基于结构树解析的网页正文抽取方法[1]

刘秉权[1] 王喻红[2] 葛冬梅[3] 李佳[4]

[1,4]哈尔滨工业大学计算机科学与技术学院，哈尔滨，150001

[2,3]黑龙江工程学院计算机科学与技术系，哈尔滨，150050

摘要

本文采用一种基于结构树解析的方法来实现中文网页正文的抽取。这种方法利用了中文网页中内容信息结构相似和分布聚集的特性，实现简单，通用性好，可以克服包装器方法需要针对特定数据源的缺点。该方法在分析网页时是利用 DOM(Document Object Model) 的树结构来进行的，通过把网页解析为 DOM 树使分散的网页有序化。目前该方法已经应用到面向移动平台的新闻信息自动分类系统中，很好地满足了系统的需求。

关键词：计算机应用；DOM；网页数据抽取；包装器

中图分类号：TP391.3

Extracting Text Content from Chinese Web Pages Based on Parsing Structural Tree

Liu Bingquan[1], Wang Yuhong[2], Ge Dongmei[3], Li Jia[4]

[1,4]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001

[2,3]Department of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, 150050

Abstract

This paper adopts an approach for extracting text content from Chinese Web pages based on parsing structural tree. By using similar structure and collective distribution of Chinese web pages content, the method is simple, accurate and easy to be implemented. In comparison with traditional methods, the method needn't construct different wrappers for different data sources. The method parsed Web pages by using tree structure of DOM, which made dispersed content orderly and extracted content from Web pages more simply. Now, the method has been applied to a News Content Auto-Classify system of mobile platform.

Keywords: computer application; DOM; web data extraction; wrapper

1 引言

随着Internet技术的发展和普及，Web网页信息迅速成为人们获取信息的主要渠道之一。而网页内容信息一般是由导航信息、网页正文、广告信息、版权信息、相关链接等部分组成的。面对这些繁杂的信息，人们通常不能迅速准确地找到自己所关心的正文信息。有效地从浩瀚的信息海洋中挖掘可以为人所用的各种知识，提取人们所需的信息，已成为信息处理领域的一个研究热点。

目前互联网上的网页绝大部分是用 HTML 语言编写的。HTML 语言提供的标记主要是用来控制网页内容的显示格式，这些标记的使用缺少规律，网页设计人员可以随便设计。但是不同网站内部的网页大多都是由一套相同的内容模板生成的，因此可以认为网页的设计是有相对规律的。经过对各大新闻网站的网页进行观察得到这样一个事实：网页中各类信息一般都具有相同的结构，即以相同的标记或格式来显示的，如集中在<table>、<p>、<li>内；而且网页中的内容信息多是以聚集的形式存在，即有用信息内容相对集中，无用信息也相对集中，如导航栏就是由多个<a>链接标记集中显示形成的，文本内容则以<p>或 等标记集中显示。

随着需求的增加，近来涌现出了多种信息抽取工具，采用的技术也各不相同，涉及多个研究领域，如：数据库、人工智能、信息检索等。文献[1]中的 WHISK 系统是利用自然语言处理的技术，根据语义项的上下文实现感兴趣信息的定位。这种基于自然语言理解方式的信息抽取技术，是将 Web 文档视为文本进行

[1] 论文得到黑龙江省自然科学基金（E200635）资助

处理的（主要适用于含有大量文本的 Web 页面），抽取的实现没有利用到 Web 文档不同于普通文本的层次特性。获得有效的抽取规则需要大量的样本学习。文献[2]介绍的 RoadRunner 工具是基于包装器归纳方式的一个完全自动化的包装器自动生成工具。它通过比较来自同一数据源的两个（或多个）样本网页的结构来为包含在网页中的数据生成一定的模式。该方法假设目标网页都是从某个数据源自动生成的，那么它就可以利用网页的标记结构重新得到网页中包含的数据的模式，所以其适用范围有一定的局限性。文献[3]中认为网页是可以收缩和扩展的，定义了一种生成文摘的策略。利用了一种语义文本单元(Semantic Textual Units)划分的方法，把网页划分为一个具有层次结构的独立的内容单元。然后把这些语义单元根据 HMTL 的格式排列为一个层次结构。这种分级结构的文本最终被显示在 PDA 或手机设备上。文献[4]中提出的方法是利用计算每个内容块的信息熵来抽取出有用的信息块。首先根据<TABLE>标记把网页分成多个内容块。在训练集上统计网页中出现特征词的概率，计算这些特征的信息熵，并根据训练集得到的信息熵的阈值决定信息块的抽取。该方法需要在大量的训练集上进行统计特征词的概率，并且假设网页内容都集中出现在<TABLE>内，因此其方法有一定的局限性。

上面介绍的几种方法在应用上都有一定的局限性。从文献[5]中,我们可以看到,目前的网页数据抽取工具,都需要针对特定的数据源来编写对应的包装器或抽取规则。即针对某一特定的网站的网页书写包装器，这样就带来了程序适应性不强的问题：如果目标网站有所改动，必须随时修改程序，而且往往修改调试起来非常繁琐；如果抽取对象是多个网站，就需要书写多个包装器，这样包装器的维护是一项很复杂的工作。

2 基于结构树解析的网页正文抽取方法的研究

本文采用了一种基于结构树解析的方法来实现中文网页正文的抽取。这种方法利用了中文网页中内容信息结构相似和分布聚集的特性，实现简单，通用性好，并结合了前面信息抽取研究方法的优点。为了分析网页，首先通过 HTML 解析器把网页解析成一个 DOM(Document Object Model)树的结构。通过把网页解析为 DOM 树使分散的网页有序化，会使内容信息的抽取更加简单和灵活。DOM 树具有很强的可塑力，能够很容易地重构一个完整的网页。在对树进行操作时可以充分地利用数据的分级层次结构，而在对一个普通结构的文件进行分析时则不会具有这种数据结构的特点。

此方法大致分为以下几个步骤：

步骤 1、网页的预处理。

HTML 文档的书写是不规范的，为了使网页能够正确地被解析必须先对网页的标记进行修复，使其符合标准的语法规范。这里主要是通过 HTMLTidy 来进行验证，对不正确的 HTML 句法进行修复，之后得到的是规范的 HTML 文档。 规范网页的要求如下：

1) “<”和“>”只能用来包含网页标记（tag），当在其它地方出现这两个符号时应该用<和>代替。

2) 所有的标记必须匹配。即每个开始标记都对应一个结束标记。

3) 所有标记的属性值都必须放在引号中。如。

4) 所有的标记必须是正确嵌套的。如<a>………是不正确的嵌套。正确的嵌套形式应该是<a>………。

步骤 2、网页解析——根据网页的 HTML 标记，把网页表示成一棵树。

这里将经过规范化处理后的网页解析为一棵标记树，采用的是 DOM 方法来解析网页，解析之后得到一棵以 HTML 为根节点的 DOM 标记树。树中的每个节点都是由网页中的所有标记属性对构成的。

步骤 3、网页正文抽取算法的应用。

在 DOM 解析树上应用算法来抽取网页正文。下面给出该算法的具体描述：

该网页正文抽取算法以一个四元组(s,w,n,p)的形式来表示树中第 s 个含有信息量为 w、标记为 n、并且其父亲节点的标记为 p 的节点。通过在树 T 中找到一个节点序列的集合来进行抽取正文信息。其中每个节点序列为

$$queue(X_1,\dots,X_n)=\left\{\begin{array}{l} X(s,w,n,p)\,|\,\forall X_i,X_j\in T,0<i<j\le n,\\ \text{满足 } s_i-s_j=j-i,\text{同时满足 } n_i=n_j,p_i=p_j \end{array}\right\}。$$

其中四元组(s,w,n,p)的位置序号 s 表示节点在树中从左到右从上到下的位置顺序，根节点位置 s 的值为

0；信息量 w 是指树中节点所含有中英文字符的数量；n 是树中该节点的 HTML 标记值；p 是树中该节点的父亲节点的标记值。在树中所要寻找的是一个节点序列的集合，其中每个节点序列中的节点标记 n、父亲标记 p 都必须相等，并且任意两个节点 X_i 、X_j 在树中所处位置的距离为他们在该节点序列中的序号之差，即 s_j-s_i=j-i，i、j 表示节点在序列中的位置序号。

根据网页标记功能的不同可以把 HTML 标记分为两类，一类是专门用来构建网页框架的结构标记，一类是专门用来修饰网页文本的格式标记。因为我们只关心网页的结构分布，所以要把对网页文本进行修饰的格式标记如字体标记<font>、文本格式标记<b>、、<small>、<big>、、<I>、<b>、<u>、文本样式标记<style>等过滤掉，在对标记节点进行四元组表示过程中，只考虑结构标记，如<p>、 、<table>、<tr>、<td>、<li>、等。

算法的语言描述为：

1) 首先广度遍历标记树 T，生成树中所有结点的四元组(s,w,n,p)表示，并求出 w 值最大的节点 Max。
2) 遍历结束后按顺序生成了一组满足上面算法条件的节点序列，之后对得到的序列集合进行处理：
   - 去掉标记为<a>或者节点信息量 w 小于 q 的节点序列（在我们的系统里 q 取值为 16）；
   - 对序列集合 set 中只含有一个节点的序列进行合并或过滤，对所有单节点序列中的节点的父亲节点来应用算法考察，把父亲节点符合算法条件的节点进行合并，生成新的队列。如果不符合条件的节点不是 Max 节点，则进行过滤删除；若是 Max，则保留该单节点。过滤结束，输出集合 set 中的剩余节点序列中所含节点的全部信息值，得到的即是抽取的网页正文信息。

3 实验方法与结果分析

为了考察方法的实际效果，我们进行了下面的实验。

实验数据：从表 1 所列的网站中随机选择了一些网页。实验结果如表 1 所示。

通过对出错的网页的考察，我们发现，大部分出错的网页有一个共同特点：它们包含的正文信息不规范，通常含有一些个人评论信息，这些评论信息不但内容很多，而且结构一致，以致得不到正确的结果。

从表1的实验结果可以看出，该方法在具有通用性的同时，保持了较高的准确性。如果网页是比较规范的新闻类网页，其准确率可以达到98%。这说明这种方法具有实用性。在实际工作中，我们对来自80个网站的网页进行了抽取，抽样统计的准确率均在93 %以上。

表 1 实验结果

数据来源	网页总数(个)	正确的结果（个）	错误的结果（个）	准确率（%）
news.qq.com	80	78	2	98
news.sohu.com	70	68	2	97
news.sina.com.cn	50	49	1	98
news.163.com	50	45	5	90
news.cctv.com	50	46	4	92
www.southcn.com	50	45	5	90
msn.ynet.com	100	96	4	96
www.zaobao.com	50	48	2	96
news.people.com.cn	100	95	5	95
www3.xinhuanet.com	100	94	6	94
总计	700	664	36	95

4 结束语

该方法目前主要是应用在面向移动互联网的新闻信息获取分类系统中，因此其中的一个阈值是在该系统的环境下设定的。阈值的选取以及对结果的影响都应该进一步的探讨，以达到适合不同应用场合的数据。在上述系统中结合了本文的抽取方法与中文信息处理的自动文本分类方法，实现了新闻的自动分类，还可以尝试把其他的中文信息处理技术应用扩展到网页处理中来。

参考文献:

1 Soderland S. Learning information extraction rules for semi-structured and Free Text [J]. Machine Learning,1999,34(1-3):233-272

2 Crescenzi,V., Mecca,G., etc. RoadRunner: Towards Automatic Data Extraction from Large Web Site. In proceeding of the 26th International Conference on very Large Database Systems[C]. 2001.

3 O. Buyukkokten, H. Garcia-Molina and A. Paepcke. Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones. In Proc. of Conf. on Human Factors in Computing Systems (CHI'01), 2001

4 Lin,SH,and Ho,JM. Discovering Informative Content. Blocks from Web Documents. In Proceedings of ACM. SIGKDD, 2002.

5 Alberto H. F. Laender, Berthier A. RibeiroNeto, Altigran S. da Silva and Juliana S. Teixeira. A Brief Survey of Web Data Extraction Tools[J] . SIG2 MOD Record. 2002 ,31(2) :84-93.