互联网是一个庞大的非结构化数据库，将数据有效的检索并组织呈现出来有着巨大的应用前景，搜索引擎作为一个辅助人们检索信息的工具成为用户访问万维网的入喉和指南。但是，这些通用写搜索引擎也存在不小的局限性，不同领域，不同背景的用户往往也具有不同的检索目的和需求。为了解决通用搜索引擎所返回结果包含大量用户不关心的页面，所以我们就需要使用爬虫。

网络爬虫应用宽度搜索技术，对url进行分析，去重。网络爬虫使用多线程的技术，让爬虫具备更加强大的抓取能力，网络爬虫要实现对特定主题的爬取。网络爬虫还要完成信息的提取任务，对于抓取回来的网页提取出来，然后获取使用者希望获取的信息，可以预见将来互联网信息抓取，挖掘，再处理，将成为人们越来越多的需求，而满足这种需求的就是各种各样的爬虫工具。传统的网络爬虫技术主要应用于抓取静态web网页，

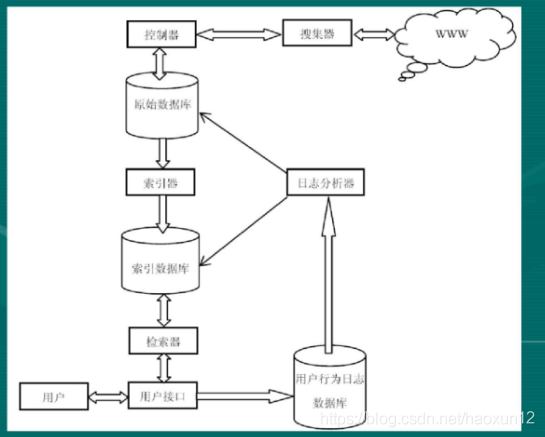


图1 爬虫工作原理

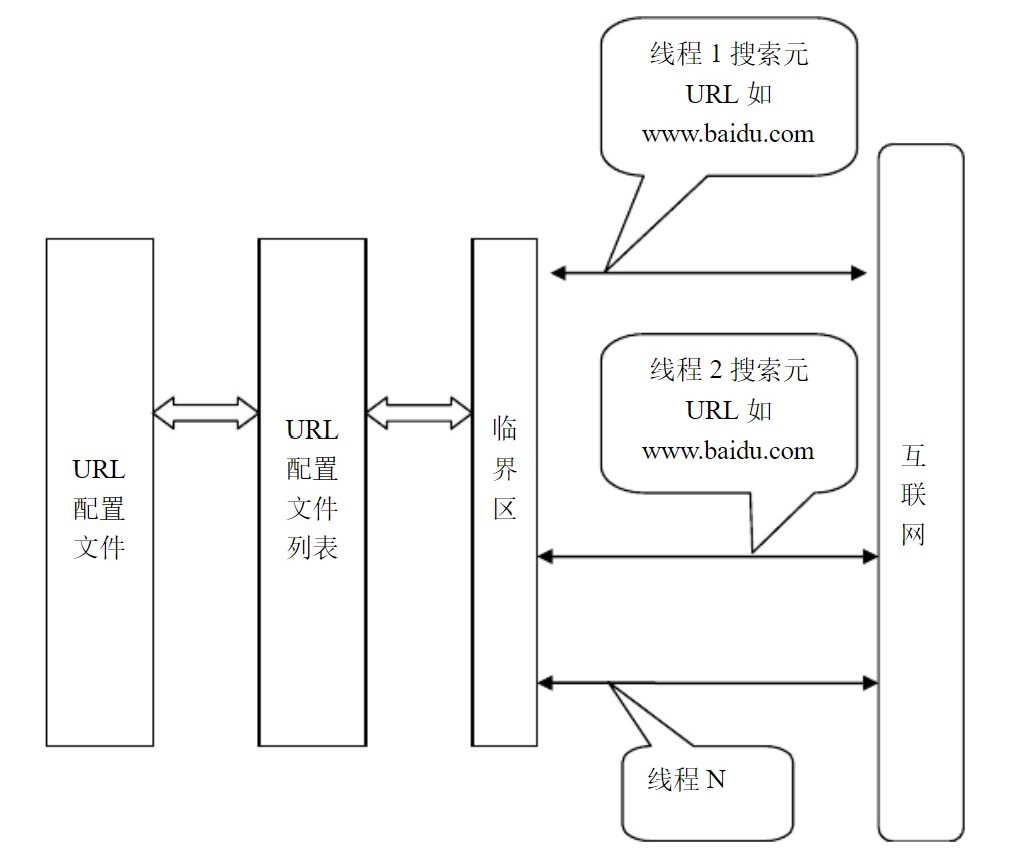


图2 多线程爬虫模型

明确地说，数据收集是一种基于既定的、系统的方式进行收集和衡量相关变量信息的过程。

在百度和Bing图像搜索中分别查询丰子恺漫画作品，收集少部分图像。再通过爬虫爬取书法开源网站获取大量丰子恺作品并保存到当前路径的文件夹中。

数据质量保证与控制。先进行重复的删除，由两个不同的搜索引擎发现的同一幅图像或者是由同一个图像的两个副本在不同网站的不同互联网位置发现，将删除完全重复的图像。接近重复的图像也会被删除

爬虫代码部分示例



图3 代码示例

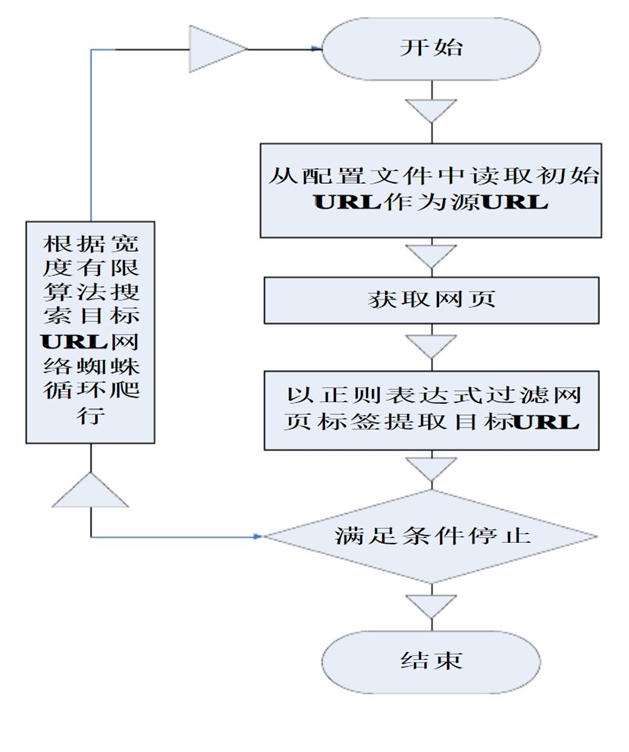


图4 爬虫工作流程图

获取成果



图5 成果展示

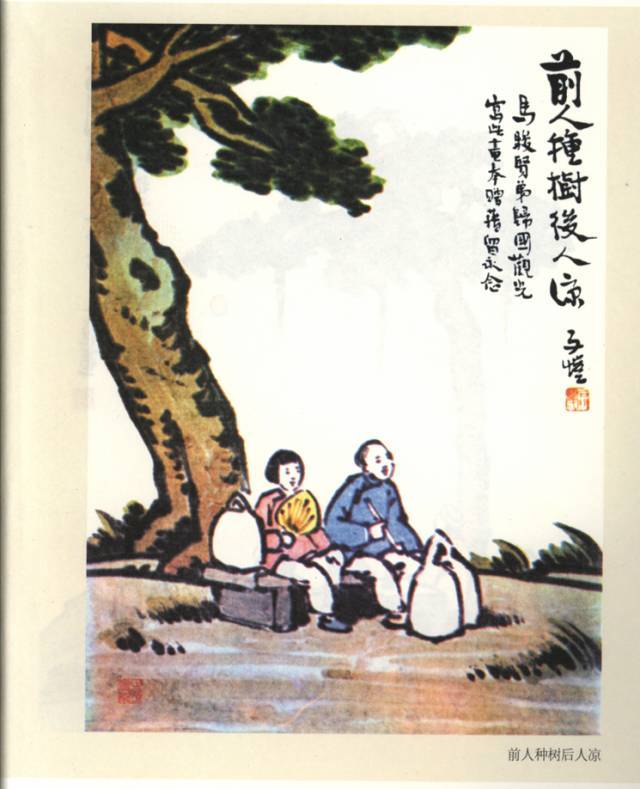


图6 成果展示



图7 成果展示

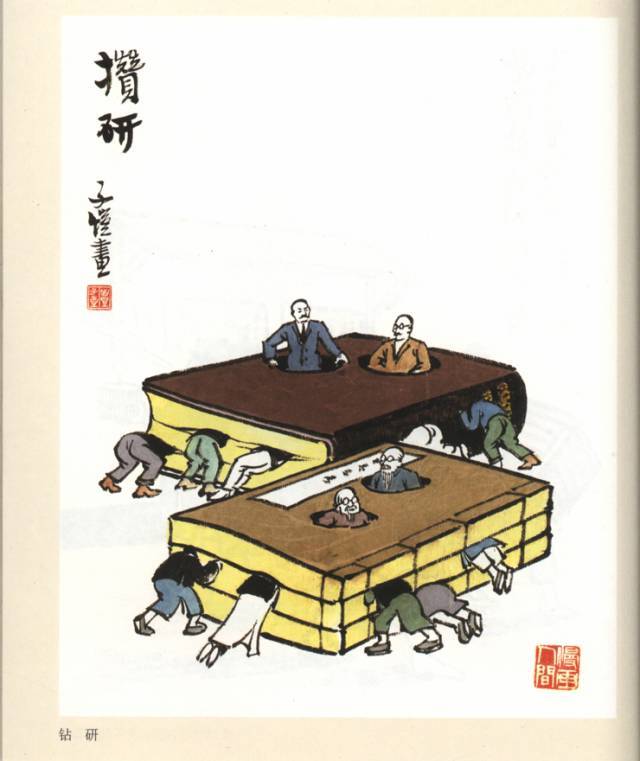


图8 成果展示



图9 成果展示



图10 成果展示



图11 成果展示



图12 成果展示



图13 成果展示



图14 成果展示