# 爬虫

## 介绍

爬虫技术说简单也简单。比如使用jdk自带的UrlConnection即可实现。

再加上一些html的解析工具，比如jsoup，nekohtml等即可实现。

而作为http的客户端，可以使用到的技术就有httpClient、OKhttp等等。

甚至可以使用xml相关的解析技术。

爬虫在日常生活中是最常见，但却是最陌生的技术。比如我们使用的搜索引擎就会使用爬虫技术实现的。

## Jdk实现

方案：

0，初始条件：当前是首页

1，下载网页得到网页内容 <━━━━━━┓

2，获取其中的所有超链接 ┃ 当没有新的链接时就停止了。

3，去掉已经下载过的链接和站外的链接 ┃

4，循环处理每一个有效的超链接 ━━━┛

// 下载网页

pulbic static String download(String urlString){

URL url = new URL(urlString);

URLConnection conn = url.openConnection();

InputStream in = conn.getInputStream();

// in流中的数据就是网页内容

}

// 获取所有的超链接

pulbic List<String> getAllLinks(String pageContent){

"<a href=(.\*)>" // mailto:.. , javascript:..

//

<!--

<a href="http://www.itcast.cn">xxx</a>

-->

// 推荐使用 DOM + XPath

String expr = "//a";

List<Element> eltList = findByXpath(expr);

for(Element e : eltList){

String href = e.getAttribute("href");

...

}

}

www.163.com

news.163.com

blog.163.com

bbs.163.com (50W个主题) 500W

1000W

问题：

1，使用递归可能会内存溢出。

2，递归方案不能使用多线程。

3，使用递归没有“断点续抓”。

LinkedList

addFirst() 和 removeLast()

addLast() 和 removeFirst()

# 反爬虫

## 介绍

与爬虫相对的就是反爬虫。反爬虫也和限流技术有一定的交集。

## 相关技术

kk-anti-reptile 是适用于基于 spring-boot 开发的分布式系统的开源反爬虫接口防刷组件。

<https://gitee.com/kekingcn/kk-anti-reptile>