Atitit 网络爬虫与数据采集器的原理与实践attilax著 v2

[1. 数据采集 1](#_Toc26645)

[1.1. http lib 1](#_Toc22525)

[1.2. HTML Parsers， 1](#_Toc3305)

[1.3. 怎么制作爬虫 2](#_Toc28717)

[1.4. 网页爬虫组成： 3](#_Toc20605)

[1.4.2. 爬虫进阶 爬虫的进阶就是需要与数据来源方斗智斗勇 3](#_Toc8357)

[2. 实现类库框架 4](#_Toc28307)

[3. 问题与难点（html转txt) 4](#_Toc21639)

[4. 参考资料 4](#_Toc21152)

# 数据采集

主要获取pagesUrls,artUrls, picUrls

可参考火车头

## http lib

## [HTML Parsers](http://best-java-source.whyandhow.org/java-cn/list/900500340/HTML-Parsers.shtml" \t "http://blog.csdn.net/java2king/article/details/_blank)，

第8章 web爬取199

8.1 一个简单爬虫算法199  
8.1.1 宽度优先爬虫201  
8.1.2 带偏好的爬虫201  
8.2 实现议题202  
8.2.1 网页获取202  
8.2.2 网页解析202  
8.2.3 删除无用词并提取词干204  
8.2.4 链接提取和规范化204  
8.2.5 爬虫陷阱206  
8.2.6 网页库206  
8.2.7 并发性207  
8.3 通用爬虫208  
8.3.1 可扩展性208  
8.3.2 覆盖度、新鲜度和重要度209  
8.4 限定爬虫210  
8.5 主题爬虫212  
8.5.1 主题本地性和线索213  
8.5.2 最优优先变种217  
8.5.3 自适应219  
8.6 评价标准223  
8.7 爬虫道德和冲突226  
8.8 最新进展228  
文献评注230

这学期去图书馆借书，无意间看到一本书**《网络机器人[Java](http://lib.csdn.net/base/17" \o "Java EE知识库" \t "http://blog.csdn.net/java2king/article/details/_blank)编程指南》**。看了下感觉如获至宝。

市面上讲爬虫的书可以说是没有，基本上只有在搜索引擎类的书里有提到，而且只是讲个思想，没有可以用的代码。

## 怎么制作爬虫

制作一个爬虫基本需要如下步骤：

* 确定目标：  
  制作爬虫首先需要清楚自己的目的，希望获取哪些信息。
* 观察目标：  
  然后通过观察目标网站或者程序，获取目标信息的入口和索引方式。
* 结构提取：  
  进入内容页面，观察感兴趣区域的页面结构，然后考虑通过正则表达式或者xpath等方式提取目标内容。
* 编码实现：  
  前3步完成之后，最后就是使用线程工具或者自己编码实现这个爬虫了

## 网页爬虫组成：

1. **一个url队列。**我们的爬虫从这个队列中读取url，并将新的url放入这个队列。这里最重要的是判重。简单的哈希就能达到判重的目的，但是为了节约空间（url的数量往往很多），一般使用bloomfilter的思想。bloomfilter与普通的哈希算法最大的不同就是bloomfilter只需要一个bit来表示某个元素是否存在，所以能节约空间。bloomfilter有一个小缺点，即准确率并不是百分百：判断一个元素是不是已经存在时，已有的有很小的可能会判断为不存在，但是没有的元素一定会判断为不存在。
2. **网页爬取模块。**需要能模拟浏览器发送请求。
3. **网页分析模块。**爬下来的是网页源码，可以用正则或者其他方法提取我们需要的信息。
4. **新的url生成模块。**生成新的url，放入队列。

#### 结构提取

scrapy用的是xpath做结构解析，所以这里需要获取到目标区域的xpath。xpath是什么参考[这里](http://www.w3school.com.cn/xpath/)。

chrome浏览器有一个很好的插件可以很方便的获取到某个html节点的xpath，[XPath Helper](https://chrome.google.com/webstore/detail/xpath-helper/hgimnogjllphhhkhlmebbmlgjoejdpjl?hl=en)，先安装好它。

打开chrome，输入URL（[https://www.baidu.com/s?wd=爬虫&pn=0](https://www.baidu.com/s?wd=%E7%88%AC%E8%99%AB&pn=0)），然后F12进入开发者模式。通过鼠标找到第1个搜索结果的html节点，然后右键：copy-Copy XPath。

### 爬虫进阶 爬虫的进阶就是需要与数据来源方斗智斗勇

* IP被封禁
* 访问次数受限制
* 爬取速度太慢
* 页面重定向
* 页面ajax生成内容
* 页面内容由复杂的javascript生成

# 实现类库框架

Httpclient webdriver

/AtiPlatf\_auto/src\_atibrow/com/attilax/dataSpider/DoubanSpider.java

DoubanSpider be = **new** DoubanSpider();

be.search(kw);

be.clickFirst();

be.processShowMainTxt();

# 问题与难点（html转txt)

通过jsoup好像不行。。。Htmlpaser好点，单好多重复的。。

貌似使用浏览器ff来save as 最好的。。

Atitit.html转换提取纯文本txt

HTML Parser 1

Jsoup 1

Browser saveas 1

## ****有移动版的网站优先爬移动版** 站点选取**

# 参考资料

有什么介绍网络爬虫的书籍推荐吗\_百度知道.html

1. 网络爬虫需要阅读的书籍\_夸父逐梦\_新浪博客.html

《用Python写网络爬虫》([澳]理查德...)

《Python网络数据采集》(...)【简介\_书评\_在线阅读】

~$itit WebDriver技术规范原理与概念.docx

Atitit 基于**scrapy的 爬虫 架构原理**

Atiend