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](#_Toc28959)

[1.1. http lib 1](#_Toc28075)

[1.2. HTML Parsers， 1](#_Toc10243)

[1.3. 第8章 web爬取199 1](#_Toc16115)

[2. 实现类库框架 2](#_Toc26215)

[3. 问题与难点（html转txt) 2](#_Toc12468)

[4. 参考资料 3](#_Toc16261)

# 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  
作者:: 绰号:老哇的爪子 （ 全名：：Attilax Akbar Al Rapanui 阿提拉克斯 阿克巴 阿尔 拉帕努伊 ）

汉字名：艾提拉（艾龙），   EMAIL:1466519819@qq.com

转载请注明来源： http://blog.csdn.net/attilax

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。

# 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

# 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

atiend