（1）Nutch（http://lucene.apache.org/nutch）, Nutch是apache的一个子项目，包括大规模爬虫工具，能够抓取和分辨web网站数据。

用在Linux中的较多，可做分布式爬虫。

（2）jsoup（http://jsoup.org/ ），jsoup 是一款Java 的HTML解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据。

（3）heritrix（http://sourceforge.net/projects/archive-crawler/files/），Heritrix 是一个由 java 开发的、开源的网络爬虫，用户可以使用它来从网上抓取想要的资源。其最出色之处在于它良好的可扩展性，方便用户实现自己的抓取逻辑。