

正文提取

2012年11月10日 15:29:56

阅读数：1023

基于统计信息

基于DOM的网页主题信息自动提取

<http://www.ccf.org.cn/resources/1190201776262/2010/05/12/h049617026.pdf>

总结：与主题无关的块总是含有大量的无关链接和极少非链接文字

[PDF] 使用特征文本密度的网页正文提取

<http://www.cqvip.com/qk/95939x/201003/32891243.html>

总结：与主题无关的块总是含有大量的“无关词”，如“版权”，“声明”，“搜索”，“首页”，“帮助”。可以计算无关词和总文本的比例。

基于标签密度的自适应正文提取方法

<http://wenku.baidu.com/view/773479eb998fcc22bcd10d12.html>

基于视觉

<http://hi.baidu.com/gghgdk/item/9d5d5e0945e3fe96a2df4308>

总结：正文节点在网页的位置总是在“中间”的，以及和其中国像元素的数量也有关联。

基于决策树

基于双决策的新闻网页正文精确抽取

<http://file.lw23.com/4/4f/4fa/4fa9ed31-f1fa-42c1-abea-51d9f143b4a9.pdf>

总结：人类识别正文段通过两个步骤：1.大概判断正文范围。2判读该正文范围内的段落是否属于正文部分。因此，机器识别可以通过全局和局部两个方面进行决策。

想法：对于决策树(暂时不理解其工作方式，求相关书目)的训练数据，可以通过这种方式获得。制作一个浏览器插件，类似于firebug或clipper的节点选择，可以选择页面的DOM元素，通过手工选取正文节点，该插件将数据传回服务器。通过这种方式将url和人工确定的正文节点对应，形成大量的训练数据。

基于包装器

通过为特定站点建立特定的包装器，即特定的正文节点获取模式，可以准确判断特定站点的正文节点。确定是需要手工确定站点。可以借助在“基于决策树”小节提到的训练数据获取方式来简化包装器的构建。

通俗来讲，就是为正文提取建立黑名单和白名单。

对当前某些插件的理解

研究了clearly的源码。源码的获取详见<http://blog.csdn.net/cattail2012/article/details/8168025>。从文件js/bulk.js的4320行起，描述的是该插件如何进行网页净化的。我称之为网页净化，因为clearly做的是这样一件事情：它从body 节点开始，对文档所有节点进行遍历处理，处理依据4419行的\$R.parsingOptions，对不同的节点进行不同处理，如保留该节点或者删除该节点，对节点的属性也进行删除或者修改，通过这种方式净化了页面元素。也就是说，clearly并没有做寻找正文节点这个工作，以此推测，对于 readability或pocket等插件，它们也都没有做提取正文节点的工作。而且对于它们的需求，也没有必要进行正文节点的获取。虽然这些插件没有进行正文提取，但是对于非正文节点的删除这个思想，可以使用在正文节点提取的算法中。

可行性分析

理论上，基于统计信息和视觉信息可以创建出可行的正文提取方案。

个人资料



cattail2012

关注

原创

4

等级：

访问： 2万+

积分： 368

排名： 22万+

粉丝

0

喜欢

0

评论

1

最新文章

JavaScript之111公使用Object dectect四个是

browser dectect

CSS @规则(at-rules)

字符集和编码

关于

字符集和编码

字符集和编码

归档

2012年11月

13篇

联系我们



请扫描二维码联系客服

webmaster@csdn.net

400-660-0108

QQ客服 客服论坛

关于 · 招聘 · 广告服务 · 网站地图

©2018 CSDN版权所有 京ICP证09002463号

百度提供搜索支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

相关文献(未读)

Machine Learning for Information Extraction in Informal Domains

<http://reports-archive.adm.cs.cmu.edu/anon/1999/CMU-CS-99-104.pdf>

[PDF]Fact or fiction: Content classification for digital libraries - Ercim

<http://www.ercim.eu/publication/ws-proceedings/DelNoe02/AidanFinn.pdf>

Two Approaches to Bringing Internet Services to WAP Devices

<http://www9.org/w9cdrom/228/228.html>

Seeing the Whole in Parts: Text Summarization forWeb Browsing on Handheld Devices

<http://lpubs.stanford.edu:8090/511/1/2001-45.pdf>

文章标签： web Web 机器学习 模式识别 [▼查看关于本篇文章更多信息](#)

[上一篇](#) gitignores手册 [下一篇](#) Javascript测试框架Jasmine

想对作者说点什么？

[我来说两句](#)



niaobirdfly 2014-03-19 09:44:00 #1楼

Machine Learning for Information Extraction in Informal Domains这篇文章不会太老了吧，那会儿的网页和现在的网页感觉不是一个级别的啊，难以想象那时的网页有多少广告、图片。。。



基于文本密度的新闻**正文**抽取方法之Python实现

1060

基于网页分析构思出的正文提取算法 回顾以上的网页分析，如果按照文本密度来找提取正文，那么就是写这么一...



WebCollector 网页**正文**提取算法(ContentExtractor)

8031

WebCollector 网页正文提取算法(ContentExtractor)WebCollector自2.10版起加入新闻网页正文自...

网页**正文**提取——Html2Article - CSDN博客

4-17

回顾以上的网页分析,如果按照文本密度来找提取正文,那么就是写这么一个算法,能够...这里所说的正文提取主要是针对新闻页面等网页的...

基于文本密度的新闻**正文**抽取方法之Python实现 - CSDN博客

8-2

基于网页分析构思出的正文提取算法回顾以上的网页分析,如果按照文本密度来找提取正文,那么就是写这么一个算法,能够从过滤html标签...

网页**正文**提取算法介绍

5229

查找发现了两个比较好的网页正文提取算法：国内：哈工大的《基于行块分布函数的通用网页正文抽取》该算...

**正文**提取 - CSDN博客

5-8

总结:与主题无关的块总是含有大量的无关链接和极少非链接文字 [PDF] 使用特征文本密度的网页正文提取 <http://www.cqvip.com/qk/95...>

基于行块分布函数的通用网页**正文**抽取算法初步认识 - CSDN博客

7-31


依据"\n"分行,若某文字行的上下存在两个空行,且此文字行长度小于阈值40,则...基于网页分析构思出的正文提取算法回顾以上的网页分析...

python 任意新闻**正文**提取

2979

在github上搜到一个正文提取程序，测试了一下基本可以对现在大多数大型新闻网站进行提取 后续我会分析一...

网页正文及内容提取算法

 5003

基于行块分布函数的通用网页正文抽取 [http://wenku.baidu.com/link?url=TOBoIHWT\\_k68h5z8K\\_Pmqr-wJMPf...](http://wenku.baidu.com/link?url=TOBoIHWT_k68h5z8K_Pmqr-wJMPf...)

一种基于文本抽取的网页正文去重算法

7-30

笔者结合二叉排序树设计了一种基于文本抽取的网页正文去重算法,本文给出了该算法...基于网页文字密度的正文信息提取算法 立即下载 ...

从HTML文件中抽取正文的简单方案 试验结果 - CSDN博客


3-9

观察样本集的数据可以发现,即使是内容型的大段文字,也有可能文本密度很低——...Python Show-Me-the-Code 第 0008 题 提取HTML正文...

 一种提取HTML网页正文的方法

 7339

这里所说的正文提取主要是针对新闻页面等网页的主体是文字的HTML页面。在做一些与文本处...

 女性得了静脉曲张变成蚯蚓腿怎么办？用这个方法坚持3个月全恢复！

水英电器 · 顶新

关于新闻博客类页面正文抽取 - CSDN博客

2-17

`get(key).getValue()); //通过页面个数和页面文字数,得到一个分值 eqKey...`...基于网页分析构思出的正文提取算法回顾以上的网页分析,如果...

网页正文提取 - CSDN博客


5-24


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...I 若从信息量角度考虑,主题块一般是含有较多文字信...

 php实现的网页正文提取算法

 2059

Html2Article-php实现的提取网页正文部分，最近研究百度结果页的资讯采集，其中关键环节就是...

 【Python】提取网页正文内容的相关模块与技术

 1821

【Python】提取网页正文内容的相关模块与技术 1、正文抽取地址 <https://github.com/b...>

 javascript 网页正文提取

 560

写这个的原因，最近在改一个网页正文提取的插件，但找遍了网站就是没有JS版的，于是乎就找...

 几个html网页提取正文的API和开源算法

 723

1. URL2io 提供网页信息提取服务 <http://blog.url2io.com/url2io-app-samples/pageless/> 2.reada...

 基于网页分析构思出的正文提取算法

 419

转自：<http://www.cnblogs.com/aisir/p/6142323.html> 参考文章链接：<http://www.cnblogs.co...>

 python beautifulsoup 抓取网页正文内容

 5738

使用python的 beautifulsoup 来抓取网页

如何抽取HTML正文

 2078

网页展现给用户的是主要内容是它的文本。因此，在获取网页源代码时，针对网页抽取出它的特定的文本内容...

从HTML文件中抽取正文的简单方案

 2.9万

2011.04.08 更新：想找此方案的代码的朋友请访问：<http://code.google.com/p/creamer> 从HTML文件中抽取...



网络爬虫之新闻页面自动提取正文

2014年05月21日855KB

下载

PHP 通用正文提取

PHP 通用正文提取，通用新闻数据提取。

3593



爬虫实战12-自动摘要及正文抽取

文章说明:

756



网络爬虫九-使用正则表达式抽取HTML正文和URL

正则表达式，又称正规表示法、常规表示法（英语：Regular Expression，在代码中常简称为rege...

1452



python爬虫进阶（十二）：自动摘要及正文抽取

一、文本长度分析 1、HTML中的换行 在HTML源码中，所有的换行都是依赖行级元素、块级元素...

483



基于行块分布函数的网页正文抽取算法代码实现

最近在做一个与资讯相关的APP，资讯是通过爬取获得，但是获取只有简单的信息，正文没有...

3041