基于标记窗的网页正文信息提取方法*

赵欣欣[1]，索红光[2]，刘玉树[2]
(1. 中国兵器工业 计算机应用技术研究所，北京 100089; 2. 北京理工大学 计算机科学技术学院 计算机科学工程系，北京 100081)

摘 要: 提出了基于标记窗的网页正文信息提取方法。该方法不仅适合于处理一个网页中所有正文信息均放在一个 td 中的情况，也适合于处理网页正文放在多个 td 中的情况，还可以处理网页正文文字短到与网页其余部分文字(如广告、导航条、版权)长度相当的情况。尤其重要的是，它能够解决非 Table 结构的网页正文提取问题。实验表明，该方法可以提高网页正文提取的准确率，适用性强。

关键词: 标记窗; 提取; 文档对象模型
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2007)03-0144-02

Web Content Information Extraction Method Based on Tag Window

ZHAO Xin-xin[1], SUO Hong-guang[2], LIU Yu-shu[2]
(1. Institute of Computer Application, China North Industries Group Corporation, Beijing 100089; 2. Dept. of Computer Science & Engineering, School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: This paper proposed a Web content information extraction method based on tag window, which could deal with some special circumstances, all the Web content information was put into one td or several tds, and the character numbers of Web content information was at most equal to that of the other information, navigation bars, advertisement, and the copyright, etc. Most especially, it could extract the Web content information which was not existed as the Table format. Experiments show that this method can improve the accuracy of the Web content information extraction and has a wide applicability.

Key words: tag window; extraction; DOM

信息抽取(Information Extraction)技术是近十几年发展起来的新领域，它起源于文本理解，属自然语言处理研究的领域。信息抽取是直接从自然语言文本中抽取事实信息，并以结构化形式描述信息，适用于信息查询、文本深层挖掘、问题自动回答等方面的应用。Web 信息进行抽取(Web Information Extraction, WebIE)是将 Web 作为信息源的信息抽取，其核心是抽取分散在 Internet 上的半结构化的 HTML 页面中隐含的信息。

万维网飞速发展，其上的信息源随之日益丰富。但是，Web 页面中经常含有广告链接、导航条、版权等非网页主题信息的内容，页面所要表达的主要信息经常被隐藏在无关的内容和结构中，限制了 Web 信息的可利用性。正确提取网页正文信息，实际上就是提取出页面要表达的主要内容。它是信息搜索(Information Search)、数据挖掘(Data Mining)、机器翻译(Machine Translation)和文本摘要(Text Digest)等 Web 信息处理的基础。

1 相关工作

传统的网页数据抽取方法通常是由包装器(Wrapper)完成的，但是，获取包装器中信息模式识别的知识是一个瓶颈问题。采用半自动化方法获取知识规则的 XWRAP 系统在进行网页抽取前，检查网页并进行预处理，最后将网页表示成一棵树[1]。

王琦等人[2]基于 DOM 规范，提出了基于语义信息的 STU-DOM 树模型，将 HTML 文档转换为 STU-DOM 树，并对其进行基于结构的过滤和基于语义的剪枝，完成了对网页主题信息的提取。

文献[3]提出在字符集中的区域提取文字的方法。该方法仅适合于主题文字集中的网页，不能提取在文字中间出现表格或链接等结构的文字。

Kristina Lerman 等人[4]提出了通过对行和列的分组，从 List 和 Table 中自动提取 Web 数据的方法。但它只在一些假定条件成立的情况下才能进行，并且需要分析许多网页之后，才能从单一的一个 List 中提出信息。

崔继馨等人[5]提出了基于 DOM 的 Web 信息抽取方法。该方法采用人工方式对样本页面附加语义信息，然后对样本页面中的样本记录进行标记，并通过机器学习的方法产生信息抽取规则；利用这些规则完成对相似结构网页的信息抽取。由于该方法需要人工参与，使得系统的可用性降低。

文献[6]提出了基于统计的网页正文信息提取方法。它根据网页中的 HTML 标记将网页表示成一棵树，然后利用树中的每个节点包含的中文字节数从中选择包含正文信息的节点。该方法适用于网页中所有正文信息均放在一个 td(td 是 HTML 语言提供的表格标记)中的情况下的网页正文提取。

收稿日期: 2006-01-18; 修返日期: 2006-04-05 基金项目: 国家预研资助项目(40104116)

作者简介: 赵欣欣(1978-)，女，辽宁人，博士研究生，主要研究方向为 Web 挖掘、个性化服务、人工智能; 索红光(1966-)，男，山东人，副教授，博士研究生，主要研究方向为自动文摘、Web 挖掘; 刘玉树(1941-)，男，山东人，教授，主要研究方向为数据挖掘、人工智能、计算机图形图像处理.

通过分析可知,现存的大多数网页正文提取方法不能处理网页正文部分被存放在多个 td 中的情况;不能处理一个 td 中含有不同内容的情况,即不能处理一个 td 中存放的不仅仅是网页正文的情况;不能处理网页正文文字长度很短,短到与网页其余部分文字(如广告、导航条)长度相当的特殊情况。而且,大多数方法均限于提取存放在 td 中的网页正文信息。但经过统计,存在大量未采用 Table 结构存放正文信息的网页,对于这种情况,上述方法都无能为力。基于此,本文提出了基于标记窗(Tag Window)的网页正文获取方法来解决上述问题。

2　基于标记窗的网页正文提取方法

基于标记窗提取网页正文信息的方法不仅适合于处理一个网页中所有正文信息均放在一个 td 中的情况,也适合于处理网页正文放在多个 td 中的情况。尤其重要的是,它能够解决非 Table 结构的网页正文提取问题。其实现简单,通用性好。

定义　称 HTML 中成对出现的标记为标记对,称 HTML 格式的网页中出现在 Title 之后的显示内容非空的标记对为标记窗。

基于标记窗提取网页正文信息的方法流程图如图 1 所示。具体步骤如下:

(1) 对网页进行规范化处理。如果一个网页满足如下条件[6],则它是规范化的网页:

①在除了网页标记 tag 外的地方出现的“〈”和“〉”用 < 和 > 代替。

②所有标记的属性值放在引号中,如〈a herf = " www. bit. edu. cn" 〉。

③所有的标记都是匹配的,即每个开始标记均对应着一个结束标记,如〈body〉和〈/body〉。

④所有的标记都是正确嵌套的,如〈a〉…〈b〉…〈/b〉…〈/a〉。

(2) 获取 Web 的 Title 及其各级标题〈h_1〉,…,〈h_g〉的标题内容。应用中国科学院的 ICTCLAS 分词系统对它们进行分词处理,去掉停用词、虚词,得到只包含实词的标题词序列 $S_{title} = \{W_{t_1}, W_{t_2}, \cdots, W_{t_m}\}$ 及各级标题词序列 $S_{h_i} = \{W_{h_{i_1}} W_{h_{i_2}}, \cdots, W_{h_{i_n}}\}$ $(i = 1, \cdots, g)$。

(3) 提取网页正文。

①找出所有的标记窗 tw,对每一个标记窗 tw_i $(i = 1, \cdots, N)$,去掉其中的 HTML 标记,得到不含任何 HTML 标记的字符串。

②对每个标记窗 tw_i 内的字符串分词,得到字符串序列。为了简化计算,仅取出 S_{tw_i} 中的实词,得到 $S_{tw_i} = \{W_{tw_{i_1}}, W_{tw_{i_2}}, \cdots, W_{tw_{i_q}}\}$。用 Levenshtein Distance 公式[7](式(1))计算标题词序列 S_{title} 与字符串词序列 S_{tw_i} 的距离:

$$L((x^1, \cdots, x^p), (y^1, \cdots, y^q)) = \begin{cases} p & q = 0 \\ q & p = 0 \\ \min(L((x^1, \cdots, x^{p-1}), (y^1, \cdots, y^q)) + 1, \\ L((x^1, \cdots, x^p), (y^1, \cdots, y^{q-1})) + 1, \\ L((x^1, \cdots, x^{p-1}), (y^1, \cdots, y^q - 1)) + \\ Z(x^p, y^q)) & \text{否则} \end{cases} \tag{1}$$

其中, $Z(x, y) = \begin{cases} 0 & \text{如果 } x = y \\ 1 & \text{否则} \end{cases}$

③比较 L 和 q 的大小,如果 $L < q$,则标记窗 tw_i 中的文字是正文信息,将其提取;否则,舍弃标记窗 tw_i 中的文字。

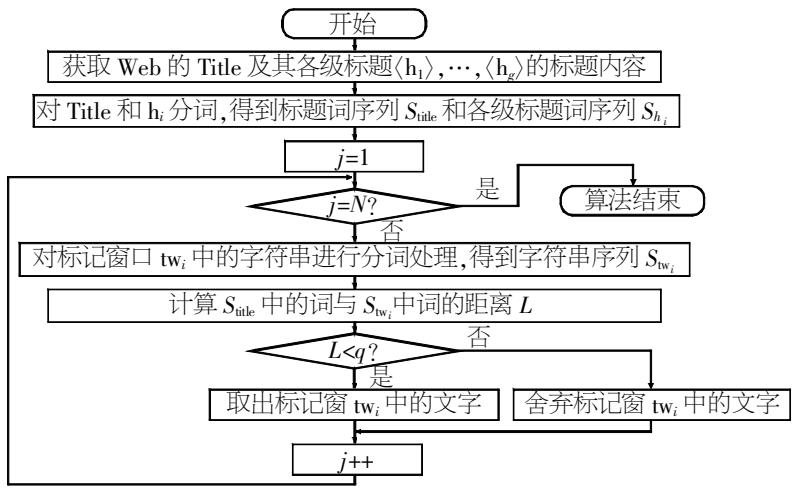

图 1　基于标记窗的网页正文提取方法流程图

对于标记嵌套的情况,先处理最里层的标记对,抽取其中的正文,并且清除此标记对及其中间的正文;然后处理外一层的标记对,依此类推。

下面用一个实际的例子来说明基于标记窗提取网页正文内容的方法。该例子所用的网页是 http://weather.sina.com.cn/news/2005/1216/12339.html,该网页的标题是“新疆发布大雾黄色预警信号”。

过滤掉该网页对应的 HTML 文件中显示内容为空的标记对后,用得到的标记窗中的四个标记窗为例进行说明。对每个标记窗分词之后得到词序列(仅保留实词)及标题词序列与它们之间的距离,如表 1 所示。按照前述算法,应该提取标记窗 1 ~3 中的文字信息,舍弃 4 中的信息。

表 1　网页实例

标题窗序号	标题词序列 S_{title}	对标题窗内的文字进行分词处理得到的词序列 S_{tw_i}	S_{title} 与 S_{tw_i} 的距离	q
1	新疆 发布 大雾 黄色 预警 信号	新疆 发布 大雾 黄色 预警 信号	0	6
2		新疆 气象台 今日 15 时 30 分 发布 大雾 黄色 预警 信号	4	12
3		今天 中午 乌鲁木齐 市区 开始 出现 浓雾 能见度为 300 m 预计 浓雾 将 维持 到 明天 上午 希 各 有关 单位 做 好 防雾 工作 注意 交通 安全	28	30
4		责任 编辑 潆 绕	6	4

3　实验与结果分析

为了考查本文提出方法的实际效果,随机选择来自 www.sina.com.cn, www.sohu.com, www.bit.edu.cn, www.people.com 和 www.mop.gov.cn 网站的 788 个网页进行实验。实验结果如表 2 所示。正确提取的网页数是指将网页的全部正文信息正确提取出来的网页个数;错提、少提正文信息的网页都是错误提取的网页。准确率 = 正确提取的网页数/网页总数。

表 2　实验结果

网页来源	网页总数	正确提取的网页数	错误提取的网页数	准确率
www.sina.com.cn	198	183	15	92.4%
www.sohu.com	186	178	8	95.7%
www.bit.edu.cn	80	78	2	97.5%
www.people.com.cn	211	195	16	92.4%
www.mop.gov.cn	113	109	4	96.5%
合计	788	743	45	94.3%

通过对结果的分析发现,之所以会出现对 (下转第 180 页)

为了解决生词复现率低的问题,编辑器提供了查询词频的功能。该功能可以实现某一课、某几课或整部教材中用了哪些词语、词语出现的频率有多少以及在各个章节中的分配情况等的查询。编者可以根据这些量化标准来调整编著过程。

为了解决词汇量过大的问题,编辑器按照汉语水平等级大纲(HSK) 来统计每章中出现的甲乙丙丁和超纲词各占多大比例,并且对每种不同等级的词汇以不同颜色来标示。所以编者可以很直观地从统计结果看出章节的编写是否科学。

2.4 其他功能

除上述汉语教材编著的专用功能外,编辑器还提供了以下功能:①生成词汇表。词汇表生成是对外汉语教材编写的一大需求,词汇表中要列出出现的所有词的字形、拼音以及词性等。如果通过人工方式去做,效率将会很低;而通过系统的此功能,在几秒之内即可完成。②章节合并。这一功能可以把多个文档合并成一个文档。③生成课件。编写的教材往往不只是为了出版用,而且教材中插入的一些多媒体信息如动画、声音、视频都无法出版。因此可以把编写的教材生成网络课件的形式,用于网络教学。④安全保存。保存功能不仅能把课件保存到服务器上,而且还能把课件保存到本地磁盘上。

对外汉语教材编辑器界面如图 4 所示。

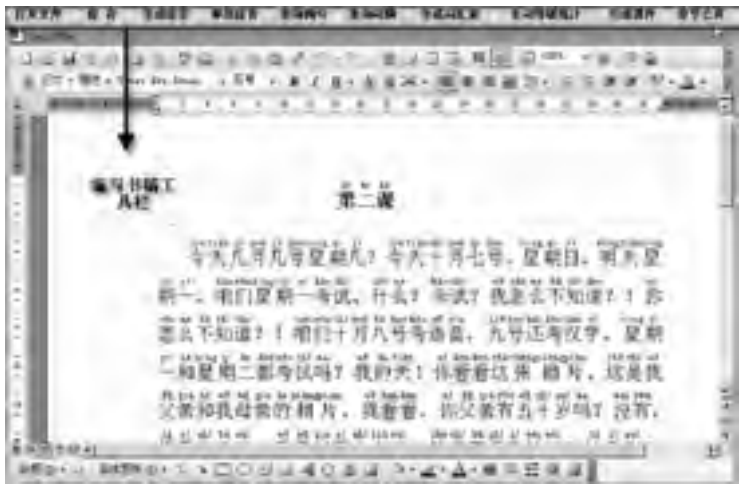

图 4 对外汉语教材编辑器界面

3 结束语

本文提出一个新型的对外汉语教材编著系统。这个系统不仅提供了协同编著过程中所需的功能,而且还有很多致力于解决汉语教材编著过程中遇到的难题的功能。整个系统是基于 J2EE 技术的松耦合设计的,因此具有很好的可维护性和可扩展性。但是系统中还有一个问题尚待解决,即一个用户只有当其在线时才能感知其他用户对文档在哪一部分作了哪些修改,但在其离线时则无法感知,这给协同编著带来了一定的不便。因此,笔者目前正在考虑实现当用户离线时,仍能收到系统自动发送的邮件。该邮件中应该包含哪个用户对文档的哪个部分作了怎样的修改。这一功能的实现,能在一定程度上提高系统的协同性。

参考文献:

[1] 卢伟. 基于 Web 的对外汉语教材多媒体协同编著系统设计与开发: 第八届国际汉语教学讨论会[C]. [出版地不详]: [出版者不详], 2005.

[2] 王琛, 赵正德, 殷小科. 基于 J2EE 的新型协同编著系统[J]. 计算机工程, 2003, **29**(13): 95-98.

[3] 裴云彰, 史元春, 徐光佑. 分布式计算模式下的协同编著系统[J]. 清华大学学报: 自然科学版, 2000, **40**(1): 76-79.

[4] 梁军华, 钟荣柏, 黄干平. 基于 CSCW 的协同编著系统的设计[J]. 武汉大学学报: 理学版, 2003, **49**(3): 341-344.

[5] 谢树煜, 张利, 安常青. 分布式多媒体协同编著系统的设计与实现[J]. 清华大学学报: 自然科学版, 1997, **37**(9): 20-23.

[6] 王长胜, 刘群. 双语语料库的检索和管理[J]. 计算机工程与应用, 2002, **38**(7): 113-114, 196.

[7] DECOUCHANT D, FAVELA J, MARTINEZ-ENRIQUEZ A M. PINAS: a middleware for Web distributed cooperative authoring: proceedings of the 2001 Symposimon Applications and the Internet, Jan. 8-12, 2001[C]. [S.l.]: [s.n.], 2001.

[8] DECOUCHANT D, MARTINEZ A M, MARTHEZ E. Alliance Web: cooperative authoring on the WWW: proc. of the CRIWG '99, International Workshop on Groupware[C]. Cancun: IEEE Computer Society, 1999.

[9] 鲁东明, 鲍宏伟, 潘云鹤. 多媒体文档协同编著系统的关键技术研究[J]. 通信学报, 1999, **20**(9): 41-46.

[10] 吕学强, 任飞量, 黄志丹. 句子相似模型和最相似句子查找算法[J]. 东北大学学报: 自然科学版, 2003, **24**(6): 531-534.

(上接第 145 页) 网页正文信息的错提、少提,是因为网页设计者想法的不同,导致他们可能使用一些修辞手法,如比喻、拟人等手法吸引 Web 访问者的浏览注意力。这样就导致了标题词序列中的词根本没有在网页正文中出现,造成了对网页正文信息的错误提取,影响了网页正文信息提取的准确率。

4 结束语

Web 上的数据抽取技术是目前点的热研究方向。虽然国内外的研究在一些技术上较为成熟和完善,但仍没有一个产品或系统能在各方面符合人们对 Web 信息抽取的要求。随着新技术和新思想的介入, Web 信息抽取技术处于不断的更新和发展中。本文提出的基于标记窗的网页正文提取方法解决了网页正文存放在多个 td 中的情况和网页正文文字短的网页正文提取问题。尤其重要的是,它能够解决非 Table 结构的网页正文提取问题。与文献[2, 6] 不同的是,本方法无须将网页表示成一棵树,只需利用正则表达式,就可以直接提取出网页中标记对之间的正文,这大大降低了算法的复杂度。实验表明,该方法性能好、适用性强。下一步的工作就是将语义信息结合到该方法中,更好地"理解"网页,进一步提高 Web 网页正文信息提取的准确率。

参考文献:

[1] IU L, PU C, *et al*. XWRAP: an XML-enable wrapper construction system for the Web information source: proceeding of the 16th IEEE International Conference on Data Engineering[C]. San Diego: [s.n.], 2000: 611-620.

[2] 王琦, 唐世渭, 杨冬清, 等. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展, 2004, **41**(10): 1786-1792.

[3] FINN A, KUSHMERICK N, SMYTH B. Fact or fiction: content classification for digital libraries: the 2nd DELOS Network of Excellence Workshop on Personalization and Recommender Systems in Digital Libraries[C]. Dublin: [s.n.], 2001.

[4] LERMAN K, KNOBLOCK C, MINTON S. Automatic data extraction from lists and tables in web sources: Automatic Text Extraction and Mining Workshop(ATEM-01)[C]. Seattle: [s.n.], 2001.

[5] 崔继馨, 张鹏, 杨文柱. 基于 DOM 的 Web 信息抽取[J]. 河北农业大学学报, 2005, **28**(3): 90-93.

[6] 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.

[7] LEVENSHTEIN V I. Binary codes capable of correcting deletions Insertions and Reversals[J]. **Soviet Physics-Doklady**, 1966, **10**(8): 705-710.