Chinese Web Content Extraction Based on Naïve Bayes Model

Wang Jinbo, Wang Lianzhi, Gao Wanlin, Yu Jian, Cui Yuntao

Chinese Web Content Extraction based on Naïve Bayes Model

Wang Jinbo[1,a], Wang Lianzhi[1,b], Gao Wanlin[1,c], Yu Jian[1,d], Cui Yuntao[1,e]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China
[a]wangcau@163.com, [b]ndjsj862@cau.edu.cn, [c]gaowlin@cau.edu.cn, [d]yj@cau.edu.cn, [e]674853800@qq.com

Abstract. As the web content extraction becomes more and more difficult, this paper proposes a method that using Naive Bayes Model to train the block attributes eigenvalues of web page. Firstly, this method denoising the web page, represents it as a DOM tree and divides web page into blocks, then uses Naive Bayes Model to get the probability value of the statistical feature about web blocks. At last, it extracts theme blocks to compose content of web page. The test shows that the algorithm could extract content of web page accurately. The average accuracy has reached up to 96.2%.The method has been adopted to extract content for the off-portal search of Hunan Farmer Training Website, and the efficiency is well.

Keywords: Web Content Extraction; DOM Tree; Page Segmentation; Naive Bayes Model.

1 Introduction

Web content extraction is to extract the text which describe the page content; and it's also known as web theme block extraction [1]. It can be used for web data mining, classification, clustering, keyword extraction and the deep processing of web information. Web is semi-structured pages, so it contains a lot of advertising links, scripts, CSS styles, navigation and useless information. The main message is often hidden in the unrelated content or structure; and the noise makes it very difficult to extract page content. Therefore, how to quickly and accurately extract text content pages has been the focus of research at home and abroad [2].About web page text extraction, there is also a lot of research and methods. Now, three main web content extraction algorithms are as follows:

1. Wrapper-based approach, this method is to extract required information from specific web information sources and be expressed in a particular form. Wrapper-based approach can be accurate extracting and have high accuracy. But due to the complexity and irregular of web structure, a wrapper implementation generally for one website, it is difficult to meet for different web information extraction tasks [3].

2. Machine learning methods, by analyzing the structure of the page, and constantly generates new template and creates template library. Literature [4] takes machine learning methods for web thematic information extraction. Web page content extraction based on templates has a relatively high degree of automation and is convenient for users. However, if you encounter a web page cannot find the corresponding template, the extraction will fail. As the template library continues to increase, the template library management will become increasingly complex [5].

3. Visualization layout classification method, a classical algorithm is VIPs put forward by Microsoft Asia research institute. It uses visual characteristics of the page structure excavation and makes full use of the web page background color, font color and size. However, due to the complexity of visual web, heuristic rules are so ambiguous that need to manually adjust the rules constantly. So how to ensure consistency of the rules is a difficulty [6].

The methods mentioned above all have some short comings and limitations. So this paper on the basis of predecessors' work and combining with the nature of html page in statistics and observation, according to the characteristics of the different features with different importance, it proposes an algorithm that uses Naïve Bayes Model [7] to train the block attributes eigenvalues of web page. After denoising the web page [8], divides web page into blocks and gets the statistical characteristics of the web block. The algorithm is easy to implement, without artificial participation and can extract web contents quickly and accurately.

2 Algorithm Framework

The algorithm is divided into training phase and testing phase. The training phase includes pretreatment of web pages and builds Naïve Bayes Model. The testing phase is based on the web pages pretreatment, using Naïve Bayes model which is built in training phrase to extract web content. Algorithm framework shows in figure 1.

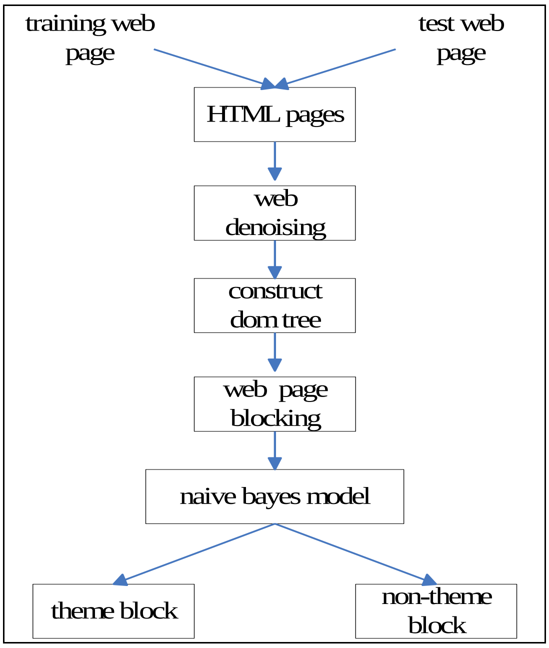

Fig. 1. Algorithm Framework

3 Web Page Pretreatment

Step one: Downloads news web pages respectively from Sohu, Netease, Sina, People's daily, Tencent. And each source downloads 200 web pages, then extracts web page manually, 500 as the training set, 500 as the testing set.
Step two: Denoising web pages and uses regular expression to delete CSS, scripts, comments on pages.

Table 1. Web Denoising Regex.

noise type	regex
css styles	<[\\s]*?style[^>]*?>[\\s\\S]*?<[\\s]*?\\/[\\s]*?style[\\s]*?>
script	<[\\s]*?script[^>]*?>[\\s\\S]*?<[\\s]*?\\/[\\s]*?script[\\s]*?>
comments	<!—(.*?)-->

Step three: Resolve web page into a DOM tree [9]. Read the web page without noise into memory and use NekoHTML to modify the tags which is not regular, then resolves web page into a DOM tree [10]. The html in figure 2(a) corresponds to the DOM tree in figure 2(b) below:

```
<html>
  <head>
    <title>四大名著</title>
  </head>
  <body>
    <div><table><tr>
        <td>三国演义</td>
        <td>水浒传</td>
        <td>西游记</td>
        <td>红楼梦</td>
    </tr></table></div>
  </body>
</html>
```

Fig. 2(a) HTML Web Page

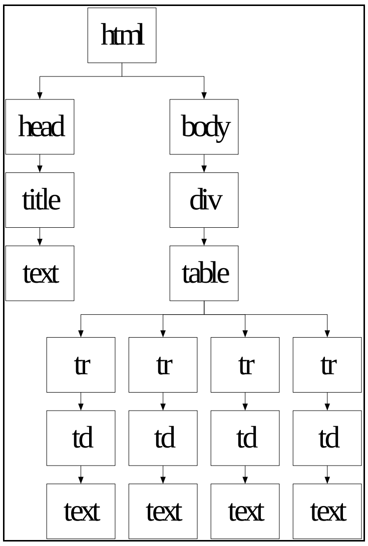

Fig. 2(b) HTML Page's DOM Tree

Html document and DOM tree is a one-to-one relationship, and the DOM tree makes computer more convenient to process semi-structured html document, easier to block the web pages.

Step four: DOM tree blocking. By observing the website, finding that the text area is usually use tags such as table, td, tr, div to divide each block of text. So this article compares the above tags to DOM tree node properties, using the bottom-up approach to block the DOM tree. The block rules are as follows:

(1) Let DOM tree leaf node enter the queue.

(2) Scanning the DOM tree leaf node in turn, if the leaf node text is empty or is not block node , continue to scan the node's parent node until it encounter a block node whose text is not empty. Recording the node and compose it and its affiliated tags into blocks.

(3)Scanning the leaf node again, the same as (2), if it encounter a block node whose text is not empty. Recording the node and compose it and its affiliated tags into blocks. If the node and block node in (2) is sibling nodes, merges the block and block in (2).

4 Model Design

4.1 Naïve Bayes Model

Naive Bayes Model (Naive Bayesian Model, NBC) is the most widely used classification algorithm, it needs less estimated parameters, less sensitive to missing data, and its time complexity is low, classification is efficiency and stability.

Whether the web block is content or not is a Binary Classification Problem [11]. We use an n-dimensional feature vector $X = \{x_1, x_2, \ldots, x_n\}$ to represent a block, describing the n metrics about samples corresponding to the attributes $A_1, A_2, \ldots A_n$. $c_i \in C = \{c_1, c_2\}$ is a class variable, c_1 indicate that the page belongs to theme block, c_2 indicate that the page does not belong to theme block. To simplify the calculation, assume $x_1, x_2, \ldots, x_n$ are independent. That is, attribute values is independent between each other. It's why we call Naïve Bayes Naive [12]. A web block belonging to c_i classification's Naïve Bayesian formula shows as (1) [13]:

$$p(c_i \mid x_1, x_2, \ldots, x_n) = \frac{\prod_{j=1}^{n} p(x_j \mid c_i) * p(c_i)}{\prod_{j=1}^{n} p(x_j)} \tag{1}$$

4.2 Block feature extraction

In the paper, probability based statistical training webpage probability feature of each block is to determine the probability of block extraction test webpage of theme block. A lot of features affecting a block becoming subject block. By analyzing the structure of web pages, we can draw the conclusion that the following theme blocks [14] have several notable features:

(1)Hyperlinks are less, but navigation information blocks, advertising block contains a number of hyperlinks generally more.

(2)More text is in block; theme block is the region web information centralized, so the number of characters contained within the block is more. The noise block contains fewer amounts of characters.

(3)Theme block is used to describe the main content of a webpage, so it contains more punctuation, and noise block generally doesn't contain punctuation.

(4)<p> acts as paragraph mark, the theme contains a lot of information, and practical <p> labels often used to segment, while the noise block generally doesn't contain paragraph marks.

Based on the above characteristics , this article uses the number of characters within the block unlink , the ratio of the number of link characters and the total number of characters , the ratio of total number of punctuation and link characters , and the total number of <p> as Web page block's feature items.

Among this paper, *unlinktextsum* stands for the number of unlink characters, *linktextsum* stands for the number of link characters, *textsum* stands for the total number of text, and *puncsum* stands for the total number of punctuation.

Then the ratio of *linktextsum* to *textsum* named link shows as (2):

$$\mathrm{link} = \frac{\mathrm{linktextsum}}{\mathrm{textsum}} \tag{2}$$

The ratio of *puncsum* to *linktextsum* named *punc* shows as (3):

$$\mathrm{punc} = \frac{\mathrm{puncsum}}{\mathrm{linktextsum}} \tag{3}$$

4.3 Model training

After generating DOM tree and blocking the training web pages [15], to work out the unlink characters number, ratio of unlink characters to character number, ratio of punctuation to link characters number, <p> tags number.

4.3.1 Unlink Characters Number

The more unlink characters in a block, the richer information the block contains, then it has a higher probability to be a theme block. Because hyperlinks are generally less than 20 characters, and blocks more than 100 characters are mostly theme block. In this article we will divide the number of block's unlink characters into 6 levels, that is, the number of unlink characters is less than 20, above 100, and four equal parts between 20 and 100. The unlinked character number scatters in an interval belonging to non-theme blocks and theme blocks' probability shows as (4), (5):

$$p(linktextsum_n \mid c_1) = \sum_{i=1}^{n} p(linktextsum_i \mid c_1) \qquad 1 \le i \le n \tag{4}$$

$$p(linktextsum_n \mid c_2) = \sum_{i=1}^{n} p(linktextsum_i \mid c_2) \qquad 1 \le i \le n \tag{5}$$

That is, each block *linktextsum*'s probability is the probability of *linktextsum* less than the block and the block's sum, and $n = \frac{linktextsum}{20} + 1$.

4.3.2 Link Characters and Character Number ratios Probability

The lower link characters and character number ratios within the block, the higher probability a block to be a theme block. Navigation links blocks and advertising blocks of characters and the total number of characters ratio is generally greater than 50% and some even higher than 80%. So this article will divide link characters and character number ratios into 4 copies, that is, less than 10%, 10%-50%, 50%-80%, above 80%. The link characters and character number ratios scatters in an interval belonging to non-theme blocks and theme blocks' probability shows as (6), (7):

$$p(link_i \mid c_1) = \frac{linktextsum_i}{textsum_i} \tag{6}$$

$$p(link_i \mid c_2) = \frac{linktextsum_i}{textsum_i} \tag{7}$$

4.3.3 Punctuation and link characters number ratios probability

Theme block contains much punctuation, but link text generally doesn't contain punctuation. The higher punctuation and link characters number ratios within the block, the higher probability a block to be a theme block. This article will divide ratios of punctuation number to link characters number into 3 copies, that is, less than 2%, 2%-10%, above 10%. The ratios of punctuation number to link characters number scatters in an interval belonging to non-theme blocks and theme blocks' probability shows as (8), (9):

$$p(punc_n \mid c_1) = \Sigma_{i=1}^{n} p(\frac{puncsum_i}{linksum_i} \mid c_1) \quad 1 \le i \le n \tag{8}$$

$$p(punc_n \mid c_1) = \Sigma_{i=1}^{n} p(\frac{puncsum_i}{linksum_i} \mid c_2) \quad 1 \le i \le n \tag{9}$$

That is, each block *punc*'s probability is smaller than the block *punc* and the block's *punc* probability sum.

4.3.4 <p> tags number probability

Web content containing much information, it often uses <p> tags for a paragraph replacement. So theme block contains many <p> tags. This article will divide <p> tags number into 3 levels, that is, 0 <p> tag, 0-3<p> tags, above 3 <p> tags. The <p> tags number scatters in an interval belonging to non-theme blocks and theme blocks' probability show as (10), (11):

$$p(psum_i \mid c_1) = psum_i \tag{10}$$

$$p(psum_i \mid c_2) = psum_i \tag{11}$$

4.3.5 Block overall probability

According to the formula (1) - (11):
The probability of a block to be a theme block shows as (12):

$$p(c_1 \mid unlinksum, link, punc, psum) = \frac{p(unlinksum, link, punc, psum \mid c_1) * p(c_1)}{p(unlinksum, link, punc, psum)}) \tag{12}$$

According to the formula (1), (12):

$$p(c_1 \mid unlinksum, link, punc, psum) = \frac{p(unlinksum \mid c_1) * p(link \mid c_1) * p(punc \mid c_1) * p(psum \mid c_1)}{p(unlinksum, link, punc, psum)} * p(c_1) \tag{13}$$

p(c1) indicates the probability of a theme block in the training set, is a constant, and the denominator is also a constant.
Therefore, the probability of a block to be a theme block can be expressed as (14):

$$p(c_1 \mid unlinksum, link, punc, psum) = p(unlinksum \mid c_1) * p(link \mid c_1) * (punc \mid c_1) * (psum \mid c_1) \tag{14}$$

Similarly, the probability of a block to be a non-theme block can be expressed as (15):

$$p(c_2 \mid unlinksum, link, punc, psum) = p(unlinksum \mid c_2) * p(link \mid c_2) * (punc \mid c_2) * (psum \mid c_2) \tag{15}$$

If in a block p(c1|unlinksum,link,punc,psum)>= p(c2|unlinksum,link,punc,psum), that is, the probability a block to be a theme block is bigger than the probability a block to be a non-theme block. Extract the block, put it into theme block queue, and output the block in queue.

5 Testing and Verification

In order to verify the effectiveness of the algorithm, we use java language to implement and test the proposed algorithm. Test procedure is as follows:

Download 100 pages respectively from *Sohu*, *Netease*, *Sina*, People's Daily and *Tencent*, totaling 500 web pages. These pages cover sports, entertainment, education, practical, financial and some other themes, almost all kinds of news.

Using the algorithm to extract text of the following web page from *Sina* Finance and Economics, the page to be extracted is shown in figure3:

The page URL is

http://finance.sina.com.cn/review/jcgc/20130606/182015723399.shtml .

首页 | 新闻 | 体育 | 娱乐 | 财经 | 股票 | 科技 | 视频 | 微博 | 博客 | 读书 | 教育 | 女性 | 育儿 | 星座 | 收藏 | 尚品 | 汽车 | 房产 | 更多 注册 | 登录

sina 新浪财经 新浪财经 > 评论 > 2013年国内成品油价三次下调 > 正文

新闻 成品油 搜索

油价微调对物流行业影响甚微 整体需求欠佳

2013年06月06日 18:20 隆众石化网 我有话说(6人参与)

2013年，经济恢复缓慢，直至6月份尚未见成品油、煤工矿业完全恢复至去年同期水平。物流运输行业也因此受到连带影响，运营较差。截至2013年5月底，成品油市场经历了两涨两跌，尽管下调累计幅度高于上调幅度，但是物流行业未能因此发生变化。2013年物流行业的主要影响因素受到那些方面影响，对此，隆众石化网对国内多家大型的物流企业进行了调查。

河北一家拥有80台汽柴油运输车辆的物流公司张先生表示“2013年，我们的运费压得很低，只能保本，活太少了，尤其是运输汽柴油的业务量比去年减少了很多”。张先生还称“为了增加竞争力，我们今年运费压比去年低了接近200块，现在从河北去一趟内蒙，运费只收240元/公里，回来的时候再捎些甲醇(2638,-6.00,-0.23%)，要不然就赔钱，回来的时候甲醇的活多一点，运费可以到300元”。另外一家的公司的姜先生告诉说“我们现在有100辆车，有70辆挂靠的，自己只经营了30多辆，现在成本增加货太少，司机的工资，过路费等费用都很高，现在很难干。”

隆众还了解到，物流运输公司日常成本中人工工资、过路费用、燃烧用油费用占比较高的三个部分，但是，成品油目前的价格对物流成本的影响较小，主要还是取决于运营里程。也就是说，此次油价下调幅度较小，对物流运输公司的成本影响甚微。隆众石化网成品油分析师薛群表示，油价下调短期依旧不会增加物流行业的利润空间，有一些物流企业也在发展物流运输的同时，增加油品等商品的贸易实现多元化发展。

【新浪财经手机客户端下载 安卓客户端下载】

标签： 成品油 下调 需求

已有4条评论，共6人参与

头条推荐

- 【财经】铁路系统称刘志军好色 对钱倒真不太在乎
- 【科技】360将推便携路由器 电脑周围建立WiFi
- 【博客】美国最著名“逃犯”和蝙蝠侠式反恐
- 【尚品】时尚大师亿元豪宅 世界最令人艳羡工作
- 【读书】泰华公车私用祸幸民政高官：中堂(全)
- 【教育】中国大学50强榜单出炉 出状元最多的中学

强劲动力中的宁静驾乘体验
来自别克英朗GT

Fig. 3 the Original Page

Text extraction result is shown in Figure 4:

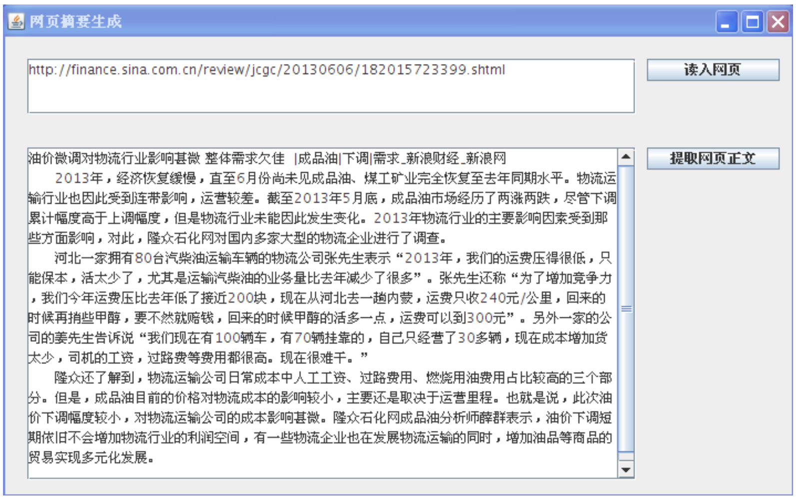

Fig. 4 Extraction Results

We divide the obtained theme information into three levels: (1) Excellent: the obtained web text is consistent with the text manually labeled. (2) Good: compared to the text manually labeled, there is only 1-2 sentences lost, or the text contains 1-2 noise blocks. (3) Poor: The text contains many mistakes. Specific test results are shown in table 2[16]:

Table 2 Algorithm Experimental Results in This Paper

web pagesource	web page number	excellent	good	poor	excellent rate(%)	good rate(%)
Sohu	100	36	60	4	36%	96%
Sina	100	38	59	3	38%	97%
Netease	100	35	61	4	35%	96%
People’s daily	100	38	59	3	38%	97%
Tencent	100	32	63	5	32%	95%

Table 3 <table> to Block Web Page Extraction Algorithm Results

web pagesource	web page number	excellent	good	poor	excellent rate(%)	good rate(%)
Sohu	100	25	70	5	25%	95%
Sina	100	28	68	4	28%	96%
Netease	100	26	62	12	26%	88%
People’s daily	100	30	66	4	30%	96%
Tencent	100	27	63	10	27%	90%

In the tables above, excellent rate is the proportion of excellent level result in all result data; good rate is the proportion of both excellent and good level in all result data.

The algorithm in this paper compares to the method only use <table> to block web page, both its good rate and excellent rates are significantly improved.

6 Conclusion

This paper proposed an algorithm using Naïve Bayes Model to train the block attributes eigenvalues of web page. Then it extracts theme blocks and composes content of web page. The method has been adopted to extract content for the off-portal search of Hunan Farmer Training Website, and the efficiency is well. Counting the good web pages extracted, the average accuracy rate is up to 96.2%. For some well-structured web pages, the accuracy rate will be even higher. An existing deficiency is the block tags considered relatively less, therefore, if consider more block tags, the accuracy of the system will also be enhanced. In future work we will do research for semi-structured web pages.

References

1. Li Xueqing. Harmonious man-machine environment [M]. 1. Beijing: Tsinghua University press, 2008-1-1 :101-107.
2. Wu Qi,Chen Xingshu,Tan Jun. Content Extraction Algorithm of HTML Pages Based on Optimized Weight [J]. Journal of south China university of technology (natural science edition), 2011, 39(4):32-37
3. C-H. Hsu. Initial Results on Wrapping Semi-structured Web Pages with Finite-State Transducers and Contextual Rules[C]. Workshop on AI and Information Integration, in conjunction with the 15'th National Conference on Artificial Intelligence (AAAI-98), Madison, Wisconsin, July 1998
4. Bar-Yossef Z,Rajagopalan S. Template detection via data mining and its applications[C]// 11th World Wide Web Conference(WWW 2002).Hawaii,USA:[s.n.],2002
5. Yang Jun,Li Zhishu. DOM-based information extraction for WEB-pages topic [J]. DOM-based information extraction for WEB-pages topic, 2008, 45(5):1077-1080
6. DENG C; YU SP; WEN JR VIPS: A Vision-Based Page Segmentation [MSR-TR-2003-79] 2003
7. Christopher D.Manning, Prabhakar Raghavan, Hinrich Schutze. Introduction to Information Retrieval [M]. 1. Beijing: People's Posts and Telecommunications Press, 2010-01-01 :175-182.
8. Dong Shoubin, Yuan Hua. The network information retrieval[M]. 1. Xi'an: Xi'an Electronic and Science University press, 2010-4-1 :93-99.
9. MacDonald M. WPF Programming book [M]. 1. Beijing: Tsinghua University press, 2011-06-01 :694-698.
10. HTML resolve [EB/OL] http://litertiger.blog.163.com/blog/static/824538200693093340410/
11. Borenstein M. Meta. Analysis: An Introduction [M]. 1. Beijing: Science Press, 2013-01 .
12. Zhao Yanqi, Xie Xiaoxi,Xun Yuchang. Application of Naive Bias classification [J]. Electronic production,2013,(7).
13. J.Kupiec, J.Pedersen, F.Chen, A Trainable Document Summarizer in Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995:68-73,Seattle,Washington,July
14. Wu Qi,Chen Xingshu,Tan Jun. Webpage content extraction algorithm based on optimized weight [J]. Journal of South China University of Technology (Natural Science Edition),2011,(4).
15. Wang Chao,Xu Jiefeng. Webpage blocks and blocks of text extraction research based on the CURE algorithm [J]. Microcomputer and its application,2012,(12).

16. Guo, Yan ; Tang, Huifeng; Song, Linhai; Wang, Yu; Ding, Guodong Source: Advances in Web Technologies and Applications - Proceedings of the 12th Asia-Pacific Web Conference, APWeb 2010, p 314-320, 2010, Advances in Web Technologies and Applications - Proceedings of the 12th Asia-Pacific Web Conference, APWeb 2010.