目前互联网上公布出来的正文提取算法

齐码代码 2014-03-13 **2597** 阅读

目前互联网上公布出来的正文提取算法，大家可以综合比较下，一起来测试下哪个更好用。 词网–北京词网科技有限公司http://demo.cikuu.com/cgi-bin/cgi-contex 猎兔网页正文提取 http://www.lietu.com/extract/ PHP版网页正文提取http://www.woniu.us/get_content_demo/ 网页正文提取分析(DEMO) http://61.128.196.27/txt 个人认为 http://61.128.196.27/txt 这个提取最牛，基本上无论什么页面都能提取出来，而且能有效的保持原文风格、图片、链接。
http://code.google.com/p/joyhtml/
看看这个效果不错
http://www.likeshow.net/article.asp?id=92
我一年前写的玩意 虽然不完善 但尚可用之在新闻和BLOG 论坛提取上 提取的正文对于BLOG和BBS包含评论及回复 具体原理也写很清楚了
如题，想从html源码中提取正文内容，<P></P>之间的内容，但是<P>的写法不规则。除了正则表达式的方法，还有其它的提取方法吗？谢谢！
最新下载
在线演示和最新下载：
http://www.shoula.net/ParseContenthttp://www.pudn.com/downloads152/sourcecode/internet/search_engine/detail668443.html

Google Code开源网页正文提取cx-extractor2010-05-19 12:31基于行块分布函数的通用网页正文抽取：线性时间、不建DOM树、与HTML标签无关
简述：
对于Web信息检索来说，网页正文抽取是后续处理的关键。虽然使用正则表达式可以准确的抽取某一固定格式的页面，但面对形形色色的HTML，使用规则处理难免捉襟见肘。能不能高效、准确的将一个页面的正文抽取出来，并做到在大规模网页范围内通用，这是一个直接关系上层应用的难题。
作者提出了《基于行块分布函数的通用网页正文抽取算法》，首次将网页正文抽取问题转化为求页面的行块分布函数，这种方法不用建立Dom树，不被病态HTML所累（事实上与HTML标签完全无关）。通过在线性时间内建立的行块分布函数图，直接准确定位网页正文。同时采用了统计与规则相结合的方法来处理通用性问题。作者相信简单的事情总应该用最简单的办法来解决这一亘古不变的道理。整个算法实现不足百行代码。但量不在多，在法。
项目网址：http://code.google.com/p/cx-extractor/
算法描述：基于行块分布函数的网页正文抽取算法.pdf
欢迎大家提出意见~

http://www.ngiv.cn/post/204.html
VIPS算法对搜索引擎的意义
http://blog.csdn.net/tingya/archive/2006/02/18/601954.aspx

基于视觉的Web页面分页算法VIPS的实现源代码下载
http://blog.csdn.net/tingya/archive/2006/04/28/694651.aspx
作者信息：飞跃,javascript教程-技术之家博客的博主

http://www.madcn.net/?p=791

我这里有个开源的项目，还不错，你上googlecode搜索joyhtml。
http://gfnpad.blogspot.com/2009/11/blog-post.html
下面几个是一些开源的程序：
1.一个python的基于文本密度的程序：
http://ai-depot.com/articles/the-easy-way-to-extract-useful-text-from-arbitrary-html/
ps：里面有bug，要稍加改动。 另外，对于没有对html注释部分进行处理
2.Java 开源项目： Gate
http://gate.ac.uk/

其实可以利用Dhmtl对象进行编程分析，已获得所要的数据文件，详细请看我的程序
http://www.vbgood.com/thread-94788-1-1.html
http://download.csdn.net/source/568439

一．标题块
l 分块节点：td，div，h，span
l 一般位于Head/Title的位置
l 当前单元含有<h1>-<h3>，<b>，<i>，等标签
l 样式，一般class包含title，head等字符
l 文字长度，一般大于3个字符，小于35个字符

二．发表时间块
l 分块节点：td，div， span
l 文字长度，一般小于50个字符
l 包含日期格式（2010-08-09）的字符串
l 包含以下关键字：来源，发表

三．主题块
l 分块节点：td，div
l HTML网页中有一些特殊标签，通常只出现在网页主题块中，如<P>
等。因此，主题块中往往包含着特殊标签。
l 主题块内容含有较多的句子，因此具有较多逗号、句号等标点符号（>5）。
l 若从信息量角度考虑，主题块一般是含有较多文字信息。
l 主题块的 标签密度=1000*标签数/文字数 应在小于一个范围。
l 主题块的 文本密度=len(文本)/len(HTML代码) 较大
l 不应该包含 “上一篇”，“下一篇”
l 包含以下字符串的内容块，判定为包含版权信息，需减权：“ICP备04000001号”，“版权所有”，“Copyright”
l 主题块序号在标题块之下
l 主题块序号在发表时间块之下
l 主题块序号在相关链接块之上

四．相关链接块
l 分块节点：td，div
l 文字应为“相关链接”、“相关新闻”、“相关报道”等敏感词，且连接比例很高。
l 链接数小于20

实现：
根据以上信息块特征，采用特征提权算法，C#（3.5）编程实现，命名为QD正文提取组件。经测试，对Html格式规范的以文字为主的内容页，正确提取率在85%以上，各大门户的新闻页面在95%以上。 例子下载(需要安装Microsoft .NET Framework 3.5)

注：QD正文提取组件 不开源，需要源码的朋友可选择付费获取。

这时挑选出的正文一般也就是到位了，但是问题是很可能在头尾残留了一些块广告。我认为这些块广告与正文中广告有很大的不同。这些广告的马脚就是其父节点，它们的父节点要么也包含了正文所在区域，也就是和正文平级，要么本身就是正文所在区域的一个子节点，很难是正文节点本身的。那么对疑似正文节点进行一次扫描，剔除那些父节点文字内容过大（包含了广告以及正文，即和正文平级）的块，也剔除那些父节点文字内容过小的块。
经过这样的处理，得到的内容基本上就是我们需要的正文了。下面就是要提取标题。
在代表整个网页的document中扫描一次，寻找那些有font字体的，strong的，h1的，title的节点，提取他们的信息。然后将得到的文字内容分词，查验分出来的词有多少是被正文包含的，包含最多的一半就是标题。但是这里要注意，有时候找到的节点本身是正文节点的子节点，那么无论怎么分，分出来都是完全包含的，所以要剔除那些本身是正文一部分的疑似标题。这样做对大部分网页也是有效了，但是对仅有的标题就在正文节点里的那些页面，目前为止我还没有特别好的想法。
这些日子也研究了一些别人的论文，有很多思想都非常好，也有很多人想到用马尔科夫，人工神经来训练。也许以后我会考虑用用看吧。现在这样也还可以，呵呵。
？

这个算法我也写了一下，不过是用C++写的。
我不太懂楼上讨论的分页是什么意思，我通过分析dom树然后用文中提到的规则进行dom结点处理以及后续的处理。
我主要是想把网页中的内容按网页框架分开，把正文部分合在一起，然后用贝叶斯决策计算正文特征支持率
提取网页内容。
现在VIPS基本写完。
但是却也发现了些问题，
比如说有些结点的坐标提取出来会有提取不出分隔条，这是因为有少数坐标有些重叠。这里涉及到一个坐标的确定问题。
然后是结点分割规则问题，现在的页面是大部分是通过DIV来组织页面。而VIPS似乎更合适TABLE组织的页面，我试过用TABLE组织的页面，分得相当不错。
另外，TINYA上面的翻译似乎改了些规则，还有部分翻译不是很准确。比如虚拟文本的定义部分与原文有些出入，不知道TINYA有没有注意到。
最后，很感谢TINYA 对这个算法的介绍。

本文来自CSDN博客，转载请标明出处：http://blog.csdn.net/tingya/archive/2006/02/18/601836.aspx

作者：齐码代码

点赞

记录我的程序员之路
原文地址：目前互联网上公布出来的正文提取算法, 感谢原作者分享。

←通用的用户登录过滤器（SessionFilter）

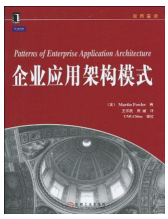
→浏览器调用windows本地应用程序

发表评论

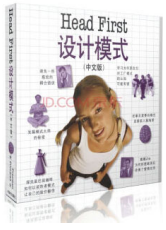
发表评论

好书推荐

代码整洁之道
[亚马逊]



企业应用架构模式
[京东 亚马逊]



您可能感兴趣的博客

您可能感兴趣的代码