Journal of Wuhan Polytechnic

基于文本标签属性的网页信息抽取方法研究

沈　娜

（宿迁开放大学，江苏　宿迁　223800）

摘　要：伴随着互联网的飞速发展，网络上的信息资源呈现出井喷态势，如何从海量的信息中抽取出自己需要的信息已经变得越发的困难。在分析现有 Web 信息抽取技术现状及面临的挑战的基础上，设计了一种基于文本标签属性的 Web 新闻信息抽取模型。主要介绍了基于标签的 Web 信息抽取技术的算法，给出了信息抽取的具体实现过程，对基于 DOM 树节点遍历的文本标签过滤算法进行了描述，并选取了主流的新闻网站进行了抽取实验，验证了算法的可行性。

关键词：HTML DOM 树；文本标签属性；Web 新闻；信息抽取

中图分类号：TP391.1　　**文献标识码：**A　　**文章编号：**1671-931X（2016）01-0062-04

一、引言

互联网的飞速发展使得 WWW 成为一个庞大的信息空间，为人们提供着丰富的信息资源。大量的信息资源通常是以网页的形式呈现出来，网页中存在着大量的与我们所关注的内容无关的信息，如广告信息、版权信息以及导航条等等，这些我们称之为“网页噪音”，通常以链接导航的形式出现在主题内容周围或者主题内容的中间，这些“噪音”严重影响着人们对于信息的准确获取，如何从网页中抽取出正文内容，避开一些不相干的信息干扰，已经成为 WEB 智能领域中的一个重要课题。

二、Web 信息抽取技术

（一）Web 信息抽取技术研究现状

1.基于统计理论

基于统计理论的信息抽取技术获取网页的信息主要通过两种方式：第一种是通过对各个标签所包含的信息量的多少进行统计，将包含信息量最大的标签对比出来，通过获取这些标签的内容进而得到网页的主要信息；第二种是计算网页中链接文本和普通文本的比值来确定抽取的阈值[1]，一般情况下，链接文本多为信息噪音应该过滤掉。基于统计理论的技术克服了数据源的限制，不需要针对某个具体的网页进行定制，节省了工作量，具有一定的普遍性。该技术主要适用于那些以文本为主的网页，如果网页中含有大量图片或者视频内容将使抽取的结果产生较大偏差，文本的数量应该具有明显的优势。同时，基于统计理论的信息抽取技术虽然理论上易于实现，但其核心为针对不同的页面要应用不同的阈值，而合理阈值的确定是一个相对困难的过程，阈值的大小将会对抽取的结果产生直接的影响，如果阈值设置不当，信息抽取将会产生很大的偏差[2]。

2.基于视觉特征

基于视觉特征的信息抽取技术依赖于网页的组织结构，该技术主要适用于那些结构清晰、没有过多标签错误并且符合一般设计习惯的网页。如何对网页进行合理并且准确的分块是基于视觉特征的信息抽取技术的核心内容，该方法考虑到了网页上可以

收稿日期：2015-05-11

作者简介：沈娜（1984-），女，江苏宿迁人，工程硕士，宿迁开放大学讲师，研究方向：数据库技术、网络安全与应用技术。

直观看到的分块结构，比单纯只考虑网页标签结构的信息抽取方法更为科学[3]。然而，该技术受网页自身的错误影响较大，对于被浏览器容错的标签错误较为敏感，而且需要人工对页面区域进行分块，网页设计中的结构不规范、视觉分块的标准不统一的问题都会对该抽取方法产生巨大的影响，导致该抽取方法的实现过程较其他信息抽取的方法较为繁琐。

3.基于 DOM 树

HTML 标签的可嵌套性使得从网页标签抽象得到的 DOM 模型通常以树状结构呈现，从而克服了网页数据源的限制[4]。由于基于 DOM 树结构的信息抽取技术是根据网页自身结构的特点，通过对网页 DOM 树的对比操作来确定页面的主题信息的，因而受主观因素的影响相对比较小，相比基于视觉特征的信息抽取方式更易于实现，并且在实现更为高效的情况下，操作也更为准确[5]。它不仅可以用来处理单正文的页面，对于多正文的页面信息抽取同样有效。但是该技术需要对每个页面都进行类似的处理，没有充分利用已经抽取得到的结果对类似的页面进行统一的处理方法，在这个方面效率较低。

4.基于模板

对于经由统一的模板，通过动态查询数据库来产生页面的网页而言，其特点是网页相似度较大，能够找到统一的模板，基于模板的信息抽取技术正是利用该类网页的这种特点提出来的，其优点在于该方法通过对相似页面的分析总结出一套统一的抽取模板，可以大大提高抽取效率，然而对于不同种类的页面需要设置不同的阈值，因此模板的生产具有相当的难度。而且该技术只能针对单正文的网页内容，不适用于多正文的网页，模板的通用性问题还有待提升。

（二）Web 信息抽取技术面临的挑战

1.已有的基于 HTML 网页结构模式的信息抽取存在的问题

共存问题。Web 信息抽取性能的精确性、自动化程度、通用性均有要求，如何解决其共存的问题成为 Web 信息抽取首先要解决的问题[6]。

自动抽取。Web 网页具有海量异构的特点，对于手工配置的信息抽取提出了挑战[7]，如何实现 Web 信息的自动抽取是 Web 信息抽取技术的一大难点。

适应范围。硬性的抽取算法不能适应 HTML 文档的变化，对于不同的 Web 页面进行定制需要更多的工作量，如何提高抽取算法的适用范围成为 Web 信息亟待解决的问题之一[7]。

准确性。Web 网页中有的网页结构复杂，对于复杂页面信息抽取的准确性要求也是 Web 信息抽取中的另一个重要的环节。

2.解决方法

现在的新闻网页在一般情况下都具有相似的页面布局结构，这就为同一类的网页定制抽取算法成为可能。并且，经过实例分析可以发现，网页的内容同其对应的网页 DOM 树有着密切的联系，通过分析 DOM 树的属性可以分析并获取网页的内容，从而可以根据定制的算法抽取出网页中有价值的信息。

在研究静态网页结构特性，分析网页 DOM 树结构的基础上，设计了基于文本标签属性的信息抽取模型。基于文本标签属性的关键就是要根据给定网页的特点，找到所要抽取信息所在标签的位置。一般情况下，内容不同的标签的属性也不相同，这就是基于标签属性信息抽取的出发点。通过分析标签的属性之后，针对各网页定制相应的抽取算法。

三、基于文本标签属性的网页信息抽取

通过分析研究各大新闻网站的源码，我们发现非新闻内容通常存在于新闻正文内容区域周围或是新闻内容区域中间，由于新闻网页的布局结构与风格基本相似，针对某一具体新闻网站的研究发现，新闻正文的 HTML 标签非常相似，非新闻内容也具有相似的 HTML 标签，因此，本文设计了一种基于文本标签属性的信息抽取模型，由用户或是应用程序提交一个初始的 URL 作为输入，最终将过滤后的 HTML 网页内容返回回来。该模型是建立在一个简单的网络爬虫程序的基础上的，首先通过爬虫获取新闻页面的内容，然后根据网页文本标签的属性对获取的新闻内容进行处理、过滤，从而取得最终的抽取结果。

网络爬虫在实现中可借助于一些网页分析工具对网页进行遍历和获取页面中的文本内容，本文使用的是 HtmlParser[8]分析工具。HtmlParser 通过对目标网页建立其逻辑结构，然后采用 HtmlParser 过滤器定位指定的 HTML 节点的方式实现对网页的信息提取。

以获取网页新闻内容为例，主要思路是：首先获取新闻网页中所有的新闻标题和新闻链接；然后根据新闻链接去获取新闻正文内容。具体的操作过程分成三步：

第一步，输入一个 URL 作为一个 Parser，使用这个 Parser 作为一个 Visitor；

第二步，进行节点的遍历，并获取 Visitor 遍历后得到的数据。这个过程主要通过

Parser.visitAllNodeWith（Visitor）语句实现，Visitor 通过 visitor.beginParsing（）做解析之前的事情，每取到一个节点 Node，都会让该节点接受该 Visitor；

第三步，Visitor 通过 visitor.finishedParsing（）做解析后的事情。

抽取页面新闻链接的关键代码如下：

```
Parser parser=new Parser(url);
```

```
parser.setEncoding("gb2312");
TagNameFilter filter=new TagNameFilter("A");
NodeList nodeList=parser.extractAllNodesThatMatch
(filter);
Node[] nodes=nodeList.toNodeArray();
for(int i=0;i<nodes.length;i++)
{
Node f_parent=nodes[i].getParent();
Node parent=f_parent.getParent();
if (parent.getText ().startsWith ("ul class=\"f14
l26 ic_l\"")
||parent.getText ().startsWith ("ul class=\" f14
l26 Q-iDotBlue")
||parent.getText().startsWith("ul class=\"f14 l26
Q-iDotBlue"))
{
LinkTag link=(LinkTag)nodes[i];
String linkUrl=link.getLink();
String text=link.getLinkText();
try {
rs=stmt.executeQuery("select * from title where
name='"+text+"'");
if(! rs.next()){
stmt.executeUpdate("insert into title(name,src)
values('"+text+"','"+linkUrl+"')");
}
} catch (SQLException e){
//TODO Auto-generated catch block
e.printStackTrace();
}
System.out.println(linkUrl+"******"+text);
}
}
```

由于提取新闻链接需要过滤掉非链接的部分，留下链接的内容，因此在做 Parser 时需要通过 TagNameFilter filter=new TagNameFilter("A") 过滤掉非链接的内容。由于页面中的链接并非都是新闻链接，可能会出现广告链接等一些噪音信息，需要进一步对获取的链接进行过滤。

（一）基于文本标签属性的信息抽取模型

基于文本标签属性的信息抽取模型如图 1 所示，其运行的步骤描述如下：

步骤 1：遍历给定的 HTML 网页，分析并得到 HTML 页面的所有节点；

步骤 2：将节点抽象成 HTML DOM 树，根据 Tag 进行节点类型的划分，例如经过划分可以得到 LinkTag，ImageTag，ParagraphTag，InputTag，FrameTag 等等；

步骤 3：分析给定网站，找出新闻正文与非新闻正文的 HTML 节点的属性，根据找到的属性定制网站，过滤掉与新闻正文无关的部分；

步骤 4：存储新闻正文文本。

本模型的关键在于通过对特定网页新闻正文的标签属性的分析，定制针对具体网页的正文抽取算法，由于是针对具体网页进行定制，该模型的执行正确率达到 100%，缺点是定制不具有普遍适用性，针对不同的网页都要进行分析进而获得准确的标签，对定制人员的要求比较高。

（二）基于文本标签属性的信息抽取算法

该算法的核心思想是：确定要抽取的 HTML 标签属性，针对具体的网页进行定制。实现过程中需要对网页文档解析树进行分析，根据抽象 HTML 文档解析树的节点属性，做进一步的抽取。具体的实现过程如下：

第一步：分析网页 HTML 源码

以腾讯新闻网页为例，通过分析网页正文，我们发现新闻正文的内容是在“ ”“ ”之间，根据“Cnt-Main-Article-QQ”div 标签的 id 属性就可以定位到腾讯新闻正文所在位置。但是腾讯新闻的正文中除了文本内容外，还可能存在图片或视频等

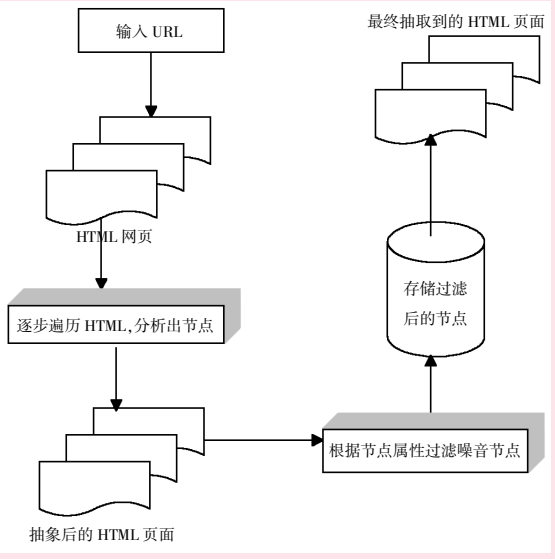

图 1 基于文本标签的信息抽取模型

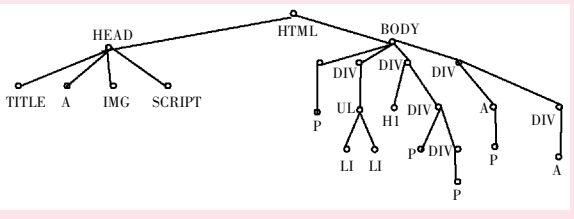

图 2 网页文档解析树

表 1 Web 新闻抽取系统功能列表

内容属性	1.新闻标题;2.新闻来源;3.新闻类型;4.新闻发布时间;5.新闻原文链接;6.新闻正文;7.新闻图片;8.新闻参与评论数
相关抽取函数	1. ExtxactTitle (Stringurl);2. ExtractSource (Stringurl);3. ExtxactTime (Strlng);4. ExtxactContent(Stringurl);5. ExtxactImage(Stringurl);6. ExtxactCmt(Stringurl)
应用类型选项	1.国际新闻;2.国内新闻;3.社会新闻;4.军事新闻;5.美图欣赏;6. RSS 阅读

内容,为了区分这些内容,可以再进一步分析 div 下面的标签,通过分析,纯文本内容是在标签“<P>”“</P>”之间,而图片内容是在<img>标签下显示。由于腾讯新闻的标题是在“<h1>”“</h1>”之间,通过过滤掉除<h1>之外的标签可以获得新闻的标题。

第二步:抽象 HTML 节点

根据网页的 HTML 标记,将网页表示成一棵树型结构,可以将任何一个 WEB 网页的 HTML 标签抽象成一棵网页文档解析树,每个文本标签都可以看成是树中的一个节点。如图 2 所示。

本模型暂时忽略各个标签间的差异,将所有的 HTML 标签都抽象成节点 Node,当需要进行节点过滤的时候,再按照类型对节点进行分类,划分得到 LinkTag,ImageTag,ParagraphTag,InputTag,SelectTag 和 FrameTag 等类,进而得到具有特定属性特征的 Tag,最后对抽象得到的 Tag 属性进行过滤操作。

第三步:定制网页抽取算法

具体的实现方法描述如下:

步骤 1:用一个 URL 或是页面 String 做一个 Parser;

步骤 2:用这个 Parser 做一个 Visitor;

步骤 3:使用 Parser.visitAllNodeWith(Visitor)遍历节点,并获取 Visitor 遍历得到的数据;

步骤 4:根据标签类型的不同抽象节点类型;

步骤 5:根据标签属性的不同定制过滤参数。

例如:

```
public boolean accept(Node node) {
return ((node instanceof Tag)
&&! ((Tag)node).isEndTag()
&&((Tag)node).getTagName().equals("H1")
&&((Tag)node).getAttribute("id")! =null
    &&((Tag)node).getAttribute("id").equals("
artibodyTitle"));}
```

这段代码的功能是获得类型为 H1 的节点,过滤的条件有两点:一是标签名称为 H1;二是标签的 id 属性不为空且必须是“artibodyTitle”,使用这个过滤算法便可以将新浪新闻的正文标题抽取出来。

表 1 给出了抽取时从网页中抽取的元素和抽取元素所使用的相关函数信息。在实际应用中,由于各大新闻网站有所不同,可能会导致某些元素的抽取出现空的情况发生,这时,在抽取显示页面中将不予显示。

四、结语

本文提出的 WEB 新闻正文抽取模型主要针对网页中的广告等垃圾信息进行数据清洗,将网页中用于显示特殊效果或是广告内容的脚本去掉,而对于那些动态显示网页有用信息的脚本(如新闻评论数)则不会被过滤掉。其抽取工作主要分成两步:第一步,抽取新闻网页中所有的新闻标题及新闻链接,过滤掉不相关的链接,如广告链接等等;第二步,对于第一步所抽取到的新闻链接,抽取其中的新闻正文内容,将其它不相关的噪声信息过滤掉,从而抽取出网页中所有有价值的信息。系统分别选取了新浪网、腾讯网、凤凰网、人民网、搜狐网、环球网和网易新闻七大主流新闻网站进行了抽取实验,针对不同网站正文的不同特征,系统定制了相应的抽取算法,实验结果表明,正文抽取的正确率都能够保持在 95%以上。

参考文献:

[1] 林子熠,沈备军.基于统计的自动化 Web 新闻正文抽取[J].计算机应用与软件,2010,27(12):232-235.

[2] 周佳颖,朱珍民,高晓芳.基于统计与正文特征的中文网页抽取研究[J].中文信息学报,2009,23(5):80-85.

[3] 胡东东,孟小峰.一种基于树结构的 Web 数据自动抽取方法[J].计算机研究与发展,2004,41(10):1607-1613.

[4] 于琨,蔡智,糜仲春,蔡庆生.基于路径学习的信息自动抽取方法[J].小型微型计算机系统,2003,24(12):2147-2149.

[5] 朱永盛,武港山.基于 Web 的新闻信息抽取[J].计算机工程,2006,32(10):74-76.

[6] Li, J.Q., Zhao, Y.PathRank:Web page retrieval with navigation path,Proc[J].ECIR,2009,09:350-361.

[7] Gulhane, P., Rastogi, R., Sengamedu,S.H.,Tengli,A. Exploiting Content Redundancy for Web Information Extraction[J].Proc.VLDB Endow,2010,03:578-587.

[8] 章栋兵.互联网舆情分析关键技术的研究与实现[D].武汉:武汉理工大学,2010.

[责任编辑:胡大威]

(下转第 73 页)

（上接第 65 页）

Research on Information Extraction of Webpage Based on Text Tagging Attributes

SHEN Na
（Suqian Open University, Suqian223800, China）

Abstract: With the rapid development of Internet, online information resources present a blowout situation. At the same time, it has become increasingly difficult to extract information from huge amounts of the information we need. After studying the existing Web information extraction technology and the challenges faced, we design a Web news information extraction model based on text tagging attribute. This paper mainly introduces the Web information extraction technology based on the attribute of text tag, presents the specific implementation process of information extraction, describes the traversal algorithm of the filtering text labels based on DOM tree node, and chooses the mainstream news sites to carry out the extraction experiment and to verify the feasibility of the algorithm.

Key words: HTML DOM Tree; text tagging attributes; Web news; information extraction