灵活结构网页的正文提取

殷 彬，杨会志
（电子科技大学 中山学院，广东 中山 528400）

摘 要：在Web数据挖掘中，由于网页大多都含有指向其他页面的超链接等噪音信息，为了减少噪音信息对Web数据挖掘效果的影响，有必要对网页进行净化处理，提取其中的正文，同时，现实中很多网页的代码结构不是特别规范，对此，提出一种对灵活结构网页适用的正文抽取算法。将网页用HTML标签分割成节点形式，找出其中含有正文内容的一个节点，以此节点为基础向前和向后进行余下正文内容的抽取。实验结果表明，本算法的适用性强、正确率较高。

关键词：Web数据挖掘；网页内容提取；正文节点；超链接节点；节点权值；链接密度
中图分类号：TP391 **文献标识码**：A **文章编号**：1673-629X(2011)09-0111-03

Content Extraction Based on Unknown Structure Web

YIN Bin, YANG Hui-zhi
(Zhongshan Institute, University of Electronic Science and Technology of China, Zhongshan 528400, China)

Abstract: There is often some useless information in the Web page, such as hyperlinks, copyright, which will affect the accurateness of Web data mining results. Extracting useful text content from a Web page for the mining is necessary. On the other hand, some pages' HTML codes are not standard. To solve this problem, propose an approach of Web information extraction based on unknown structure Web. It splits a Web page into a lot of nodes using HTML tags, then finds out one of the nodes which contained valuable information, and searches out other informative content nodes in front or back of the node, finally extracts the article from the Web page after connecting all found nodes' contents together. Experiments show that the arithmetic can deal with unstructured Web pages and is effective.

Key words: Web data mining; Web information extraction; content node; hyperlink node; node weight; link density

0 引 言

对于Web数据挖掘来说，一项非常基础和重要的工作就是把网页中有价值的正文内容与其他噪音内容分离开。因为噪音内容往往会影响挖掘结果的准确性。

对于网页正文内容提取已经有一些研究，主要分为两大类，第一类主要针对一些有共同模式的网页集进行提取。比如黄豫清和杨建武等人的从Web文档中构造半结构化信息的抽取[1,2]，仲华等人的基于XML架构的信息抽取[3]，赵金仿等人的根据同类网页自身结构特点的网页信息抽取[4]，BURGET R的基于网页视觉特征提取正文[5]，李石君的基于HTML模式代数的Web信息提取[6]，LI Yu等人的基于标签序列和树匹配正文抽取[7]，这些算法先在训练网页集中总结出暗含的模式，然后按照此模式对其余网页的内容进行抽取和处理。

第二类主要对于没有固定模式的各种单独的网页的正文内容进行抽取。如GENG Hua等人的基于DOM树的Web信息提取[8,9]，LIN Shian-Hua等人的基于<table>标签的正文信息块的抽取[10]，孟军等人的节点频度和语义距离相结合的网页正文信息抽取[11]，宋明秋的根据中文标点符号确定正文内容的算法[12]。

文献[8~12]对于无固定模式的单独的网页能进行有效的正文内容抽取，但它们要求被抽取网页的代码要规范，即代码中如果有一个HTML标签<tag>开始，则必需有一个对应的标签</tag>结束，而且要求HTML标签对不能错误嵌套。但网络上存在大量的代码不规范的网页。这样要使用文献[8~12]中的算法对不规范代码的网页进行正文抽取的话，必须提前对网页代码进行规范化修正，而代码规范化本来也是一件比较有难度的事情，所以，文中在不修正网页代码的情况下，提出一种对代码结构灵活网页的正文内容抽取算法。

收稿日期：2011-02-22；**修回日期**：2011-06-04
基金项目：中山市科技计划项目(20092A210)
作者简介：殷 彬(1978-)，男，讲师，硕士，研究方向为Web数据挖掘、Web商务智能；杨会志，教授，博士，研究方向为数据仓库，数据挖掘。

1 文中算法

文中认为网页文件是由很多节点构成，节点记为 HTMLNode，其结构为 HTMLNode{存放标签的 tag；存放节点内容的 content；存放节点权值的 nodeWeight；存放该节点是否为超链接节点的布尔变量 isALink}。其中节点权值 nodeWeight 的概念参见定义 1，在计算节点权值时，为了英文字母和汉字计算统一，一个汉字的长度计为 2 。以图 1 中的网页源代码为例，图中一共有 8 个节点组成。分别为：

HTMLNode(1){tag = "<html>"; content = ""; nodeWeight = 0; isALink = false}

HTMLNode(2){tag = "<body>"; content = ""; nodeWeight = 0; isALink = false}

HTMLNode(3){tag = "<h1>"; content = "人文社科系召开 2009 年度年终总结大会"; nodeWeight = 1156; isALink = false}

HTMLNode(4){tag = "</h1>"; content = "人文社科系 2009 年度年终总结大会于 2010 年 1 月 11 日下午……"; nodeWeight = 8100; isALink = false}

HTMLNode(5){tag = ""; content = "更多新闻"; nodeWeight = 0. 1451; isALink = true}

HTMLNode(6){tag = ""; content = ""; nodeWeight = 0; isALink = false}

HTMLNode(7){tag = "</body>"; content = ""; nodeWeight = 0; isALink = false}

HTMLNode(8){tag = "</html>"; content = ""; nodeWeight = 0; isALink = false}

```
1. <html >
2. <body>
3. <h1>人文社科系召开 2009 年度年终总结大会
4. </h1> 人文社科系 2009 年度年终总结大会于 2010
年 1 月 11 日下午……
5. <a herf = 'mores. asp'>更多新闻
6. </a>
7. </body>
8. </html>
```

图 1 网页文件源代码例子

定义 1(节点权值) 节点中含有该节点的节点标签和对应的正文内容，记节点标签的字符长度为 β，记节点正文内容的字符长度为 μ，则节点权值 nodeWeight 的计算公式为：

$$\text{nodeWeight} = \left(\frac{\mu}{1 + (1 - \lambda) * \beta}\right)^2$$

当节点标签为 $< a >$ 时，$\lambda = 0$；

当节点标签不为 $< a >$ 时，$\lambda = 1$。

网页中的正文节点的权值一般都很大，超链接节点的权值一般很小，故可以借助节点权值辅助判断某节点是否为正文内容的一部分。

在流行的网站随机选取了 893 个含有正文的网页，对它们进行了节点权值最大的节点是否属于网页的正文进行统计，其结果如表 1 所示。从表 1 可以得出观察 1。

表 1 权值最大节点内容是否属于正文的统计结果

网页来源	网页数量(个)	属于正文(个)	属于率
QQ 论坛 (bbs. qq. com)	146	145	99.3%
百度知道论坛 (zhidao. baidu. com)	246	244	99.1%
新浪网 (www. sina. com. cn)	263	263	100%
新华网 (www. newhuanet. com)	179	179	100%
教育网表格网页 (www. edu. cn)	59	58	98.3%

观察 1 通常情况下，节点权值最大的节点是正文节点，它所含的内容属于正文的一部分。

基于观察 1，只要先找出一个含正文内容的正文节点，然后以此节点出发向前和向后搜索紧挨着的其余正文节点，最后把所有正文节点所含的内容按顺序连接起来就形成了网页的正文。详细的算法见图 2 。

按照上述思路寻找其余正文节点，什么条件下寻找终止呢？终止的情形可以有 2 种：

(1)遇到一个超链接节点或全部节点都被搜索完毕；

(2)遇到较多超链接节点或全部节点都被搜索完毕。

对于第 1 种情形，如果正文中嵌入少量超链接，则终止后，正文节点没有找完，抽取的网页正文内容就不完整。而且现在很多情形下，正文中都有少许超链接。所以这种方法适应性不强。

对于第 2 种情形，很难确定到底遇到多少个超链接节点才比较合适？采用链接密度的概念，当链接密度超过了规定的阈值，则停止寻找。链接密度的概念如定义 2。

定义 2(链接密度)以第一个出现的超链接节点开始计算，选出此后的 $k-1$ 个内容非空的节点，在选出的共 k 个节点中，则链接密度 LDensity_k 的计算公式为：

$$\text{LDensity\_}k = \frac{\sum_{i=1}^{k} \lambda_i \mu_i}{\sum_{i=1}^{k} \mu_i}$$

当节点 i 的标签为 < a > 时,$\lambda_i = 1$;

当节点 i 的标签不为 < a > 时,$\lambda_i = 0$。

当链接密度比较大时,说明节点块中超链接非常多,相应的节点块应为链接块,而非正文内容块;相反,相应的节点块很可能为正文块,属于网页正文内容的一部分。

选取合适的链接密度阈值,当寻找过程中链接密度大于该值后,就停止寻找过程。

通过链接密度的方法,可以很好从正文含少许超链接的网页中抽取出正文,具有很强的适应性。

```
输入:网页代码
输出:网页中的正文
H=read(Html);
H=clearHtml(H); //清除 H 中的噪音标记(脚本/css/注释/表单/版权/免责声明)
vHTMLNodes=splitIntoHTMLNode(H); //把 H 按照节点 HTMLNode 的形式存储到向量 vHTMLNodes 中
LinkT=LinkT0; //设定链接密度阈值
Jma=findMax(vHTMLNodes); //从 vHTMLNodes 中选出权值最大的节点编号
front=1;
for(i=Jma-1;i>=1;i--){
LDensity_5=computeLD(vHTMLNodes[i]);//计算链接密度
if(LDensity_5>=LinkT){
    front=i;
    exit for;
  }
}
after= vHTMLNodes.length;
for(i=Jma+1;i<=vHTMLNodes.length;i++){
LD=computeLD(vHTMLNodes[i]);//计算链接密度
if(LDensity_5>=LinkT){
    after=i;
    exit for;
  }
}
for(i=front;i<=after;i++) //打印正文信息
  print vHTMLNodes[i].content;
```

图 2　正文抽取算法

2　实验与分析

对典型的站点随机选取了一些网页,同时构造了一个如图 3 所示的不规范的网页。并用此算法对这些网页进行正文抽取,程序均由 Java 语言实现。经过多次测试,k 取值为 5、链接密度阈值取 0.7 比较合适。通过与人工提取出的正文内容做比较分析,抽取的结果如表 2 所示。从结果可以看出,本算法正确率是较高的。对于不能正确抽取的网页一个特征为正文节点的内容非常短,另一特征为正文中连续嵌入了多个超链接。

```
<html >
<body>
<h1>人文社科系召开 2009 年度年终总结大会
</h1> 人文社科系 2009 年度年终总结大会于 2010 年 1 月 11 日下午……
```

图 3　代码不规范网页文件源代码例子

表 2　算法测试结果

网页来源	测试数量(个)	抽取正确数量(个)	抽取错误数量(个)	正确率
新浪网 www.sina.com.cn	263	261	2	99.2%
腾讯网 www.qq.com	288	285	3	98.9%
新华网 www.newhuanet.com	179	170	9	94.9%
教育网 www.edu.cn	227	216	11	95.1%
北京大学 www.pku.edu.cn	167	163	4	97.6%
代码不规范网页(图 3)	1	1	0	100%

3　结束语

网页的正文内容抽取可作为网页分类、网页聚类和搜索引擎等应用的预处理过程,文中提出的算法能较好地抽取网页正文。算法具有如下特点:

(1)适应性强。

算法[8~12]的正确运行都需要网页的 HTML 标签对<tag>和</tag>要成对出现且要正确嵌套,本算法不需要这样的前提要求。另外,算法[10,11]要依赖特定的标签 或<table>,但这两个标签并不是所有网页中都有的,本算法不需要被特定标签所限定,这样对于未来可能用到一些新标签的网页也同样适用。

(2)正确率较高。

运用本算法对典型网站随机选取的网页进行的正文内容抽取,正确率最小都达到了 94% 以上。

本算法对正文节点的内容非常短和含连续超链接的网页的正文抽取效果不佳。下一步的研究方向主要对此算法进行改进,使其能适用前两种特征的网页的正文内容抽取。

参考文献:

[1]　黄豫清, 戚广志, 张福炎. 从 WEB 文档中构造半结构化信息的抽取器[J]. 软件学报, 2000, 11(1):73-78.

(下转第 117 页)

$$\frac{\alpha S}{1-C}\cdot\frac{|(b_3-b_1)\times(b_0-2b_1+b_2)|}{|b_3-b_1|^3}$$

$$=4\cdot\frac{|(p_2-p_0)\times(p_0-2p_1+p_2)|}{|p_2-p_0|^3}$$

又由式(11)得

$$\left|\frac{1}{4}\cdot\frac{\lambda^2\alpha S}{1-C}\cdot(b_0-2b_1+b_2)-(p_0-2p_1+p_2)\right|\times(p_2-p_0)=0$$

所以,存在μ使得

$$\frac{1}{4}\cdot\frac{\lambda^2\alpha S}{1-C}\cdot(b_0-2b_1+b_2)-(p_0-2p_1+p_2)=\mu(p_2-p_0)$$

令$v=\frac{1}{4}\cdot\frac{\lambda^2\alpha S}{1-C}$,则

$$v(b_0-2b_1+b_2)-(p_0-2p_1+p_2)=\mu(p_2-p_0),$$

$$b_0-2b_1+b_2=\frac{1}{v}(p_0-2p_1+p_2)+\frac{\mu}{v}(p_2-p_0)\qquad(12)$$

综上,两曲线曲率连续拼接的条件为式(10~12)。由上述方法,可根据工程需要让两种曲线在拼接点达到G^1,G^2和曲率连续。

5 结束语

C-B样条曲线与三次均匀B样条曲线有着几乎相同的性质,与NURBS有着类似的同样有效的算法。文中给出了C-B样条的任意分割法,并给出了C-B样条之间G^1光滑拼接条件且应用于花瓶旋转曲面造型,最后给出C-B样条与均匀B样条之间光滑拼接的结论,得到的结果可直接应用于工程。

参考文献:

[1] 施法中.计算机辅助几何设计与非均匀有理B样条[M].北京:高等教育出版社,2001.

[2] 刘焕章,刘旭敏.基于结点调整B-spline曲面G^1/G^2的光滑拼接[J].计算机工程与应用,2007,43(14):60-63.

[3] 王文涛,汪国昭.带形状参数的双曲多项式均匀B样条[J].计算机辅助设计与图形学学报,2005,17(4):625-633.

[4] 蒋大为,刘　哲.等距曲线B样条光顺逼近[J].计算机辅助设计与图形学学报,1994,6(2):89-94.

[5] 谌炎辉,周良德.N二次曲面的NURBS表示及控制顶点和权因子的计算[J].湘潭大学自然科学学报,2005,27(2):151-154.

[6] Zhang Jiwen. C-curves: An extension of cubic curve[J]. Computer Aided Geometric Design,1996,13(3):199-217.

[7] Zhang Jiwen. Two different forms of C-B-splines[J]. Computer Aided Geometric Design,1997,14(1):31-41.

[8] Mainar E. Shape-preserving alternatives to the rational Bezier model[J]. Computer Aided Geometric Design,2001,18:37-60.

[9] Morin G. A subdivision scheme for surfaces of revolution[J]. Computer Aided Geometric Design,2001,18:483-502.

[10] 陶淑一,吴庆标.基于约束优化的B样条曲线形状修改[J].计算机工程与应用,2006,18(6):37-39.

[11] 韩旭里,刘圣军.三次均匀B样条曲线的扩展[J].计算机辅助设计与图形学学报,2003,15(5):576-578.

[12] 郭清伟,朱功勤.两相邻Bezier曲线近似合并的一种方法[J].中国科学技术大学学报,2003,33(5):518-523.

(上接第113页)

[2] 杨建武,陈晓鸥.半结构化文档集的结构模式提取的研究与实现[J].计算机工程,2001,27(10):19-21.

[3] 仲　华,崔志明.基于XML的信息抽取和多层向量空间技术研究[J].计算机技术与发展,2007,17(7):49-52.

[4] 赵金仿,赵　艳,缪建明.网页信息抽取及其自动文本分类的实现[J].计算机技术与发展,2008,18(10):37-39.

[5] Burget R. Layout Based Information Extraction from HTML Documents[C]//The Ninth International Conference on Document Analysis and Recognition. [s. l.]:[s. n.], 2007.

[6] 李石君,于俊清,欧伟杰.基于HTML模式代数的Web信息提取方法[J].计算机研究与发展,2006,43(9):1644-1650.

[7] Li Yu, Meng Xiaofeng, Li Qing, et al. Hybrid Method for Automated News Content Extraction from the Web[C]//Web Information Systems Engineering (WISE2006). Wuhan:[s. n.], 2006.

[8] Gupta S, Kaiser G, Neistadt D, et al. DOM-based Content Extraction of HTML Documents[C]//The 12th International Conference on World Wide Web. [s. l.]: [s. n.], 2003.

[9] Geng Hua, Gao Qiang, Pan Jingui. Extracting Content for News Web Pages Based on DOM[J]. International Journal of Computer Science and Network Security, 2007, 7(2):124-129.

[10] Lin Shian-hua, Ho Jan-ming. Discovering informative content blocks from Web documents[C]//ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. [s. l.]: [s. n.], 2002.

[11] 孟　军,刘秋水,王秀坤.节点频度和语义距离相结合的网页正文信息抽取[J].计算机工程与应用,2009,45(1):140-143.

[12] 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.