Atitit.数据检索与网络爬虫与数据采集的原理概论

[1. 信息检索 1](#_Toc8590)

[1.1. 《信息检索导论》(（美）曼宁...)【简介\_书评\_在线阅读】 - dangdang.html 1](#_Toc7042)

[1.2. 《现代信息检索(原书第2版)（由信息检索领域的代表人物撰写，及时掌握现代信息检索关键主题的详细知识）》(（智）贝泽耶茨...) 2](#_Toc28971)

[2. 网络爬虫 2](#_Toc15117)

[2.1. 第8章 web爬取199 3](#_Toc31875)

[2.2. 《用Python写网络爬虫》([澳]理查德...) 4](#_Toc7102)

[3. 数据采集 4](#_Toc4198)

[3.1. 《Python网络数据采集》(...)【简介\_书评\_在线阅读】 - dangdang.html 4](#_Toc985)

[4. 爬虫框架与工具 5](#_Toc7154)

[5. 参考资料 5](#_Toc2112)

# 信息检索

## 《信息检索导论》(（美）曼宁...)【简介\_书评\_在线阅读】 - dangdang.html

第1章　布尔检索  
第2章　词项词典及倒排记录表  
第3章　词典及容错式检索  
第4章　索引构建  
第5章　索引压缩  
第6章　文档评分、词项权重计算及向量空间模型  
第7章　一个完整搜索系统中的评分计算  
第8章　信息检索的评价  
第9章　相关反馈及查询扩展  
第10章　XML检索  
第11章　概率检索模型  
第12章　基于语言建模的信息检索模型  
第13章　文本分类及朴素贝叶斯方法  
第14章　基于向量空间模型的文本分类  
第15章　支持向量机及文档机器学习方法

## 《现代信息检索(原书第2版)（由信息检索领域的代表人物撰写，及时掌握现代信息检索关键主题的详细知识）》(（智）贝泽耶茨...)

第1章　引言

第2章　用户搜索界面

第3章　信息检索建模

第4章　检索评价

第5章　相关反馈与查询扩展

第6章　文档：语言及属性

第7章　查询：语言及属性

第8章　文本分类

第9章　索引和搜索

第10章　并行与分布式信息检索

第11章　Web检索

第12章　Web爬取

第13章　结构化文本检索

第14章　多媒体信息检索

第15章　企业搜索

第16章　图书馆系统

第17章　数字图书馆

# 网络爬虫

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。

有什么介绍网络爬虫的书籍推荐吗\_百度知道.html

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：艾龙，  EMAIL:1466519819@qq.com

转载请注明来源： http://blog.csdn.net/attilax

## 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

## 《用Python写网络爬虫》([澳]理查德...)

　　第1章 网络爬虫简介 1

　　第2章　数据抓取　23

　　第3章　下载缓存　39

　　第4章　并发下载　57

　　第5章　动态内容　69

　　第6章　表单交互　89

　　第7章　验证码处理　103

　　第8章　Scrapy　121

　　第9章　总结　143

# 数据采集

主要获取pagesUrls,artUrls, picUrls

可参考火车头

## 《Python网络数据采集》(...)【简介\_书评\_在线阅读】 - dangdang.html

第1章　初见网络爬虫　　2

第2章　复杂HTML解析　　11

第3章　开始采集　　26

第4章　使用API　　42

第5章　存储数据　　61

第6章　读取文档　　80

第7章　数据清洗　　94

第8章　自然语言处理　　103

第9章　穿越网页表单与登录窗口进行采集　　120

第10章　采集JavaScript　　128

第11章　图像识别与文字处理　　139

第12章　避开采集陷阱　　154

第13章　用爬虫测试网站　　164

第14章　远程采集　　174

# 爬虫框架与工具

# 参考资料

Atitit 网络爬虫的原理与实践attilax著