一种基于特征符号的网页主题信息抽取方法

王 舒[1], 朱 敏[1], 张 明[2], 牛 颢[1], 赵 瑜[1]
(1. 四川大学 计算机学院, 成都 610064; 2. 四川省计算机研究院, 成都 610041)

摘 要: 随着 Internet 网络的日益普及,Web 上的海量数据给文本挖掘尤其是网页主题提取带来了更多的挑战,现有的文本提取方法在保证高准确率的同时无法满足 Web 挖掘方法的通用性。通过对 Web 网页结构进行研究,对网页生成树模型进行了改进,找到网页结构的通用规则,提出一种基于特征符号的提取方法 CECS(content extraction characteristic symbols),结合相关度对网页主题内容进行提取。实验证明,所提算法具有很高的准确性和通用性。

关键词: 生成树模型; 特征符号; 相关度; 主题提取
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2009)12-4539-03
doi:10.3969/j.issn.1001-3695.2009.12.039

Content extraction of Web pages based on characteristic symbols

WANG Shu[1], ZHU Min[1], ZHANG Ming[2], NIU Hao[1], ZHAO Yu[1]
(1. *College of Computer Science, Sichuan University, Chengdu* 610064; 2. *Sichuan Institute of Computer Sciences, Chengdu* 610064, *China*)

Abstract: With the popularity of the Internet, the large amounts of data on the Web provides many challenges for data mining techniques, especially for content extraction of Web pages. The existing methods can not guarantee the generality and effectiveness of Web mining approaches. By studying the internal structure of Web pages, this paper proposed an improved document tree model and discovered the general rules for analyzing it. In addition, extracted content from Web pages based on characteristic symbols. The experimental results prove that the proposed method is accurate as well as generic.

Key words: document tree model; characteristic symbols; relevance; content extraction

随着 Internet 的迅猛发展,Web 上的数据呈海量增长,使从 Web 文档中提取有价值信息变得更加困难。Web 文档中除主题信息外往往包含很多噪声内容,如广告信息、超链接、图片和 Flash 等,这些噪声给 Web 主题信息检索带来了很大干扰。研究[1]表明,通过提取主题信息可以减少一半的浏览时间。因此,网页主题内容的提取当前已经成为 Web 信息处理中的研究热点[2]。

1 相关工作

传统的网页信息抽取方法使用包装器(wrapper)来抽取网页中有关的数据。它根据一定的信息模式从特定的信息源中抽取内容,而且一个包装器只能针对一个信息源。由于网页结构复杂多变,包装器很难满足通用性的需要。

当前,国内外关于网页噪声去除的研究比较深入,已经提出了诸多方法,比较有代表性的包括微软亚洲研究院提出的基于页面可视化信息的 VIPS(vision-based page segmentation)方法来提取网页正文,它主要利用字体的大小、布局信息等一些页面的视觉特征将其分成各个视觉信息块。

文献[3~5]是对 VIPS 方法的改进。此方法对符合人们观察习惯的网页是很有效的,但是由于视觉特征的复杂性,很难找到一个通用规则集。

文献[6~8]采用了基于模板的方法,但是一个模板只能针对一类网页,如果针对多类网页就必须构造模板集,这使得模板的开发和维护等工作非常复杂。

目前效果最好的 Web 主题抽取方法是基于网页结构化信息的正文抽取方法,根据样式集构造 DOM 树,通过遍历 DOM 树提取网页主体部分。文献[9]首先提出了内容块(content block)的概念,利用〈table〉标签将网页划分成块。文献[10,11]继续深化了这一思想,并提出了一组启发式规则,利用信息检索方法,通过〈table〉标签提取网页的主题,但如何衡量正文没有给出一个明确的方法。在文献[12]中,作者将中文个数作为正文的衡量标准,假设正文中出现的中文个数是最多的;文献[13]通过字符个数与超链接个数的比值作为标准。以上两种方法错误率高,均不能作为一种普通方法。中科院计算所软件研究室提出了利用 Table 标记和视觉特征对页面进行语义块划分并识别各语义块属性的算法 TVPS(table and vision based page segmentation)[14],该方法的分块算法中只考虑了把各个最底层的〈table〉作为标记。

以上方法的分块算法通用性差,对于没有〈table〉标签的

收稿日期: 2009-01-05; **修回日期**: 2009-04-20

作者简介:王舒(1984-),女,吉林吉林人,硕士研究生,主要研究方向为计算机网络、信息系统(wangshujcferrero@163.com);朱敏(1971-),女,教授,博士研究生,主要研究方向为计算机网络、信息系统;张明(1978-),男,河北大成人,工程师,主要研究方向为计算机网络;牛颢(1983-),男,陕西西安人,硕士研究生,主要研究方向为计算机网络、信息系统;赵瑜(1980-),女,四川西昌人,硕士研究生,主要研究方向为计算机网络、信息系统.

文本无法处理,在正文提取方面也没有一个准确率相对较高的方法。

文献[15]提出了一种以中文句号为判断依据的正文提取方法,准确率相比其他方法有很大的提高,但是对 Web 文本的网页复杂结构考虑不够全面。本文在此基础上,提出了网页生成树的改进模型,对网页结构的通用规则进行了描述,并结合相关度与中文标点符号等特征对网页主题信息进行了提取。

2 信息抽取系统框架

信息抽取系统即 Web 主题文本抽取系统,由文件解析器、剪枝器、分析提取器和主题清洗器四个模块构成[13],如图 1 所示。本系统工作原理为:文件解析器负责将采集到的网页源文件解析成 DOM 树的形式,然后交给剪枝器进行处理;剪枝器将 DOM 树中与提取正文主题不相关的标签过滤掉,得到主题树;分析提取器从主题树中提取正文部分后交给主题清洗器除掉正文中混入的多余字符,最终得到只包含主题内容的文件。

其中剪枝器和分析提取器是最重要的两个模块,它们应用的提取算法会影响整个处理结果的准确性。

3 DOM 树的构造及精简

DOM(document object model)是 W3C 制定的标准接口规范,它提供了访问页面中各个元素属性与方法的接口。图 2 给出了传统的 Web 文档对应的 DOM 树,正文包含在〈body〉分支下。通过对 DOM 树的遍历,可以对页面中每一个元素进行处理。

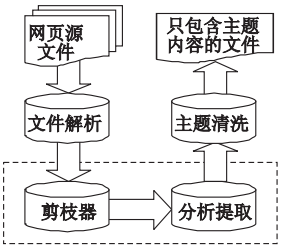

图1 信息抽取系统框架图

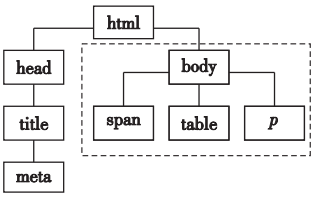

图2 Web文档对应的DOM树

Web 网页被解析成 DOM 树后,内容块是由特定的标签规划出的。通过大量研究,本文总结出正文出现位置可分为如下两类情况,构造出正文标签树如图 3 所示(*c* 代表正文)。a)正文存于〈table〉标签下的〈div〉标签中,如图 3(b)所示;b)正文存于独立〈div〉标签下,或者〈div〉标签下的〈table〉标签之中。其中正文可能与〈table〉标签嵌套出现,如图 3(c)所示。

由于构造完成的 DOM 树中包含大量与提取正文无关的内容,需要对树进行剪枝,即去掉无关的标签与注释。通过上述分析,正文文本应用如下步骤进行剪枝:

a)剔除不包含在〈body〉标签内的全部内容。

b)保留顶层〈div〉、〈table〉标签及其内部内容,剔除其余内容。

c)去掉网页的注释信息〈!--〉中的内容。

e)〈script〉〈noscript〉〈input〉〈button〉〈link〉〈style〉〈select〉〈embed〉〈object〉〈img〉〈iframe〉〈form〉〈applet〉〈textarea〉这些标签用于控制文件交互性和显示,与正文无关,剔除后可以大大提高检索速度。

e)如果一个节点不是文本节点且没有子节点,则把该节点删除。

4 主题提取算法——CECS

进行语义分析之前首先要根据分块节点将网页分割成小的信息块。分块节点的选择决定了分块粒度的大小,粒度过粗会导致提取结果不明确,粒度过细会导致抽取结果不完整。根据对正文标签树的研究,本文采用〈div〉标签对文本分块。分块流程如图 4 所示。

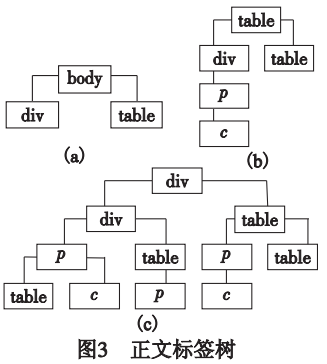

图3 正文标签树

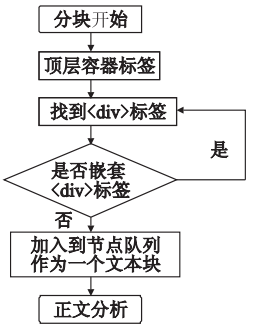

图4 分块算法流程图

分块后进一步区分正文块与噪声块。根据文献[15],中文句号总数的90%以上出现在正文中,因此本文将中文句号的个数作为衡量正文的一个重要特征,用 stopcount 属性表示。网页中经常出现大篇幅文字类广告和超链接,给正文提取带来很大的干扰,因此引入相关度的概念。

定义 相关度。一个块(block)节点的主题相关度(correla-tivity)由块内链接个数(linkcount)和块内中文句号个数(stopcount)决定,形式化定义如下:

$$\text{correlativity}(\text{block}_i)=\begin{cases}\dfrac{\text{linkcount}(\text{block}_i)}{\text{stopcount}(\text{block}_i)}\text{stopcount}(\text{block}_i)\neq 0\\ 1 \qquad\qquad \text{stopcount}(\text{block}_i)=0\end{cases} \tag{1}$$

$$\text{linkcount}(\text{block}_i)=\sum_{i=1}^{N}\text{linkcount}(\text{block}_{ij}) \tag{2}$$

$$\text{stopcount}(\text{block}_i)=\sum_{i=1}^{N}\text{stopcount}(\text{block}_{ij}) \tag{3}$$

此相关度公式是根据广告和超链接特征来判断的。在 Web 文本中,独立的广告装在单独的文本标签中,一条广告伴随着一条或多条链接。广告通常是言简意赅的句子,其中文句号个数很少甚至没有,因此用块内链接数和块内中文句号个数的比值作为判断是否是广告块或超链接块。

算法 1 主题提取算法 CECS

```
输入:block 文本块
输出:block_i//正文块
for( $i=0; $i<sum(〈div〉); $i++ ) //遍历每个〈div〉块
  {
    if( correlativity(block_i) <1) //相关度小于 1 则非广告块
   {
    get(block_i); //得到正文块
   }
  else //相关度大于等于 1 则判断为广告块,过滤掉
   {
    next;
   }
  }
```

最后将提取出的正文信息采用正则匹配去噪声的方法去掉多余字符,得到只包含主题内容的文件。

算法 2　主题清洗算法

```
输入:含有多余字符的正文
输出:不含多余字符的正文
sub clearHTMLElements //清除主题中多余字符
{
  my( $temp) = @_;
  $temp = ~ s/\<Ahref\ = \"http\:. * \" \>//gsi; //去掉多余链接
  $temp = ~ s/\< \/A\>//gsi;
  $temp = ~ s/\<BR\>//gsi; //去掉多余换行符
  $temp = ~ s/\ \;//gsi; //去掉多余空格符
  $temp = ~ s/\<B\>//gsi; //去掉多余加粗符号
  $temp = ~ s/\<SPANid\ = post\d + \>//gsi; //去掉多余 span 标签
  $temp = ~ s/\< \/SPAN\>//gsi;
  $temp = ~ s/\<FONTsize\ = \d + \>//gsi; //去掉多余 font 标签
  $temp = ~ s/\< \/FONT\>//gsi;
  return $temp;
}
```

5　实验结果

为了验证本文算法的有效性,测试数据来自于 10 个门户网站。由于同板块内部文本结构基本相同,而板块之间差异较大,本文从每个网站的新闻、财经、体育、娱乐等板块随机抽取了若干篇网页作为测试集进行测试,最后结果与人工抽取的正文内容进行对比分析。同时,本文用以往方法对上述网页进行测试,以往方法中文献[15]的准确率对比其他方法有很大提高,更具对比性。实验结果如表 1 所示。

表 1　实验结果

数据来源	网页总数	本文方法/%	文献[15]方法/%
www. sina. com. cn	777	95.6	95
www. sohu. com	651	96.0	96
www. 163. com	593	91.9	90
www. tom. com	478	97.4	96
news. china. com	115	91.3	90
www. chinaren. com	120	90.0	86
www. 21cn. com	120	83.3	83
www. yahoo. com. cn	462	96.3	86
www. chinadaily. com. cn	120	96.6	95
www. qq. com	350	88.2	79
总计	3 786	94.0	89.6

由表 1 可知,用本文提出的算法进行网页正文提取的准确率最高为 97.4%,最低为 83.3%,平均为 94.0%,相比同类算法准确率有一定提高。对比以往方法,本文测试所用网页数量多,并且板块内容丰富。与文献[15]相比,本文采用的文本分割符合绝大多数网页的组织结构和内容布局。另外,本文采用的主题提取方法应用于中文的主题内容提取更加准确。为进一步分析错误的网页原因,本文对每个网站网页的各个板块的准确率情况作了研究,以新浪网为例,如表 2 所示。

表 2　新浪网实验结果

数据来源	网页总数	正确个数	错误个数	准确率/%
社会新闻	154	150	4	97.4
军事新闻	126	125	1	99.2
财经	170	163	7	95.8
娱乐	105	94	11	89.5
体育	94	85	9	90.4
科技	128	126	2	98.4

通过测试,新闻、财经、科技板块的准确率均超过了 95%,而娱乐和体育板块相对较低。经分析、测试,由于娱乐、体育板块的新闻经常以图片作为叙述主题,其中穿插了若干文字对图片进行描述,并且伴随着大量超链接,特征符号不明显,无法正确提取出主题内容。

6　结束语

本文针对 Web 文本主题进行提取,在文本分割的基础上,提出了一种新的 DOM 树构造模型,克服了以往采用〈table〉标签进行分割的不足。在正文提取方面,提出了基于特征符号的抽取方法 CECS,并将特征符号与网页相关度结合提取出网页主题,实验结果表明该方法具有很高的准确性和适用性。未来研究工作主要针对提取不成功的网页作进一步研究,找出其结构规律,将准确率进一步提高。

参考文献:

[1] BUYUKKOKTEN O, GARCIA-MOLINA H, PAEPCKE A. Accordion summarization for end-game browsing on PDAs and cellular phones[C]//Proc of ACM Conference on Human Factors in Computing Systems. New York: ACM Press, 2001:213-220.

[2] BERRY M W, BROWNE M. Understand search engines: mathematical modeling and text retrieval philadelphia[C]//Proc of SIAM. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1999:116.

[3] CAI Deng, YU Shi-peng, WEN Ji-rong, *et al*. Extracting content structure for Web pages based on visual representation[C]//Proc of the 6th Asia Pacific Web Conference. Xi'an: Springer, 2003:406-417.

[4] YU Shi-peng, CAI Deng, WEN Ji-rong. *et al*. Improving pseudo-relevance feedback in Web information retrieval using Web page segmentation[C]//Proc of the 12th International World Wide Web Conferenceon. 2003:11-18.

[5] 时达明,林鸿飞,杨志豪. 基于网页框架和规则的网页噪音去除方法[J]. 计算机工程,2007, 33(19):276-278.

[6] WANG Ji-ying, LOCHOVSKY F H. Data-rich section extraction from HTML pages[C]//Proc of the 3rd International Conference on Web Information Systems Engineering. Singapore : IEEE Computer Society, 2002:313-322.

[7] 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:自然科学版, 2005, 45(9):1743-1747.

[8] 陈枫. 基于 TABLE 布局和隐马尔可夫模型的 Web 自由文本信息抽取[D]. 杭州:浙江大学,2007.

[9] LIN Shian-Hua, HO Jan-ming. Discovering informative content blocks from Web documents[C]//Proc of the 8th ACM SIGKDD International Conference on Konwledge Discovery and Data Mining. New York:ACM Press, 2002:588-593.

[10] 李效东,顾毓清. 基于 DOM 的 Web 信息提取[J]. 计算机学报, 2002, 25(5): 41-62.

[11] GUPTA S, KAISER G E. DOM based content extraction of HTML documents[C]//Proc of the 12th World Wide Web Conference. 2003:207-214.

[12] 黄文蓓,杨静,顾君忠. 基于分块的网页正文信息提取算法研究[J]. 计算机应用, 2007, 27(Z1):101-154.

[13] 王志琪,王永成. HTML 文件的文本信息预处理技术[J]. 计算机工程,2006,32(5):46-48.

[14] CAI Deng, YU Shi-peng, WEN Ji-rong, *et al*. VIPS: a vision-based page segmentation algorithm, MSR-TR-2003-79 [R]. 2003.

[15] 吴新涛. 基于向量空间模型的网页信息过滤方法研究[D]. 大连:大连理工大学,2008.