文章编号：1000-8608(2009)04-0594-04

网页正文信息抽取新方法

宋明秋*，张瑞雪，吴新涛，李文立

（大连理工大学 系统工程研究所，辽宁 大连 116024）

摘要： 基于包装器的信息抽取方法只能处理一种特定的信息源，而且对网页结构的依赖性强.基于此提出了一种将中文标点符号和HTML树结构作为识别网页正文内容重要特征的网页分析方法，通过统计中文标点符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内容.实验结果表明该方法能有效地剔除网页噪音并提取网页正文，具有较好的通用性和较高的准确性.

关键词： 包装器；HTML树；网页信息提取

中图分类号： TP391 **文献标志码：** A

0 引 言

随着互联网的飞速发展，Web上的网页数目正以指数级的爆炸性趋势增长.面对如此巨大的资源，在Web上检索及发现有价值的信息已成为一项重要的任务.基于Web的研究涉及信息检索、信息过滤、信息抽取、搜索引擎、网页分类等，它们研究处理的主要对象就是网页信息.在网页中除了表达主题的正文内容外，还有与主题内容无关的导航条、广告信息、版权信息以及相关链接等噪音内容.有效地清除网页噪音并抽取网页正文是提高基于Web的应用程序处理结果准确性的一项关键技术，已成为基于Web的信息系统预处理环节中一项必不可少的工作.

在较早的时候，一般都使用包装器来对网页进行正文抽取，其基本思想就是针对特定的网站来书写或者抽取相应的规则，因为同一类型数据源的网页结构都是类似的.该方法只能处理一种特定的数据源，而且对网页结构的依赖性强，这种方法最大的缺点就是规则抽取的工作量大，规则维护的代价高，无法适应网页结构的变化，可扩展性差[1,2].

此外，有大量研究是针对网页内容本身进行分析的.文献[3、4]提出了一种基于网页结构化信息的正文抽取方法，该方法先将网页表示成一棵树，然后通过遍历这棵树的〈table〉结点来获取网页正文，不过如何去衡量正文并没有一个准确的方法，并且阈值也很难确定.文献[5]提出了一种根据网页的视觉化特征来提取网页正文的方法，主要利用字体的大小、布局信息、背景颜色等一些视觉信息，根据一定的规则将页面划分成视觉块.这种方法很好地模拟了人们观察网页的习惯，当人去观察一个网页时，因为正文比较突出、醒目，并且通常在网页的正中间，所以可以很轻松地找到.不过，由于视觉特征的复杂性，很难找到一个通用的规则集.文献[6、7]提出了一种将网页中字符个数与超链接个数的比值作为权值衡量正文内容的方法.文献[8]提出了一种基于统计的网页正文抽取方法，将中文文字个数作为衡量正文内容的标准，它首先假设中文字符在网页正文中出现的次数要比在其他部分出现的次数多得多.可是在实际过程中，这种方法错误率太高，无法作为一种通用的方法.

本文在研究已有网页信息提取方法的基础上，针对中文网页布局的特点，提出一种新的网页分析方法.该方法先将网页内容结构化表示，即将HTML文件规范化以构造HTML树，并提取结构树中的文字内容及其链路结构；然后根据中文句号的出现频率来确定一部分正文内容；最后根据

收稿日期： 2007-08-03； **修回日期：** 2009-05-11.
基金项目： 国家自然科学基金资助项目(70671016).
作者简介： 宋明秋*(1967-)，女，博士，副教授.

正文内容链路结构的相似性获取其余正文内容.

1　基于中文标点符号和 HTML 树结构的网页正文信息抽取方法

HTML(hyper text markup language)是超文本标记语言，是基于标准通用标记语言(SGML)的一个庞大的文档处理系统. SGML 的基本思想是采用描述标记(Tag)来提供描述文档结构的附加信息. HTML 利用 SGML 定义了一些标记，如〈html〉、〈title〉等，用于描述文本的显示方式，并对这些标记的使用都做了格式定义，对于实体符号的显示和标记元素的结构也做了规范，使得 HTML 网页在文本格式和结构上存在一定的规律，也为网页信息的提取提供了方便.

1.1　中文标点符号在网页中的分布特征

网页可以分成两类：一类是导航型网页，该网页主要是超链接导航信息，如各种门户网站；另一类是正文型网页，是指包含有主题内容的网页. 本文只对正文型网页进行处理，因为导航型网页含有大量的超链接，很容易被处理和识别. 本文将正文网页划分为 5 个部分：网页导航信息、网页正文、内容相关链接、内容不相关链接以及版权信息. 选取 10 个不同的门户网站，如新浪、网易、搜狐等，每个网站随机选取 10 个正文型网页，以统计中文标点符号在网页正文中出现的次数 N_1 和在网页页面中出现的次数 N_2.

由表 1 可见，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高的. 究其原因，主要在于网页正文部分大多由一个个句子组成，所以句号出现比较多；导航信息大多是两字短语；链接部分一般都取自所链接文章的标题，标题中一般不会出现句号；版权部分也基本都没有成行的句子，所以句号较少.

表 1　网页正文和网页页面中标点符号的分布

Tab. 1　Statistics of Chinese punctuation in main content blocks and in whole web page

中文标点符号	N_1	N_2	N_1/N_2
句号	2 285	2 389	96%
逗号	4 798	5 225	92%
问号	192	698	28%
叹号	97	295	33%

由此可见，使用中文标点符号，尤其是句号，可以作为网页正文区别于其他部分的特征.

1.2　网页内容结构化表示

HTML 文件是自描述的半结构化数据，数据的结构和内容混在一起，没有明显的区分；它们具有一定的结构性，但这些结构化的信息并没有提供足够的语义信息. 由于半结构化的数据很难被应用程序直接使用，为了从 HTML 文件中提取信息，必须先将其结构化.

一般情况下，HTML 元素相互嵌套，因此最适合用树型结构存放. 但由于 HTML 元素并不完全递归嵌套，允许有交叉的情况，而且有些元素可以没有结束标记，在将 HTML 文件组织成树型结构之前，需要先规整化，使其元素完全递归嵌套[9]. 规范化的要求如下：

(1)“〈”和“〉”只能用来包含网页标记，当在其他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〈table height=“400”〉.

(4)所有的标记必须是正确嵌套的. 如〈A〉…〈B〉…〈/A〉…〈/B〉是不正确的嵌套，正确的嵌套形式应该是〈A〉…〈B〉…〈/B〉…〈/A〉.

(5)由于文字内容有可能被修饰标记如〈a〉〈strong〉等标记隔断，为保持数据内容与修饰标记的顺序性，需要增加自定义标记〈text〉来嵌套文字内容.

在规范化之前，可以先删除〈form〉、〈script〉和〈style〉等用于控制 HTML 文件的交互性和显示的标记，这些标记不包含主题内容，剔除后可加快处理速度. 使用 HTML Tidy 工具对网页进行规范化可以实现完整的 HTML 树结构的显示，但没有针对性和可操作性，所以编写了用于本文研究的 HTML 规范化程序，然后通过 HTML 树解析程序将规范的 HTML 文件解析成 HTML 树，树中的每个结点包含了一对标记间的所有字符，结点的名字为对应标记的名字，如图 1 所示.

1.3　正文内容的结构特征

一个网页的正文内容具有很好的连续性，结构都非常相似，都具有同一个祖父结点和相似的父结点. 所以如果能根据字符特征确定一部分正文信息，找到其对应的树结点，则可以在该类结点附近寻找到结构与之相似但字符特征不明显的其他正文信息. 本文是通过提取从根结点到正文结

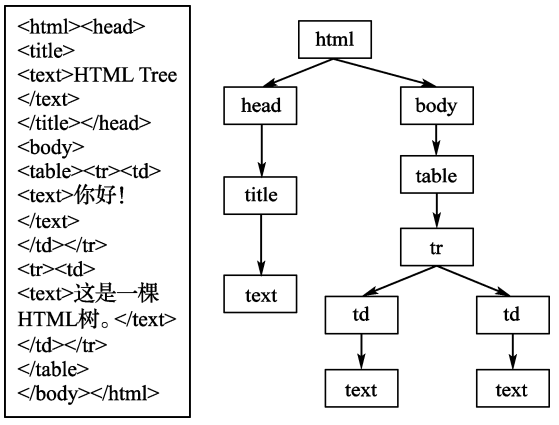

图 1 HTML 文件及其树形式

Fig. 1 HTML document and HTML tree representation

点的链路结构并比较链路结构的相似性来分辨正文内容和噪音信息的. 以图 1 为例,网页中有 3 块〈text〉文本信息,它们对应的链路结构分别是:

"HTML Tree": html→head→title→text;

"你好!": html→body→table→tr→td→text;

"这是一棵 HTML 树。": html→body→table→tr→td→text

对比 3 条链路很容易发现"你好!"和"这是一棵 HTML 树。"具有相同的链路结构. 如果可以通过某种特征确定"这是一棵 HTML 树。"是正文信息,则可以认为"你好!"也是正文信息,"HTML Tree"属于噪音信息.

然而在一般的网页结构中,正文内容的不同部分的链路并不是完全相同的,比如正文中的链接内容,或被强调的标题内容都比一般的正文信息多一层修饰标记. 为避免将这些被修饰的正文内容视为噪音信息,需要建立一个指标 $K_{i,j}$ 来衡量链路 i 与链路 j 的相似程度. 首先从根结点到叶结点依次比较链路结构 i、j 各个结点信息,记结点相同的个数为 $N_{i,j}$,记链路 i、j 的长度分别为 L_i、L_j,有 $K_{i,j}=\dfrac{2N_{i,j}}{L_i+L_j}$. 该指标具有遗传性,既与链路的祖亲程度有关,又与两叶结点的深度有关. 显然当它们的根结点不相同时 $K_{i,j}=0$;当两链路各结点都匹配,而且长度相等时,$K_{i,j}=1$.

1.4 网页正文内容的提取步骤

(1)整个正文内容的提取过程都是在网页 HTML 树中进行的,所以首先要对整个网页数据进行处理,提取出 HTML 树结构.

(2)遍历 HTML 树,提取所有的〈text〉结点,同时记录从根结点到每个〈text〉叶结点的链路. 在这里,使用递归算法遍历 HTML 树,递归函数如图 2 所示.

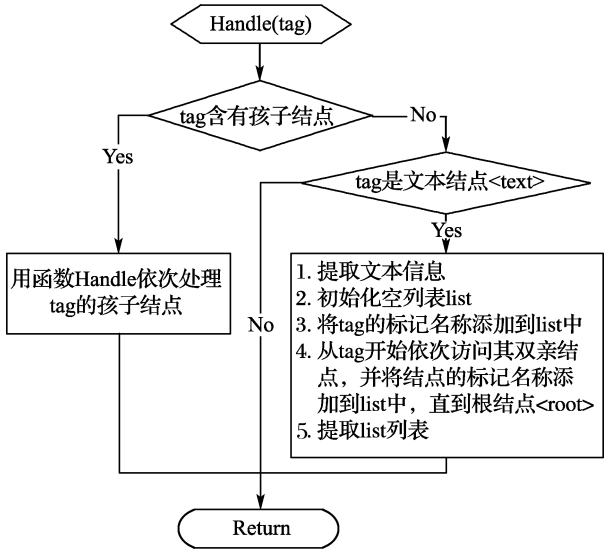

图 2 提取文本信息及其链路的递归算法

Fig. 2 Recursive algorithm for extracting text and its link path

(3)检查每段文字内容,记录各自的中文句号数,并选择句号数最大的文字内容对应的链路为结构样本.

(4)选取其他中文句号较多的文字内容,将其与结构样本进行链路结构的比较,以确定结构样本的正确性.

(5)选择阈值,将每段文字对应的链路与结构样本作比较,选取相似度大于阈值的链路结构作为正文内容. 相似度阈值的选择主要是考虑正文内容中经常出现的链接和被修饰的文本块. 考虑到最坏的情况,就是一部分文本块同时被链接和修饰,则该结点对应的链路结构与样本链路结构匹配的结点数目减少 1,链路长度和增加 2,从而相似度为

$$K_0=\frac{2N_{i,j}}{L_0+L_0+2}=\frac{2(L_0-1)}{2L_0+2}=\frac{L_0-1}{L_0+1}$$

其中 L_0 为样本正文的链路长度.

2 实验结果与分析

为了验证该方法的实际效果,本文选取了 10 个具有代表性的综合门户网站,包括新浪、搜狐、雅虎等网站,用此方法提取网页的正文内容,程序均由 Python 语言实现. 通过与人工提取出的正文内容做比较分析,得出实验结果如表 2 所示.

表 2　实验结果

Tab. 2　Experiment results

数据来源	网页总数/个	正确数目/个	错误数目/个	准确率/%
www. sina. com. cn	100	98	2	98
www. sohu. com	100	97	3	97
www. 163. com	100	95	5	95
www. tom. com	50	47	3	94
news. china. com	50	46	4	92
www. yahoo. com. cn	50	42	8	84
www. comxc. com	40	38	2	95
www. xfocus. net	40	34	6	85
www. 51job. com	40	37	3	92.5
www. chinadaily. com. cn	40	39	1	97.5

由表 2 可知，用基于中文标点符号的网页正文信息抽取方法进行网页正文提取的准确率最高为 98%，最低为 84%，平均为 94%，适用于绝大多数网站，保持了较高的准确性.

3　结　语

网页正文信息提取是网络信息资源的预处理过程，它的准确性关系着基于 Web 的应用程序文本处理结果的精确性. 实验结果表明，基于中文标点符号 HTML 树结构的网页正文信息抽取方法通用性强、准确率高，可作为搜索引擎、网页分类、信息抽取、信息过滤、信息重构与主题信息采集等的预处理，能大大提高网页语义内聚性和基于 Web 的应用程序文本处理的效率.

参考文献：

[1] GRESCENZI V, MECCA G, MERIALDO P. Roadrunner: towards automatic data extraction from large web site [C] // **Proceedings of the 27th International Conference on Very Large Database Systems**. Roma: Morgan Kaufmann, 2001

[2] YI Lan, LIU Bing. Web page cleaning for web mining through feature weighting [C] // **Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)**. USA: Lawrence Erlbaum Associates Inc., 2003

[3] LIN Shian-hua, HO Jan-ming. Discovering informative content blocks from web documents [C] // **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD'02)**. New York: ACM, 2002

[4] 常育红，姜　哲，朱小燕. 基于标记树表示方法的页面结构分析[J]. 计算机工程与应用，2004，**40**(16)：129-132

[5] CAI Deng, YU Shi-peng, WEN Ji-rong, *et al*. VIPS: a vision - based page segmentation algorithm [R] // Microsoft Technical Report MSR-TR-2003-79. Redmond: Microsoft Corporation, 2003

[6] GUPTA S, KAISER G E, NEISTADT D, *et al*. DOM based content extraction of HTML documents[C] // **Proceeding of the 12th International World Wide Web Conference (www2003)**. New York: ACM, 2003

[7] 宋睿华，马少平. 一种提高中文搜索引擎检索质量的 HTML 解析方法[J]. 中文信息学报，2003，**17**(4)：19-26

[8] 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

[9] 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)：46-67

A new approach to content extraction from web page

SONG Ming-qiu*, ZHANG Rui-xue, WU Xin-tao, LI Wen-li

(Institute of Systems Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: The approach to data extraction based on wrapper is limited to one specific information source, and greatly depends on web page structure. A new web page analysis method is proposed, which can recognize web page content according to the number of Chinese punctuations and HTML tree structure. It can eliminate noise and extract content from web page effectively. Parts of contents are confirmed by Chinese punctuations, while other parts are found by the similarity among contents. Experimental results show that this method is accurate and suitable for most web sites.

Key words: wrapper; HTML tree; web information extraction