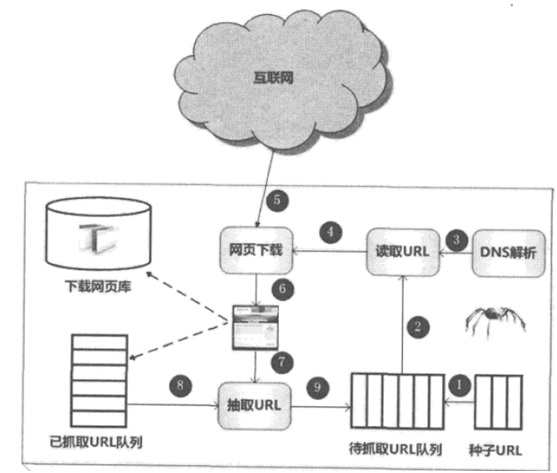
**搜索引擎爬虫工作原理**

搜索引擎的处理对象是互联网网页，日前网页数量以百亿计，所以搜索引擎首先面临的问题就是：如何能够设计出高效的下载系统，以将如此海量的网页数据传送到本地，在本地形成互联网网页的镜像备份。

网络爬虫即起此作用，它是搜索引擎系统中很关键也根基础的构件。这里主要介绍与网络爬虫相关的技术，尽管爬虫技术经过几十年的发展，从整体框架上已相对成熟，但随着联网的不断发展，也面临着一些有挑战性的新问题。

[版纳论坛](http://www.ban-na.com/forum.php?mod=forumdisplay&fid=2)下图所示是一个通用的爬虫框架流程。首先从互联网页面中精心选择一部分网页，以这些网页的链接地址作为种子URL，将这些种子URL放入待抓取URL队列中，爬虫从待抓取URL队列依次读取，并将URL通过DNS解析，把链接地址转换为网站服务器对应的IP地址。

然后将其和网页相对路径名称交给网页下载器，网页下载器负责页面内容的下载。对于下载到本地的网页，一方面将其存储到页面库中，等待建立索引等后续处理;另一方面将下载网页的URL放入已抓取URL队列中，这个队列记载了爬虫系统已经下载过的网页URL，以避免网页的重复抓取。对于刚下载的网页，从中抽取出所包含的所有链接信息，并在已抓取URL队列中检查，如果发现链接还没有被抓取过，则将这个URL放入待抓取URL队列末尾，在之后的抓取调度中会下载这个URL对应的网页。如此这般，形成循环，直到待抓取URL队列为审，这代表着爬虫系统已将能够抓取的网页尽数抓完，此时完成了一轮完整的抓取过程。



对于爬虫来说，往往还需要进行网页去重及网页反作弊。

上述是一个通用爬虫的整体流程，如果从更加宏观的角度考虑，处于动态抓取过程中的爬虫和互联网所有网页之间的关系，可以大致像如图2-2所身那样，将互联网页面划分为5个部分：

**1.已下载网页集合：**爬虫已经从互联网下载到本地进行索引的网页集合。

**2.已过期网页集合：**由于网页数最巨大，爬虫完整抓取一轮需要较长时间，在抓取过程中，很多已经下载的网页可能过期。之所以如此，是因为互联网网页处于不断的动态变化过程中，所以易产生本地网页内容和真实互联网网页不一致的情况。

**3.待下载网页集合：**即处于上图中待抓取URL队列中的网页，这些网页即将被爬虫下载。

**4.可知网页集合：**这些网页还没有被爬虫下载，也没有出现在待抓取URL队列中，不过通过已经抓取的网页或者在待抓取URL队列中的网页，总足能够通过链接关系发现它们，稍晚时候会被爬虫抓取并索引。

**5.不可知网页集合：**有些网页对于爬虫来说是无法抓取到的，这部分网页构成了不可知网页集合。事实上，这部分网页所占的比例很高。

根据不同的应用，爬虫系统在许多方面存在差异，大体而言，可以将爬虫划分为如下三种类型:

**1. 批量型爬虫（Batch Crawler）：**批量型爬虫有比较明确的抓取范围和目标，当爬虫达到这个设定的目标后，即停止抓取过程。至于具体目标可能各异，也许是设定抓取一定数量的网页即可，也许是设定抓取消耗的时间等。

**2.增量型爬虫（Incremental Crawler）：**[景洪论坛](http://www.ban-na.com/forum.php?mod=forumdisplay&fid=2)增量型爬虫与批量型爬虫不同，会保持持续不断的抓取，对于抓取到的网页，要定期更新，因为互联网的网页处于不断变化中，新增网页、网页被删除或者网页内容更改都很常见，而增量型爬虫需要及时反映这种变化，所以处于持续不断的抓取过程中，不是在抓取新网页，就是在更新已有网页。通用的商业搜索引擎爬虫基本都属此类。

**3.垂直型爬虫(Focused Crawter）：**垂直型爬虫关注特定主题内容或者属于特定行业的网页，比如对于健康网站来说，只需要从互联网页而里找到与健康相关的页面内容即可，其他行业的内容不在考虑范围。垂直型爬虫一个最大的特点和难点就是：如何识别网页内容是否属于指定行业或者主题。从节省系统资源的角度来说，不太可能把所有互联网页面下载下来之后再去筛选，这样浪费资源就太过分了，往往需要爬虫在抓取阶段就能够动态识别某个网址是否与主题相关，并尽量不去抓墩无关页面，以达到节省资源的目的。垂直搜索网站或者垂直行业网站往往需要此种类型的爬虫。