文章编号:1671-9352(2012)03-0038-05 DOI:10.6040/j.issn.1671-9352.2012.03.008

基于DOM树及行文本统计去噪的网页文本抽取技术

李霞,蒋盛益
(广东外语外贸大学思科信息学院,广东 广州 510006)

摘要: 首先对网页源码文本统一编码转为UTF格式,然后把HTML网页文档转换为XML文档并解析为一棵DOM树。依据XML语言特点及噪声特征规则先对DOM树的噪声节点进行过滤删除,然后依据中文标点符号统计方法提取网页正文内容,并在此基础上利用行文本统计方法去除提取出的正文中存在的噪声信息,最后得到网页正文文本。对来自结构完全不同的主流与非主流的中英文新闻网站上的2 000篇网页进行实验,结果表明本文提出的方法具有较高的抽取准确率,并具有很好的通用性和实现简单的特点,适用于针对互联网中不同网站新闻文本信息的自动采集。

关键词: 网页文本抽取; DOM树; 行文本统计; 标点符号统计
中图分类号: TP391 **文献标志码:** A

Content extraction from web page based on the DOM tree and line-text statistical noise-elimination

LI Xia, JIANG Sheng-yi
(Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006, Guangdong, China)

Abstract: As different web pages have different codes, the HTML web page first need to be encoded with the uniform code UTF8, and then translated into an XML document which is parsed into the DOM tree. After removing some noise nodes from the DOM tree according to the features of XML language and the rules of the noise characteristics, text contents are extracted from the DOM tree by the method of statistics of punctuation and noise information is continued to be eliminated from contents extracted above by the method of statistics of line-text. The result of experiments on 2000 web pages obtained from different web sites shows that our method has high accuracy, great generality, and simplicity, and can be automatically used to extract the right contents from different web sites.

Key words: content extraction from web pages; DOM tree; statistical of line-text; statistical of punctuation

0 引言

随着网络技术的发展,尤其是Web 2.0时代的到来,Web已成为一个巨大的知识宝库,如何挖掘和理解Web中的海量数据,如何在Web海洋中快速找到人们所需的信息已成为当前研究的一个热点。目前,互联网中大部分信息都存储在半结构化的网页里,而网页通常含有一些与正文信息无关的广告信息、导航链接信息以及版权信息等,这些信息通常也被称为噪声信息。噪声信息的存在是导致网页正文提取准确率不高的一个重要原因,因此识别

收稿日期:2011-11-30; 网络出版时间:2012-03-20 10:58
网络出版地址:http://www.cnki.net/kcms/detail/37.1389.N.20120320.1058.016.html
基金项目:国家自然科学基金资助项目(61070061);教育部人文社会科学研究青年基金资助项目(11YJCZH086);广州社科青年基金资助项目(11Q20)
作者简介:李霞(1976-),女,副教授,硕士,主要研究方向为数据挖掘和自然语言处理. Email:shelly_lx@126.com

和清除网页中存在的噪声文本是提高网页文本采集效果的一个关键技术。由于不同网站的模板和结构有所不同，实现一种有效、通用及实现简单的网页正文信息自动抽取技术显得尤为重要。

目前国内外学者对 Web 信息的抽取工作已经做了大量的研究，所提出的方法中主要包括基于模板的方法[1]、基于学习的方法[2-5]、基于网页内容分块的方法[6-7]和基于统计的方法[8-10]等。文献[1]通过自动抽取同类网页的 Wrapper 来对同类网页进行自动抽取。文献[2]通过将网页文本按照其显示属性的不同进行分组，以显示属性值为基础对 Web 页面文本进行分类。文献[3]将相似度高的网页归为一类，依据每类网页训练得到的网页模板对未知网页的内容进行提取，该方法适用于网页结构相似度较高的网络文献的提取，而无法适用于结构不同的 Web 网页的信息提取。文献[4-5]提出利用某些学习方法来识别广告冗余及不相关信息，但是这些技术都不是自动化的操作，需要大量手工标记训练数据和领域知识。文献[6]通过计算页面集中每个属性的信息熵，依据熵值将一个网页划分为内容块和噪声块，该方法虽然可以发现和区分出同一个站点中页面的内容信息和噪声信息，但对不同的模板的不同网站需要设置不同的最优阈值。文献[7]将 HTML 网页分为不同的文本内容块，然后通过分析块重要度和块特征来辨别出含有正文的内容块，在识别噪声信息如导航栏、网站目录等信息时采用通过敏感词过滤及广告地址过滤等方法，具有一定的局限性。文献[8-9]所提出的方法实现简单、通用性好，但存在以下缺点：(1) 只考虑了 table 节点；(2) 当某些网页所包含的正文信息很短时，因为信息量较少，所包含的正文节点往往会被过滤掉。文献[10]通过统计中文句号确定部分正文信息，然后根据正文信息在结构上的相似性确定其他正文信息的内容。该方法在某些正文信息较少的网页中，可能会将噪声判定为代表结构的正文信息，从而无法提取出真正的正文。

本文在前人工作的基础上，结合对 HTML 网页性质的观察和分析，实现了一种基于 DOM 树及行文本统计去噪的网页文本抽取方法，该方法克服了文献[8-10]中所提出方法的不足，不仅可以提取正文文本较多的网页正文，同时还能够较好地提取正文内容较少的网页正文。为了验证本文提出方法的有效性，文章对来源不同的 10 个中英文网站的 2 000个新闻网页进行了抽取实验，实验结果表明本文提出的方法确实提升了网页信息提取的准确度，并能适用于不同结构网站的网页信息的全自动提取。

1　网页标准化及 DOM 树生成

目前互联网中大多数网页仍然使用 HTML 格式，HTML 语言是一种标识语言(Markup Language)，它定义了一套标签来刻画网页显示时的页面布局。在 HTML 格式的网页中，存在标签不匹配、嵌套混乱及标签格式不规范等情况，如有〈title〉标签，而没有对应的〈/title〉标签，这种不规范有时不会对网页的正常显示有影响，但不便于正文信息的抽取，为此首先应对 HTML 代码进行预处理，将其标准化。本文采用 W3C 组织推荐的工具集 HTML Tidy[11] 来将书写不规范的 HTML 文档转换成格式良好的 XHTML 文档(XML 的子集)。对 HTML 网页规范化的过程主要包括如下步骤：

(1) 统一网页的编码形式，将编码为 GBK、GB2312、UTF-8 等不同格式的网页统一转换成 UTF-8 字符集编码格式；

(2) 使用开源工具 HTML Tidy 转化网页为标准化的 XHTML 文档；

(3) 用正则表达式替换可能引起错误和干扰的字符，如将 HTML 源码中的“ ”以空串替换，将 HTML 标签格式〈strong〉〈/strong〉、〈font〉〈/font〉、〈p〉〈/p〉以空串替换等；

(4) 通过在网页的头部添加标准的 XML 声明〈? xml version = “1.0” encoding = “UTF-8” ?〉，将标准化后的 HTML 文档转化为 XML 文件；

(5) 用 C ++ 的 Tinyxml 库把标准化后的 XML 网页源码解析成一棵 DOM 树；

(6) 将 DOM(document object model) 树中与正文提取无关的节点信息删除，这些节点包括 style、script、img、〈! —〉、iframe、object、meta、applet、link、doc 等。

其中 DOM 是一种以面向对象方式描述的文档模型，它定义了表示和修改文档所需的对象，这些对象的行为和属性以及这些对象之间的关系。DOM 树将整个页面映射为一个由层次节点组成的树形结构，其典型结构如图 1 所示。

2　网页正文内容提取

2.1　基于标点符号统计提取正文内容

通过对网页源码的分析发现，网页中包含文字

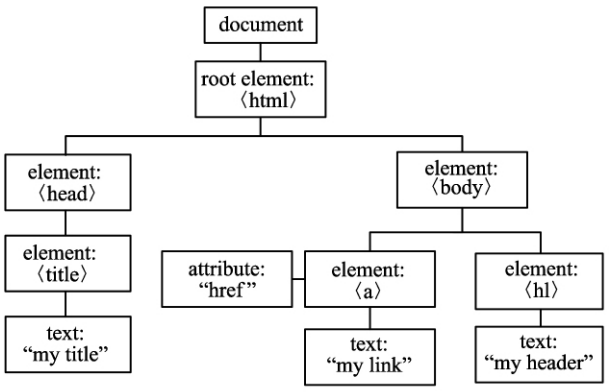

图1 DOM 树结构

Fig. 1 The structure of DOM tree

的节点通常分为两类: 一类是包含有标点符号的文字节点,这类节点大多数是正文节点和某些版权信息节点; 另一类是不包含标点符号的文字节点,这类节点通常是导航或广告链接节点。文献[10]统计发现约有96%的中文句号出现在网页正文中,是所有中文标点符号中分布最多的。该文将句号作为网页正文区别于其他部分的特征。考虑到中文句号对英文网页的不适用性,以及在某些网页的正文中只有感叹号而无句号等其他情况,本文将中英文句号、逗号、感叹号、中括号等标点符号作为区分网页正文与其他部分的特征。

设文本节点 element $=\{c_1,c_2,\cdots,c_n\}$, $c_i(i=1,\cdots,n)$ 为组成该文本节点的字符,这些字符包括中文汉字、英文字母和中英文标点。为了体现标点符号特征明显的文本为正文内容这一思想,通常是定义满足条件 $\frac{\text{文本节点 element 的标点符号个数}}{\text{文本节点 element 的所有字符总数}}>p$ 的文本节点内容为网页正文内容,这样处理将使得阈值的设置相当困难。由于不同网页的正文内容差别很大,有些正文的标点符号多,有些正文的标点符号少,有些正文虽然很长,但是一共就几个句子,所以标点符号的比值非常小,此时如果阈值 p 取得太大,则无法将正文取出。反之为了能够将长短不同的网页的正文都能取出,将阈值 p 设定得太小,就会将更多的噪声选择进来。考虑到网页正文部分文字较多而标点符号相对较少的特点,本文的处理方法是将标点符号比值落在区间 $[0,p]$ 之间的节点文本取出作为网页正文内容,这样做可以确保所有网页的正文都能够被提取出来,不会被噪声所淹没。

本文随机提取了10个不同网站的2 000篇网页,设定 p 值在0.01到0.9之间(间隔刻度为0.01)来提取正文,结果显示当 p 值设定在0.3以下时能够对所有网页全部提取正文。同时由于网页中某些版权等信息也有可能是比较长的文本信息,并且也包含一定的标点符号,如腾讯网中的噪声文本"如果你对新闻频道有任何意见或建议,请到交流平台反馈。"和新浪网中的噪声文本"¦ Copyright? 1996-2011 SINA Corporation, All Rights Reserved 新浪公司"中均含有一定比例的标点符号,这些噪声数据也会被同时提取出来作为正文内容。通过对网页的详细分析发现,相同网站的网页的噪声信息是基本相同的,如具有相同的广告信息,具有相同的版权信息等,同时这些信息一般均会出现在不同行上。基于这个特点,我们提出在初始构建 DOM 树并提取含有部分噪声的正文内容的基础上,使用 MD5 编码技术统计行文本信息,将视为噪声的行文本从已提取的正文中删除。

2.2 行文本统计去噪

信息摘要 MD(message digest)是根据公开的 MD5 算法对原信息进行数学变换后得到的一个128位的特征码,依据特征码的惟一性和不可逆性,MD5 编码值可以惟一地代表原信息的特征。行文本统计去噪的过程是,首先将节2.1提取出的正文内容依据换行标志'\n'划分为行文本集,然后计算每个行文本的 MD5 编码,并维持一个行编码表,对正文的每一行进行统计并检索 MD5 编码表,当行文本频繁出现时,判定该行文本为噪声信息,将其从正文部分删除。行文本统计去噪算法的详细描述如下:

(1) 读入一篇含有噪声的网页正文文本 Text;

(2) 按照换行标志'/n'划分网页正文文本 Text 得到行文本集 $D=\{t_1,t_2,\cdots,t_n\}$,其中 $t_i(i=1,2,\cdots,n)$ 为行文本;

(3) for $i=1$ to n do

(4) 计算 t_i 的 MD5 编码并查找 MD5 编码表;

(5) 如果编码表为空,则在编码表中插入一条新记录,该记录的 LC 值设为1;

(6) 如果编码表不空且找到 t_i 的相同编码值,则将编码表中该记录的 LC 值加1;

(7) 判断 t_i 的 LC 值,如果满足 $\text{LC}\leqslant \text{int}\left(1+\frac{\text{RC}}{L}\right)$,则保留 t_i,否则认为 t_i 是噪声,将 t_i 从当前网页正文 Text 中删除;

(8) end for

(9) 成功处理网页的数量 RC 加1,转入(1)继续下一篇网页正文的处理。

算法中的参数 LC 代表某文本行总共出现的次数,RC 代表当前已经成功处理多少篇文章,L 表示

每处理多少篇文章允许行文本噪声增加一行的范围。通常随着处理新闻网页数量的增加，部分行文本重复的概率就会增加，因此需要设定一定的范围来判定该行文本是否是噪声。本文设置 L 为 50，即处理文章个数在 50 篇以内，允许判定为噪声的行文本重复次数为 2，当处理文章个数为 100 时，允许判定为噪声的行文本重复次数为 3，依此类推。

算法中随着处理文章的增多，行文本记录在数据库中会急剧增加，这会导致查询数据库的效率降低。事实上，某些属于正文的行文本的重复出现次数会很低，这些行文本应该需要从行文本数据库中删除。本文的处理方法是当行文本的重复次数小于成功处理总得文章个数的 1% 时，即 $LC \leqslant \frac{RC}{100}$ 时，将该类行文本记录从行文本数据库中删除。这样随着采集到的新闻数量的增多，行文本数据库的大小基本维持在一个恒定的大小。

为了验证本文算法思想的正确性，作者对 10 个网站均分别自动抽取 1 万个网页的正文，实验结果显示同类网站的广告链接信息、版权信息等出现的次数比较频繁。表 1 展示了部分网站中出现的频繁度行文本信息。依据噪声数据高频繁度出现的特点可以去除很多文本信息中的干扰因素，获取准确率较高的正文内容。

表 1　8 个网站频繁行文本信息

Table 1　The highest frequency line-text information from 8 web sides

网站	频繁行文本信息
news. qq. com	如果你对新闻频道有任何意见或建议，请到交流平台反馈。
news. sina. com. cn	┆ Copyright? 1996-2011 SINA Corporation, All Rights Reserved 新浪公司
news. 163. com	网友评论仅供其表达个人看法，并不表明网易同意其观点或证实其描述。
news. cn. yahoo. com	Copyright? 2011 Yahoo. com. cn 版权所有 不得转载
news. china. com	所有评论仅代表网友意见。
news. csdn. net	无
www. cnblogs. com	找优秀程序员，就在博客园
www. chinadaily. com. cn	Today´s top news Editor´s Picks

3　实验结果

本文对来自主流和非主流及结构完全不同的 9 个中文网站和一个英文网站中一共抽取了 2 000 个网页作为测试数据，采用基于标点符号统计及结合行文本统计去噪的算法对这 2 000 个网页的正文进行提取，提取结果如表 2 所示。从表 2 统计结果可以看出，本文提出的方法在不同网站的新闻网页上抽取的结果最高为 100%，最低达到 93%，平均准确率达到 96. 8%。通过分析发现，对于某些广告噪声较多的网站如 http: //news. sina. com. cn，http: //new. qq. com 等，本文方法的优势更为明显。

表 2　本文方法的试验测试结果

Table 2　Experiment results of our method

测试网页数据来源	网页总数/个	正确抽取数/个	错误抽取数/个	准确率/%
http: //news. qq. com	200	196	4	98
http: //news. 163. com	200	192	8	96
http: //news. sohu. com	200	192	8	96
http: //news. china. com	200	189	11	94. 5
http: //news. sina. com. cn	200	199	1	99. 5
http: //news. cn. yahoo. com	200	197	3	98. 5
http: //news. csdn. net	200	200	0	100
http: //china. nba. com	200	187	13	93. 5
http: //news. cnblogs. com	200	186	14	93
http: //www. chinadaily. com. cn	200	198	2	99
合 计	2 000	1 936	64	96. 8

由于本文提取的新闻网页全部来自不同的网站，并且所提取的准确率差别不大，这证明本文提出的方法具有很强的通用性，能够适用于不同结构的网站新闻文本信息的提取。通过对抽取出错的网页进行分析发现，抽取错误的主要因素为以下几点：(1) 〈span〉、〈h〉等修饰标签过多的网站提取结果

会丢失部分正文,原因是所使用的 C ++ 的 TidyHtml 库在规范 html 源码的时候存在着一些不足,它在处理大块有着相同标签的源码的时候会误删一部分标签;(2) 由于基于行文本统计的去噪需要一个学习的过程,所以在初始的几个网页处理效果不明显,会包含一些错误的链接信息。图 2 为采集到的来自不同网站新闻网页的语料数据库的部分截图。

index	urladdr	source	title	newstime	content	updatetime
24001	http://cbachina.163.co	网易CBA官网	曝范斌曾错过执教江苏 队	2011-04-12 10:34:41	曾经想来南钢? 这个消息一	2011-04-29 22:50:31
24002	http://cbachina.163.co	网易CBA官网	江苏早已失去翻盘动力 新	2011-04-09 04:43:50	和新疆队的"绯闻"不断,而另	2011-04-29 22:50:31
24003	http://money.163.com	网易财经	第三轮融资尘埃落定 钻石	2011-03-18 08:05:34	行业开始发生裂变。	2011-04-29 22:50:31
24004	http://cbachina.163.co	网易CBA官网	杜比+辛格尔顿=场均59分	2011-04-08 23:58:04	半决赛的最后一战,新	2011-04-29 22:50:31
24005	http://fashion.163.com	网易女人	挑战你的尺度 纽约时装周	2011-02-15 14:54:48	皮革是非常富有表现力的材	2011-04-29 22:50:31
24006	http://cbachina.163.co	网易CBA官网	杜比成争冠最大筹码 总决	2011-04-08 23:45:48	其实横扫江苏进入总决	2011-04-29 22:50:32
24007	http://cbachina.163.co	网易CBA官网	老迈巴特尔撑新疆内线 死	2011-04-08 23:40:37	老迈巴特尔撑新疆内线 死敌	2011-04-29 22:50:32
24008	http://money.163.com	网易财经	钻石小鸟确认获得第三轮	2011-03-14 13:50:36	电子商务公司上海钻石小鸟	2011-04-29 22:50:32
24009	http://cbachina.163.co	网易CBA官网	江苏赛季总结:动荡后重回	2011-04-08 23:17:57	整个CBA赛季,江苏队-	2011-04-29 22:50:32
24010	http://cbachina.163.co	网易CBA官网	评分:杜比再度奉献超级数	2011-04-08 23:10:04	杜比再度轰下35分7助球	2011-04-29 22:50:32
24011	http://realestate.cn.ya	雅虎家居	实用收纳 推荐三套适合玄	2011-04-27 11:06:00	: 玄关是进出住宅的必经之	2011-04-29 22:50:33
24012	http://realestate.cn.ya	雅虎家居	为您的小户型扩大空间 5种	2011-04-22 11:24:00	小户型的家,收纳一直都是	2011-04-29 22:50:33
24013	http://cbachina.163.co	网易CBA官网	盘点加分条件:巨星、本	2011-04-11 22:38:10	盘点加分条件:巨星、本土	2011-04-29 22:50:33
24014	http://cbachina.163.co	网易CBA官网	盘点保分条件:优秀教练	2011-04-11 22:38:10	要想成为总冠军,国字	2011-04-29 22:50:33
24015	http://money.163.com	网易财经	于广洲同志任海关总署党	2011-04-09 23:37:00	于广洲同志任海关总署党组	2011-04-29 22:50:34
24016	http://fashion.163.com	网易女人	马克雅克布秀场星光熠熠	2011-02-16 14:47:54	Karen Elson,凯伦·艾尔森选	2011-04-29 22:50:34
24017	http://realestate.cn.ya	雅虎家居	家具业迎来春天 儿童家具	2011-04-13 10:47:00	中国家具协会副理事长陈宝	2011-04-29 22:50:34
24018	http://realestate.cn.ya	雅虎家居	春天打造浪漫闺房 10招教	2011-04-12 07:34:00	韩流来袭,那些唯美画面的	2011-04-29 22:50:34
24019	http://realestate.cn.ya	雅虎家居	外国时尚设计 春意盎然的	2011-03-19 08:36:00	春天,万物复苏,生机盎然	2011-04-29 22:50:35
24020	http://money.163.com	网易财经	交行:3月再现顺差因出口	2011-04-10 13:52:12	2011年3月,我国进出口总值	2011-04-29 22:50:35
24021	http://money.163.com	网易财经	中兴通讯3G手机首度突破	2010-05-22 09:09:28	本报讯5月18日,日本第	2011-04-29 22:50:35

图 2 采集到的来自不同网站新闻网页的语料数据库中部分数据截图

Fig. 2 Screenshots of part content database data extracted from different web sides

4 总结

本文在改进基于标点符号统计的网页文本信息抽取方法的基础上,引入了基于行文本统计去噪的方法,得到了效果较好的网页文本信息提取结果。从来自不同网站的中英文新闻网页的提取结果看,所提出的方法具有较高的准确率,并且实现方法简单,具有很强的通用性。该方法已经被用于作者研究的网络舆情分析原型系统的新闻网页语料的采集系统中,进一步的研究工作是对采集新闻相对应的评论做进一步的提取。

致谢 感谢颜杰龙和黄小鸿同学实现了本文部分算法,并在大量网站上进行了测试验证工作。

参考文献:

[1] 梅雪,程学旗,郭岩,等. 一种全自动生成网页信息抽取 Wrapper 的方法 [J]. 中文信息学报,2008,22(1):22-29.

[2] 汪建伟,杨冬青,高军,等. 一种基于分类算法的网页信息提取方法 [J]. 计算机科学,2008,35(3):91-93.

[3] 李文立,王乐超,宋春雷. 基于 HTML 树和模板的文献信息提取方法研究 [J]. 计算机应用研究,2010,27(12):4615-4617.

[4] DAVISION B D. Recognizing nepotistic links on the Web [C] // Proceedings of the AAAI-2000 Workshop on Artificial Intelligence for Web Search. Austin: AAAI Press, 2000: 23-28.

[5] JUSHMERICK N. Learning to remove Internet advertisements [C] // Proceedings of the 3th International Conference on Autonomous Agents. Washington: ACM Press, 1999: 1-7.

[6] LIN S H, HO J M. Discovering informative content blocks from web documents [C] // Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. NewYork: ACM Press, 2002: 588-593.

[7] 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(6):24-30.

[8] SUHIT G, GAIL K, DAVID N, et al. DOM-based content extraction of HTML documents [C] // Proceedings of the 12th International World Wide Web Conference. Budapest: ACM Press, 2003: 207-217.

[9] 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.

[10] 宋明秋,张瑞雪,吴新涛,等. 网页正文信息抽取新方法 [J]. 大连理工大学学报,2009,49(4):594-597.

[11] Dave Raggett. Clean up your web pages with HTML TIDY [EB/OL]. [2011-05-30]. http://www.w3.org/People/Raggett/tidy/.

(编辑:许力琴)