <p>content</p >
      </div>
    </div>
  </body>
</html>
```

document
html,
html/head,
html/body,
html/body/div:child(0),
html/body/div:child(1),
html/body/div/div:child(0),
html/body/div/div:child(1),
html/body/div/div/p:child(0),
html/body/div/div/p:child(1),
html/body/div/div/p:child(2)

Figure 2: News body page html tag dismantling

First, you need to initialize html, parse the html file into a document object, and split and take out each tag. As shown in Figure 2.

Then obtain the label text statistics in each div box, and calculate the label text density in each div respectively. The calculation formula is:

$$\mathrm{DT_i} = \frac{\mathrm{NC_i} - \mathrm{NPC_i}}{\mathrm{ND_i} + 1 - \mathrm{NPD_i}} \tag{1}$$

Among them: $\mathrm{DT_i}$ represents the density of text in each <div> tag, $\mathrm{NC_i}$ represents the total length of the tag (number of char), and $\mathrm{NPC_i}$ represents the length of the <p> tag in the subtag (number of <p> char), $\mathrm{ND_i}$ represents the number of all subtags (number of descendants), $\mathrm{NPD_i}$ represents the number of <p> tags in the subtag (number of <p> descendants). Next, calculate the punctuation density in each div, the calculation formula is:

$$\mathrm{DP_i} = \frac{\mathrm{NC_i} - \mathrm{NPC_i}}{\mathrm{NP_i} + 1} \tag{2}$$

Among them: $\mathrm{DP_i}$ represents the punctuation density (density_of_punctuation) in each representative <div> tag, $\mathrm{NC_i}$ represents the total tag length (number of char), $\mathrm{NPC_i}$ represents the <p> tag length in the subtag (number of < p> char), $\mathrm{NP_i}$ represents the total length of punctuation (number of punctuation).

Finally, combining the label text density and punctuation density, use logarithmic functions of different bases to compress the data, adjust the weight ratio, and calculate the text density score in each div. The div box with the highest text density score is the text content. The final text density score calculation formula is:

$$\mathrm{DS_i} = \mathrm{DT_i} \times \log_{10}(\mathrm{NPD_i} + 2) \times \ln(\mathrm{DP_i}) \tag{3}$$

Among them: $\mathrm{DS_i}$ represents the final text density score (density_score) for each representative <div> area, $\mathrm{DT_i}$ represents the density of text in each <div> tag, and <p in the $\mathrm{NPD_i}$ sub-tag >number of <p> descendants, $\mathrm{DP_i}$ represents the density_of_punctuation in each <div> tag.

4. Validation experiment

4.1. Dataset Selection

This paper collects 9313 news web pages from different large news websites with a time span of 30 days as a sample dataset for the performance evaluation of the algorithm. At the end of the experiment, 100 results extracted from news samples using artificial methods (Xpath expression parsing, regular expression matching, etc.) were compared with the results extracted automatically by the algorithm.

4.2. Validation Results

For the extraction results, this paper compares the extracted content with the standard content, and uses the evaluation indicators commonly used in machine learning such as Precision and Recall to measure the pros and cons of the algorithm. The evaluation index calculation is shown in formula 4.

$$\mathrm{R} = \frac{\mathrm{LCS(S_0,\ S_1).length}}{\mathrm{S_0.length}},\quad \mathrm{P} = \frac{\mathrm{LCS(S_0,\ S_1).length}}{\mathrm{S_1.length}} \tag{4}$$

Among them: $\mathrm{S_0}$ represents the set composed of the results extracted by the automatic extraction algorithm, $\mathrm{S_1}$ represents the set composed of the standard results extracted by the manual method, and LCS($\mathrm{S_0}$, S1) represents the longest common subsequence of $\mathrm{S_0}$ and $\mathrm{S_1}$. The text comparison process uses a modified LCS (Longest Common Sequence) algorithm[7]. The same character with the longest length after removing zero or more characters from two given strings without changing the order of the remaining characters sequence. A top-down extrapolation method is used to calculate the length of the common substring. The operation steps of the algorithm are as follows:

(1) Get the length of the string $\mathrm{S_0}$ $\mathrm{lengthS_0}$ and the length of the string $\mathrm{S_1}$ $\mathrm{lengthS_1}$

- If the length of $\mathrm{S_0}$ or $\mathrm{S_1}$ is 0, the length of LCS is 0
- If $\mathrm{lengthS_0} \neq 0$ and $\mathrm{lengthS_1} \neq 0$, create a matrix of size $(\mathrm{lengthS_0} + 1) \times (\mathrm{lengthS_1} + 1)$

(2) Set the first row and first column of the created matrix to 0, that is

$$\mathrm{matrix}_{i,\ 0} = 0,\ \ \mathrm{matrix}_{0,\ j} = 0,\ \ 0 \le i \le \mathrm{length}S_0,\ \ 0 \le j \le \mathrm{length}S_1 \tag{5}$$

(3) Initialize the matrix

$$\mathrm{matrix}_{i,\ j} = \begin{cases} 1, & S_{0_i} = S_{0_j} \\ 0, & S_{0_i} \neq S_{0_j} \end{cases} \tag{6}$$

(4) Calculate the matrix from top to bottom and from left to right according to the requirements of formula 6, and finally get $\mathrm{matrix}_{i,\ j}$

$$\mathrm{matrix}_{i,\ j} = \begin{cases} \mathrm{matrix}_{i,\ j+1}, & S_{0_i} = S_{0_j} \text{ and } \mathrm{matrix}_{i-1,\ j} == \mathrm{matrix}_{i,\ j-1} \\ & \max(\mathrm{matrix}_{i,\ j-1},\ \mathrm{matrix}_{i-1,\ j}) \end{cases} \tag{7}$$

The maximum value in the matrix represents the longest common subsequence, which is represented by LCS. The experimental results are shown in Table 1.

Table 1: News Webpage Extraction Experiment Results

Evaluation indicators	Precision	Recall
Sohu News	93.9	95.3
Eastmoney News	95.5	91.2
Netease News	92.4	96.7
Sina News	93.0	94.1

The experimental results show that the text content extracted by this algorithm is more consistent with the results of manual extraction. Due to the inconsistency in the extraction range of tags before and after HTML paragraphs, there is still a 10%-15% error. Overall, the algorithm is suitable for news website text. The intelligent extraction of content is still relatively good.

5. Conclusions

This paper proposes a web page intelligent parsing algorithm based on text and symbol density, aiming at the shortcomings of the existing methods in the process of extracting the text pages of news web pages in China. The algorithm can quickly and accurately extract the text of news web pages.

References

[1] Xue Mei et al. A method for fully automatic generation of web page information extraction Wrapper [J]. Journal of Chinese Information Processing,2008(01):22-29.

[2] Wenchao Yang et al. Adaptive Multi-Information Block Web Information Extraction Based on DOM Tree [J]. Network Security Technology & Application,2012(11):62-64.

[3] Marek Kowalkiewicz,Maria E. Orlowska,Tomasz Kaczmarek,Witold Abramowicz. Robust web content extraction[P]. World Wide Web,2006.

[4] MEHTA B, NARVEKAR M. DOM tree based approach for web content extraction[C]// 2015 International Conference on Communication, Information & Computing Technology. Mumbai: IEEE, 2015: 1-6.

[5] K.Nethra,J. Anitha,G. Thilagavathi. WEB CONTENT EXTRACTION USING HYBRID APPROACH [J]. ICTACT Journal on Soft Computing,2014,4(2).

[6] Chengjie Sun et al. Research on the Method of Information Extraction of Web Page Text Based on Statistics [J]. Journal of Chinese Information Processing,2004(05):17-22.

[7] Yongxin Wang et al. An Efficient LCS Algorithm [J]. Journal of Nanyang Institute of Technology, 2013 (6) :67-70.