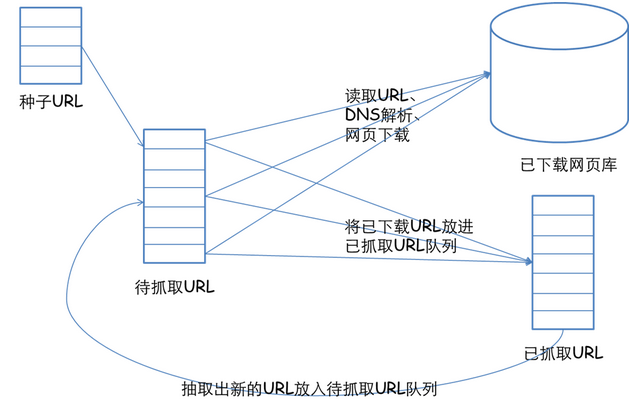
网络爬虫的基本结构和架构如下所示：

一、网络框架如下：

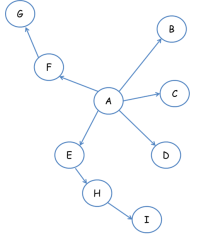


基本工作的流程如下：

1. 首先确定一个种子URL，然后找出要爬取内容的链接URL；
2. 再将要爬取URL中的链接URL放入待爬取的URL队列中；
3. 从待爬取的URL队列中取出URL，并将它们对应的网页爬取下来，利用页面解析，解析出我们需要的内容；
4. 分析以抓取的URL中的其他URL，并将它们放入待抓取的URL中，从而进行下一次循环。
5. 抓取策略

在爬虫系统中，待抓取URL队列是很重要的一部分。待抓取URL队列中的URL以什么样的顺序排列也是一个很重要的问题，因为这涉及到先抓取那个页面，后抓取哪个页面。而决定这些URL排列顺序的方法，叫做抓取策略。

1、此次我们采用了深度优先遍历策略，深度优先遍历策略是指网络爬虫会从起始页开始，一个链接一个链接跟踪下去，处理完这条线路之后再转入下一个起始页，继续跟踪链接。我们以下面的图为例：



遍历路径：A-F-G E-H-I B C D

2、更新策略

网页的更新策略主要是决定何时更新之前已下载过的页面，常见的更新策略有以下三种：

1. 历史参考页面

根据页面以往的历史更新数据，预测该页面未来何时会发生变化。一般来说，是通过泊松过程进行建模进行预测。

1. 用户体验策略

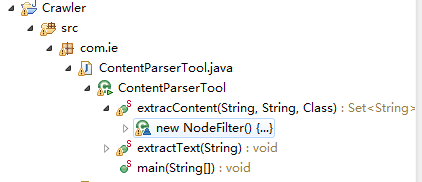
尽管搜索引擎针对于某个查询条件能够返回数量巨大的结果，但是用户往往只关注前几页结果。因此，抓取系统可以优先更新那些现实在查询结果前几页中的网页，而后再更新那些后面的网页。这种更新策略也是需要用到历史信息的。用户体验策略保留网页的多个历史版本，并且根据过去每次内容变化对搜索质量的影响，得出一个平均值，用这个值作为决定何时重新抓取的依据。

1. 聚类抽样策略

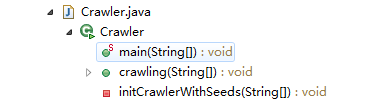
这种策略认为，网页具有很多属性，类似属性的网页，可以认为其更新频率也是类似的。要计算某一个类别网页的更新频率，只需要对这一类网页抽样，以他们的更新周期作为整个类别的更新周期。

1. API接口

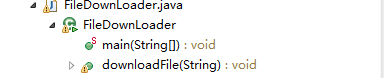
本次项目中接口的相关信息如下：



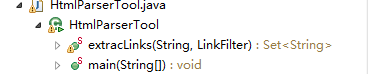
此类主要解析页面的内容，将爬取 的页面的内容中爬取出我们需要的内容，如书的作者，价格，简介等。



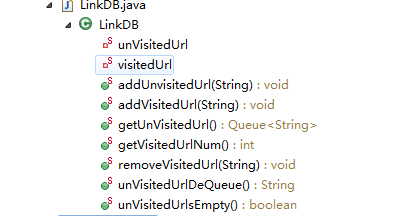
主类方法



主要实现页面的内容的下载。



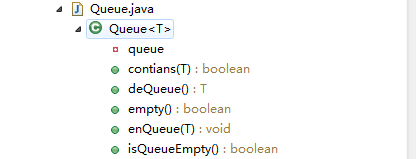
获取一个网站上的链接,filter 用来过滤链接



此类的主要功能是实现数据与数据库的连接，将数据传输到数据库中。



此功能主要实现网址URL的判断，判断此URL是否已经被爬取和待爬取的URL，将这些URL分别放入不同的队列中。



此类的主要实现对列的基本方法。