基于网页框架和规则的网页噪音去除方法

时达明，林鸿飞，杨志豪

（大连理工大学信息检索研究室，大连 116024）

摘 要： 随着 Internet 的迅猛发展，Web 网页上的信息呈现爆炸式的增长，而网页噪音是任何网页都不可避免的部分，它也是影响阅读网页和进行 Web 信息处理的一个重要因素。本文提出了一种基于网页框架和规则的网页去噪方法，该方法根据网页中 HTML 标签<table>将网页分成若干部分，并对各个 table 的长宽比属性进行比较，去掉长宽比很大的部分，然后对其余 table 中的内容进行分析，根据内部是否存在和段落文字有关的标签<p>或
等来区分主题内容和噪音内容，并在此基础上，去除噪音内容。对来自 CWT200G(Chinese Web Test collection with 200 GB web pages)中的 125 个站点的 132559 个网页进行测试后的结果表明，该方法可以有效的去除网页噪音，将该方法应用于检索，索引文件可以减少约 75%，这大大地提高了检索速度。同时，准确度也得到一定提高。

关键词： 网页；噪音；HTML；框架

An Approach to Eliminate Noise Based on Framework of Web Pages and Rules

SHI Daming，LIN Hongfei，YANG Zhihao

(Information Retrieval Laboratory of Dalian University of Technology, Dalian 116024)

Abstract: With the development of the Internet, the information on the Web has exploded, and the "noise content" is an inevitable part in any web pages. It is an important factor that influences reading web pages and web information processing, too. This paper presented an approach to eliminate noise based on framework of web pages and rules. This approach separated a page into several parts according to HTML tag <table> in a web page, then compared the ratio of width and height attributes of every table and deleted the part of bigger ratio. To the rest tables, topic and noise content were differentiated according to tag <p> or
 related to paragraph, then the noise content was eliminated based on this way. Experiment results on a set of 132559 web pages from 125 different sites from the CWT200G show that this approach can eliminate noise content of web pages effectively. What's more, this approach can decrease the size of index files to about 75%, in this way, the retrieval speed can be improved obviously. Furthermore, the accuracy of retrieval can be improved, too.

Keywords: web pages; noise content; HTML; framework

基金资助：国家自然科学基金资助项目（60373095）

作者简介：时达明（1981-），男，硕士研究生. E-mail：david2004@163.com。林鸿飞（1962-），男，教授. E-mail：hflin@dlut.edu.cn。杨志豪，男，博士，E-mail：yangzh@dlut.edu.cn。

1 引言

随着互联网的迅猛发展，Web 上的网页数目已经呈现爆炸式增长。目前，Web 上的网页已经成为人们日常生活中学习知识、获取信息必不可少的来源。然而，在网页中，除了主题信息以外，还存在大量的与主题无关的导航条、广告信息、版权信息以及修饰信息等内容，这些相对于主题内容来说就是噪音内容。例如：出于商业目的而加入的广告，为了使网页美观而加入的修饰内容等等。这些内容的存在，使得准确地识别并清除网页中的噪音内容成为提高 Web 处理准确性的一项重要技术。

1.1 相关研究

在网页噪音去除方面，目前已经有大量的研究工作。Shian-Hua Lin and Jan-Ming Ho[1]提出首先根据 table 标签将网页分成若干内容块(content block)，然后将词作为特征抽取出来，并计算每个特征词的熵值，接着根据内容块中每个特征词的熵值进而计算每个内容块的熵值，最后通过与熵值的阈值比较来划分出主题内容块和噪音内容块。此种方法将页面看成是由 table 分割的集合，不过对于无 table 的网页则很难成立。

张志刚、陈静、李晓明[2]提出以一组启发式规则为基础，利用信息检索的技术以及 Web 网页的特征，提取网页的主题以及和主题相关的内容。

欧健文、董守斌、蔡斌[3]提出一种基于模板化的网页主题提取方法，该方法采用机器学习方式生成网页集的模板，以网页链接关系中的锚点文本作为提取目标对模板进行标记，生成对应模板的提取规则，依据模板的提取规则对网页主题信息进行提取。但是该方法只对模板型网页集效果显著。

封化民等[4]提出了一种新型的 Web 页面分析和内容提取框架，该框架既包括一种新型的含有位置信息的坐标树模型，还包括能反映空间关系的图模型，将 HTML 文档转化为坐标树，并结合位置特征和空间关系对网页进行分析和提取内容。

荆涛、左万利[5]提出利用网页的布局信息对页面进行划分，并在此基础上消除噪音。

孙承杰，关毅[6]提出了一种依靠统计信息从中文新闻类网页中抽取正文内容的方法。该方法先根据网页中的 HTML 标记把网页表示成一棵树，然后利用树中每个结点包含的中文字符数从中选择包含正文信息的结点。

上述方法各有各的优点和应用领域。通过参考上述文献中的方法以及对网页结构和框架的观察，本文提出了一种基于网页框架和规则的网页去噪方法。该方法首先对网页的框架结构进行分析，即通过标签<table>将网页分割成各个部分，并对 table 的长和宽进行比较，去掉长宽比很大的部分，对剩余的部分，根据定义好的一套规则，来区分主题内容和噪音内容，最后生成只含有主题内容的文本文档。

2 相关概念

2.1 网页噪音概述

网页噪音是指在一个页面内与页面主题无关（浏览者不关心）的区域及项。这些噪音包括广告栏、导航条、修饰成分等。

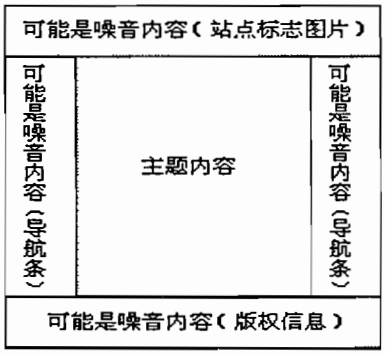


图 1 网页框架

Fig.1 Framework of web pages

网页都是有一定布局的，比如分左右两边或是中间和边缘。网页中的噪音一般都在页面中的次要位置，而将重要的内容放在网页的中间部分，这符合设计者突出网页主题的做法，同时也符合人的浏览习惯。同时，噪音部

分(例如导航条、广告、版权信息等)一般是以比较狭长的方式出现，这样在比较长宽比时可以轻易的去除，这也为本文的去除噪音算法带来了便利。如图 1 所示。

2.2 HTMLParser 类简介

HTMLParser 是一个对现有的 HTML 进行分析的快速实时的解析器。它是一个开源项目，通过它，可以准确高效地对 HTML 文本中的格式、数据进行处理。利用它可以很容易地对网页的内容进行分析、过滤和抓取。它的主要功能有：文本信息抽取、链接提取、资源提取、链接检查和内容检验等。

HTMLParser 类虽然并没有对以上提到的一些功能进行专门的处理，但是它完全可以胜任上面提及的功能，在实际应用中如果遇见上面提及的问题可以使用它内部的一些方法来处理。本文应用 HTMLParser 类来分割网页中的各个 table，得到各个 table 的 width 和 height 属性值，并对去噪后的网页进行去标签，也就是将其转化成为文本文件。

3 基于网页框架和规则的网页去噪算法

3.1 算法实现依据的规则

根据图 2 所示的一般网页的 HTML 文档以及观察其它网页的格式，可以得到一些启发式规则，如下：

i. 标签<table>和</table>之间如果有标签<p>或
，可以看做是正文内容。也就是说认为网页的主题通常是用成段的文字来描述；

ii. 若标签<table>的 width 或 height 属性为其占页面的百分比，则需要根据这个百分比的值来确定其是否为主题内容。若 width 或 height 属性的百分比数值较大，则认为有可能是主题内容；

iii. 对于多层嵌套的标签<table>，认为只在其中某一层 table 中存在主题内容；

iv. 对于没有标签<table>的网页，即不是由表格分割的网页，如果存在段落文字，则认为是主题内容。

3.2 算法思想

对于有标签<table>的网页，认为重要的信息都放在网页的中间区域，而且该区域长度和宽度都比较大。而网页边缘区域的重要性相对于中间区域都很弱，而且该区域比较狭长；对于没有标签<table>的网页，只是根据其是否存在段落文字来判断是否为主题内容，并没有考虑更多，这是因为对于大多数网页设计者来说，通常他们会先进行网页布局的设计，即先用表格等标记在页面上描绘出页面内容分布的区域，然后再在每个区域内部进行详细的内容设计，把需要的元素加进去。所以没有表格(table)的网页现在越来越少，这必然是以后网页制作发展的主流方向。本文就是根据这样一种思想来进行网页去噪，然后将提取出的主题内容变为文本文件，为以后对网页的一些处理，如分类、检索等提供了很大的方便。

3.3 算法流程

(1) 由于本文使用的是天网的 CWT200G 语料，而天网存储语料的格式又有其自身的特殊性，所以，首先根据天网存储网页的格式进行语料的切分，将其切分成单一网页的形式。

(2) 切分完毕后，开始进行去噪工作。对于没有 table 的网页，处理方法如图 3 所示。

```
If (无标签<table>){
    If (存在段落文字){
        认为是主题内容，保留
    }
}
```

图 2 无 table 网页的去噪算法

Fig.2 eliminating noise content algorithm of no table web pages

(3) 对于有 table 的网页，处理方法如图 4 所示。

```
For (对于每一个 table 表格){
    得到 table 的 width 和 height 属性
```

```
If (width 和 height 属性有一个不存在且 width 不以百分比的形式出现){
    If (width 属性值> δ1)  // δ1 为 width 属性阈值
      If (有段落文字)
          认为是主题内容，保留
}
Else if (width 和 height 属性都存在且 width 属性以百分比的形式出现){
    If (width 百分比数值> δ2)  // δ2 为 width 百分比数值的阈值
      If (有段落文字)
          认为是主题内容，保留
}
Else if (width 和 height 属性都存在且 height 属性以百分比的形式出现){
    If (height 百分比数值> δ3)  // δ3 为 height 百分比数性阈值
      If (有段落文字)
          认为是主题内容，保留
}
Else if (width 和 height 属性都存在且 width 和 height 属性值以数值的方式出现){
    计算 width 与 height 的长宽比
     If (长宽比很小)
        If (有段落文字)  //该层判断是为了防止出现长宽比很小的图片等
                         //非主题内容
          认为是主题内容，保留
 }
 Else{  //对于没有段落文字，即对于目录型网页或图片型网页
    认为 table 长宽比很小，但长和宽的值占页面很大的部分为主题内容，保留
 }
}//endfor
```

图 3 有 table 网页的去噪算法

Fig.3 eliminating noise content algorithm of table web pages

(4) 将去噪后的网页去标签，并转化为文本文件。

3.4 实验结果及分析

本文对来自CWT200G(Chinese Web Test collection with 200 GB web pages)中的125个站点的132559个网页进行测试。经过测试，网页大小由去噪前的6.04G减少到了去噪后的1.45G，网页噪音部分竟然占了网页约75%的大小！可想而知，对 Web 网页进行分类或检索等其它应用时，如果不事先进行消除噪音，那么对于分类或检索的速度和精度将会有多么大的影响。

同时，本算法在 Windows XP 系统、Pentium(R) 4 CPU 2.8GHz、1G 内存的计算机上进行了实验。经过实验统计，对所有实验数据去噪的平均速度为 18 个/秒。可见，该算法能较快速地完成网页噪音消除工作。

在准确率方面，由于测试的网页较多，不可能对每个网页都进行准确率检查，所以随机抽取 2000 个网页进行手工检查。现将检查结果用“优、良、中、差”四个标准进行判断。其中“优”代表网页主题内容正确提取，且噪音基本去除；“良”代表网页主题内容正确提取，噪音存在一部分；“中”代表网页主题内容基本能正确提取，噪音存在较多；“差”代表网页噪音基本没有消除或者主题内容没有正确提取。实验结果如表 1 所示。

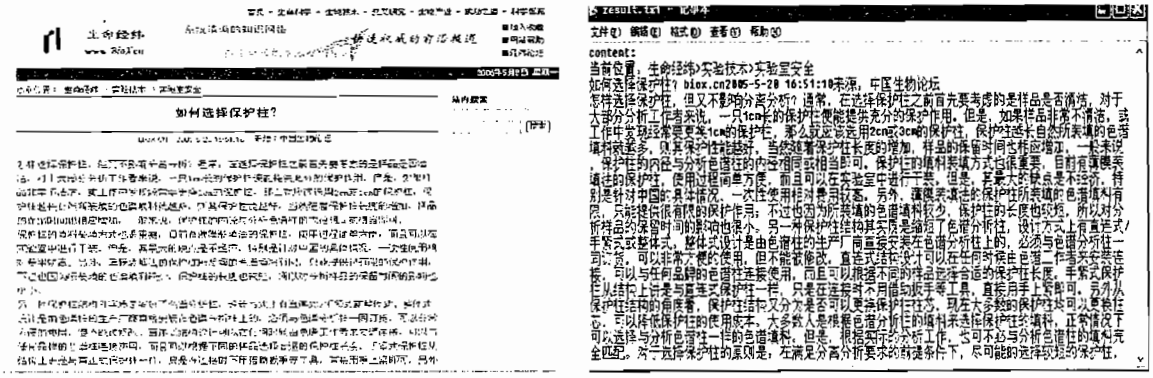
表 1 去除噪音结果

Tab.1 Results of eliminating noise content

测试标准	结果(%)

优	49.2
良	32.1
中	15.9
差	2.8

下面以 CWT200G 中的一个网页为例来显示去除噪音前的网页和去除噪音后的文本文件，如图 5 所示。



2.1. 处理前　　1.2. 处理后

图 4 处理前后结果对比

Fig.4 Comparison before and after processing

在对去噪效果较差的网页的观察中，可以发现以下一些原因：首先，网页的主题内容不明显，即主题内容只有一句话，或者是网页中的主题是图片信息等等。其次，一少部分网页的噪音中也存在一些段落文字的内容，也就是说存在<p>或
等认为是段落文字的标签，这样会误把这部分内容看做主题内容保留下来而没有被清除掉。算法已经对类似这样的错误进行了一定的处理，因为噪音中的段落文字通常比主题内容中的段落文字少，即噪音中的标签<p>或
的个数会比主题内容中的少，所以，针对这种情况，可以根据标签<p>或
的个数来判断该部分是噪音还是主题内容。

4 结束语

针对目前对 Web 信息处理的大量应用，本文提出了一种基于网页框架和规则的网页去噪方法。实验结果表明，该方法可以迅速地从网页中提取出主题内容并清除噪音，且清除噪音的准确率较高。同时，本文的方法也符合网页设计者的设计习惯，即将主题内容放在网页中间部分，且占用篇幅较大；而将噪音部分放在网页边缘，这些区域占用篇幅较小且比较狭长。实验结果也证明本文的方法是有效的。将本文的方法应用到搜索引擎方面，可以大大地减少索引量、提高搜索引擎的检索速度和检索的准确度；应用到分类方面，可以将 Web 网页中的主题内容提取出来，存放到文本文件中，然后就可以很方便地应用目前现有的分类器进行自动分类。但是，本文的方法还有一些网页不能处理或处理的效果较差，同时算法中的阈值是在不断地实验中得出的，其合理性还有待进一步实验和观察。因此，如何完善本文算法的不足，使其具有更强的通用性和适应性，将是今后的努力方向。

参考文献：

[1] Lin S-H，Ho J-M. Discovering informative Content Blocks from Web Documents [A]. Proceedings of the ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining (SIGKDD'02) [C]. 2002.

[2] 张志刚，陈静，李晓明. 一种 HTML 网页净化方法[J]. 情报学报，2004，23(4)：387-393.

[3] 欧健文，董守斌，蔡斌. 模板化网页主题信息的提取方法[J]. 清华大学学报，2005，45(S1)：1743-1747.

[4] 封化民，刘飚，刘艳敏，等. 含有位置坐标树的 Web 页面分析和内容提取框架[J]. 清华大学学报，2005，45(S1)：1767-1771.

[5] 荆涛，左万利. 基于可视布局信息的网页噪音去除算法[J]. 华南理工大学学报（自然科学版），2004，32：84-87.

[6] 孙承杰，关毅. 基于统计的网页正文信息抽取方法的研究[J]. 中文信息学报，2004，18(5)：17-22.