A Web Content Extraction Method Base on Punctuation Distribution and HTML Tag Similarity

Nan Gong, Chunxiao Fan, Yuexin Wu, and Yue Ming

Abstract Currently, web content extraction methods mostly focus on single-theme pages and have poor adaptability for multi-theme pages. In order to overcome this issue, this paper proposed a web content extraction method based on the punctuation distribution and HTML tag similarity. According to the characteristic that most of the punctuation appeared in the main text areas but rarely appeared in noise areas of web pages, an algorithm of obtaining minimum text area was presented. Furthermore, in the case of multi-theme pages, this paper proposed an approach to extract the titles and contents from each theme by further dividing the minimum text area into sub theme areas based on the tag similarity. Experimental results showed that the proposed method can effectively and accurately extract web content in different themes.

Keywords Web content extraction • Punctuation distribution • Tag similarity • Tag tree

1 Introduction

With the rapid development of the technology and scale of the Internet in recent years, the network has become a very significant channel which people obtain kinds of information. Noise such as advertising links will affect the further analysis and utilization of Web information, because the main content is often submerged in noise. To solve this problem, web content extraction technology came into being.

N. Gong (✉) • C. Fan • Y. Wu • Y. Ming
School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China
e-mail: masteryodacn@gmail.com; fcx100@gmail.com; wuyuexin@263.net; mingyue35875235@gmail.com

R. Zhang et al. (eds.), *LISS 2013: Proceedings of 3rd International Conference on Logistics, Informatics and Service Science*, DOI 10.1007/978-3-642-40660-7_120,

Web pages can be divided into three categories in general, which are content-dominated pages, navigation-dominated pages and multimedia-dominated pages. Content-dominated pages can be divided into single-theme pages, which have but one content block, and multi-theme pages containing multiple content blocks (Shaokang Wang et al. 2010). At present, lots of web content extraction methods focus on the single-theme web pages, and they cannot be applied to multi-theme pages to extract the contents of the different themes in the same page. But for web information processing, like information classification, clustering and automatic text summarization, more detailed extraction would be needed than before. This paper proposed a web content extraction method based on punctuation distribution and HTML tag similarity, which can extract web content separately by different themes with low complexity.

2 Related Work

There are many related research on the web content extraction, which is the basis of the web information processing. The most three popular algorithms are template-based method, statistic-based method and visual-based method.

Y. Kim proposed an enhanced tree matching algorithm that improved the tree edit distance method by considering the characteristics of HTML features (Yeonjung Kim et al. 2007). An HTML pattern was done by obtaining the maximum mapping values of two trees. Li Lei put forward a FFT-based extraction algorithm (Lei Li et al. 2007). This method calculated the weight of every possible range by applying window-segmentation, statistics theory and FFT. The best one was selected as solution. C. H. Li proposed an Extraction based on VIPS (Li Cunhe et al. 2010). A classification model was trained with some manually annotated blocks that were partitioned based on VIPS. Then the informative blocks can be extracted through the model.

However, some deficiencies exist in the current web content extraction methods. It can be seen that most web content extraction methods with insufficient processing ability on multi-theme pages cannot adapt to the Internet increasingly diverse and complex pages. To solve the problem that web content extraction methods mostly focusing on the single-theme pages and poorly adapting to the multi-theme pages, this paper proposed a web content extraction method based on punctuation distribution and HTML tag similarity.

3 Web Content Extraction Method Base on Punctuation Distribution and Tag Similarity

Web content extraction is the basis of the analysis of web information. This paper proposed a new method of web content extraction which included 3 steps: (1) constructing a tag, (2) obtaining the minimum text area, and (3) extracting the

sub-theme contents. In Sect. 3.1, a tag tree from HTML tag was constructed, and some information of tags and texts was added to each tree node for the convenience of extraction. In Sect. 3.2, a minimum text area obtaining method which was based on the distribution of punctuation was utilized for distinguishing between the content and noise. In Sect. 3.3, a method was proposed to further divide the minimum text area; this method was based on the similarity of tags, and could be used for distinguishing between single-theme page and multi-theme page effectively and extracting contents separately by different themes.

3.1 Constructing the Tag Tree

Constructing a tag tree from web page is a necessary step of many web content extraction methods (Yuhong Chang et al. 2004). In this research, the HTML tag was used to construct the corresponding tag tree. Meanwhile, some information, such as tag attributes and the amount of punctuation was added to each tree node so that contents could be extracted easily in step 2 and 3.

According to the characteristics of the HTML tag and text, the properties of the tag node were defined as {tagName, type, parent, childrenList, attribute, Data, punNum, textNum}. Property tagName was the name of the tag; property Type denoted the type of each node, and nodes were divided into branch nodes, text nodes and hyperlink nodes (<a> tag represented a hyperlink node directly instead of a branch node and a text node); property parent was the parent node; property childrenList was a set of successors; property attribute was the map of the attributes of the HTML tag; property Data was the text content of the node; punNum and textNum denoted the total numbers of punctuation and words in all the descendants of each node.

The calculation of punNum and textNum should follow the following rules:

Rule 1: punNum and textNum of each text node or hyperlink node equal to the amounts of punctuation and words in the node.
Rule 2: punNum of each branch node equals to the sum of its successors' punNum, and textNum of branch node equals to the sum of its successors' textNum.
Rule 3: For a text node or a hyperlink node, if punNum>textNum, this node is then considered to be noise, and punNum is set to 0.

To reduce the amount of tree data, pruning and merging is necessary and should be carried out under the following rules:

Rule 1: If a branch node has no successor, this node should be deleted.
Rule 2: If the tagName of a node is in set (script, style, iframe, object, select), this node and all of its descendants should be deleted.
Rule 3: If <p> node contains a text node and the number of characters is greater than 5, it means that <p> node contains a paragraph. Thus, all of the text nodes, hyperlink nodes, nodes and nodes below the <p> node should be merged into a text node.

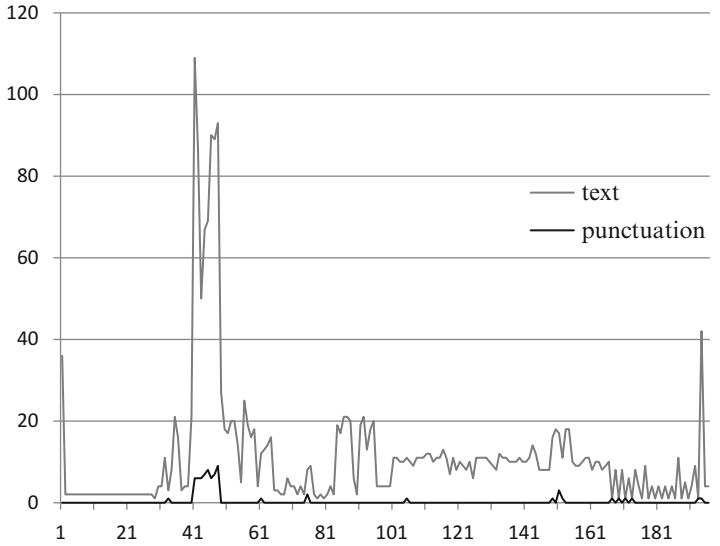

Fig. 1 Distribution of punctuations and words

3.2 Obtaining the Minimum Text Area According to the Distribution of Punctuation

Noise such as advertisement and hyperlink, should be removed in order to get the minimum text area, which contains all the text information but no noise. The minimum text area performed as a node in the web page structure; it contains the entire text contents while any successor should not contain the full text contents.

"Punctuation marks are symbols that indicate the structure and organization of written language, as well as intonation and pauses to be observed when reading aloud." Totally 50 content-dominated pages were selected randomly for statistics. Results showed that the punctuation in the content area constituted 76 % of the total amount of punctuation in the web page, while the ratio of words was only 22 %.

Figure 1 shows the distribution of punctuation and words in a Sina news web page. The abscissa arranged in the order of the appearance of text nodes. It is evident that the numbers of words and punctuation in the text region were larger than those in other regions. The fluctuations of the amounts of words were in high level in noise area. Conversely, the amounts of punctuation were almost zero in noise area. Instead of the word amount in every line, punctuation amount is more appropriate to distinguish between the major text area and the noise area.

Based on the analysis above, the main conclusion is as follows:

Conclusion 1: The amount of punctuation can be used as a basis of making judgments about whether the area is major text area or noise area.

According to this conclusion, we judge whether a node is in noise area on the ratio of the amount of punctuation contained in this node to the amount in its parent node. If the ratio is greater than the threshold Q, the sub-node should be determined the major text area which contains the entire information of its parent node, while the remaining sub-nodes are noise areas. If the ratio is less than the threshold Q, it shows that the punctuation in this node is widely dispersed, and this node is the smallest node containing the entire text information of the page, and cannot be divided.

The minimum text area can be obtained by filtering web tag tree from top to bottom layer. The detailed procedure is.:

1. Define pnode equal to body node.
2. Find out the snode which is a sequence of pnode and has the largest punNum.
3. If snode.punNum/pnode.punNum> threshold Q, then define the pnode equal to snode, and repeat step 2.
4. Pnode is then defined as the minimum text area.

3.3 Sub-themes Content Extraction

The minimum text area should be divided into several sub-themes areas in order to extract contents separately by different themes. By analyzing the structural characteristics of several popular multi-theme web pages, we found that the structures and layouts of titles and contents in these sub-themes webpages were identical, and the locations, fonts, background colors and styles of web pages in different themes were basically the same. After analyzing the observation, the following three conclusions were suggested to be used for multi-theme extraction.

1. In multi-theme web pages, the tag tree structures describing each text area are almost the same; in addition, all of the tag trees are the successors of the minimum text node. Figure 2 shows a paradigm of tag tree structure in multi-theme page. Node <div#1> and <div#2> both contains a theme main text contents, their structures are identify and they are both the successors of the minimum text area <div*>.
2. Currently, almost every website uses CSS to control the display of internet web pages. CSS, namely Cascading Style Sheets, is a language used to perform file styles, i.e. HTML. Almost every CSS uses the attribute class to describe the style for reuse. Only in some extreme cases the attribute id is used to control the style of a certain element, which can be ignored due to no impact on the extraction.

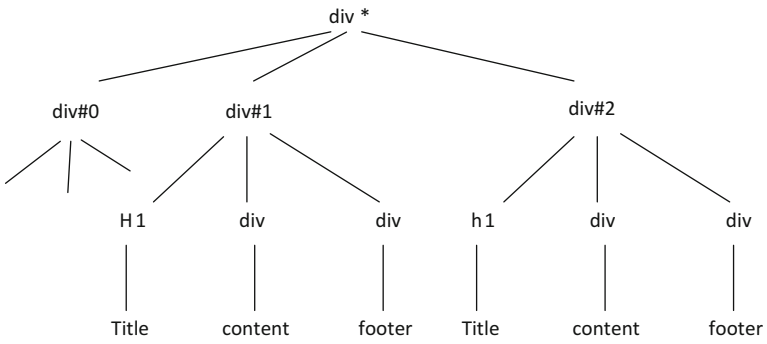

Fig. 2 Paradigm of tag tree structure in multi-theme page

Fig. 3 Several groups of similar tags in the minimum text area

```
<div*>
  <div.a>ad</div>
  <h1.h>title1</h1>
  <div.c>content1</div>
  <div.a>ad</div>
  <h1.h>title2</h1>
  <div.c>content2</div>
</div>
```

Therefore, for two tags which have the same attribute class and the same tagName, we can get the conclusion that the displays of the two tags are the same, without browser rendering.

3. The text information contained in each theme text area should be the same as that contained in the minimum text area. That means text information cannot be dropped when we extract sub-theme information, in order to ensure the integrity of the extracted information.

We defined the concept of similar tags based on the conclusion 1 and 2.

Definition 1. if tag A and tag B have the same parent nodes and their attributes class are the same, tags A and B can be considered similar, and they are similar tags.

According to the analyses above, we preliminarily judge whether a page is single-theme or multi-theme by analyzing the similarity of the successor nodes of the minimum text node. If there are no similar tags in the successor nodes of the minimum body text, the page is then determined as a single-theme page; otherwise it is determined as multi-theme page. Since the minimum text area may still contain a group of noise areas with similar tags, as the paradigm showed in Fig. 3, the information of each group divided by similarity should be compared with the information contained in the minimum text area. The group of tags, of which the ratios are greater than the threshold T, is defined as each theme area of the multi-theme page. If the ratios of each group are less than threshold T, this web page is determined as a single-theme page.

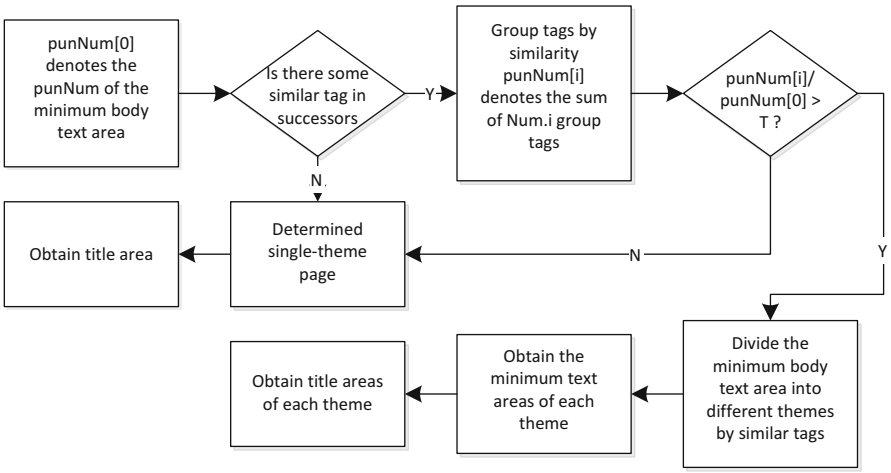

Fig. 4 Flow chart of the content extraction in minimum text area

We used the method described in Sect. 3.2 to obtain the minimum text areas of each theme. The title area is obtained by fining the node, which is <h1> or <h2> tag or contained string 'title', in the minimum text area. If there are no title areas found in the text area, we then search the title area in front of the minimum theme text area.

Figure 4 shows the flow chart of the content extraction in minimum text area.

4 Experimental Results

To verify the web content extraction method proposed, we measure the actual effect of the method by using recall and precision in information retrieval technology.

$$\text{Recall} = \text{Correct}/\text{Actual} * 100\ \%. \tag{1}$$

$$\text{Precision} = \text{Correct}/\text{Extracted} * 100\ \%. \tag{2}$$

Where, Correct is the number of correctly extracted themes; Actual is sum of themes in all pages; Extracted is the number of extracted themes.

We randomly selected 759 web pages in some well-known websites including 399 single-theme pages and 360 multi-theme pages. The results are shown in the following table (Table 1).

Experimental results show that the proposed method for web content extraction can extract content by different themes automatically, while maintaining higher recall and precision rate.

Table 1 The content extraction results of some central web sites

Type	Website	Pages	Actual	Extracted	Correct	Recall (%)	Precision (%)
Single-theme	news.sina.com	153	153	153	151	98.69	98.69
	news.sohu.com	142	142	141	138	97.18	97.87
	www.bbc.co.uk/news	104	104	110	98	94.23	89.09
Multi-theme	bbs.hupu.com/	129	1883	1767	1734	92.09	98.13
	blog.sina.com	122	845	815	792	93.73	97.18
	forum.ubuntu.org.cn/	109	983	915	847	86.16	92.57

5 Conclusion

This paper proposed a web content extraction method which was able to extract contents separately by different themes. This method is easy to implement. Besides, it has been tested to be universality and adaptability, and is independent to page data sources where prior sample learning is unnecessary. This work has laid a good foundation for the web-based text mining in the future.

Acknowledgments The work presented in this paper was supported by Beijing Municipal Commission of Education Build Together Project.

References

Lei Li, Jinlin Wang, He Bai et al (2007) Research and implementation of FFT-based extraction algorithm of webpage content main body. Comput Eng Appl 43(30):148–151

Li Cunhe, Dong Juan, Chen Juntang (2010) Extraction of informative blocks from web pages based on VIPS. J Comput Inf Syst 6:271–277

Shaokang Wang, Kejun Dong, Baoping Yan (2010) Web content information extraction using density of feature text. Comput Eng Appl 46(20):1–3

Yeonjung Kim, Jeahyun Park, Taehwan Kim et al (2007) Web information extraction by HTML tree edit distance matching. Int Conf Converg Inf Technol 2:2455–2460

Yuhong Chang, Zhe Jiang, Xiaoyan Zhu (2004) Web page structure analysis based on Tag tree method. Comput Eng Appl 40(16):129–132