**网络爬虫技术的原理、分类、应用**

1网络爬虫的历史背景和定义[1]

网络爬虫（外文名Web Crawler），又被成为网络蜘蛛，蚂蚁，网页追随者，它从万维网上获取网页，并通过网页中的其他链接在万维网上采集信息资源，是按照一定的人为规则自动识别并抓取万维网信息的程序和脚本。网络爬虫通常从一些精挑细选的种子URL开始“爬行”，将这些URL放入待抓取的URL队列，从待抓取的URL队列中取出URL，解析DNS并且得到主机IP,并将URL对应的网页下载下来，存储进已下载网页库中，直到遍历Web，这种行为被成为网络爬行（Web Crawler）。

2网络爬虫的分类及原理介绍[2]

网络爬虫按照系统结构和实现技术，大体上可以分为如下几类：通用爬虫（General Purpose Web Crawler）、聚集网络爬虫（Focused Web Crawler）、深层网络爬虫（Deep Web Crawler）。在实际运用中往往是多种技术的综合运用。

2.1通用网络爬虫

通用爬虫是从一个或者若干个初始网页开始，爬取全网的网页，所以主要用来所谓各大网站的搜索引擎，但是由于商业原因，这些技术细节很少公开，例如Google、百度等。其主要组成部分大致有URL链接库、文档内容模块、文档解析模块、URL过滤模块，当通用爬虫工作时，它先从预定的若干精选的URL开始，获取初始页面上的URL列表，并且在爬取URL的同事下载该页面，再通过文档解析模块将需要的内容保存到对应的数据库中，并再获得当前页面的新的URL，保存到URL链接库中，直到满足停止条件。

2.2聚集网络爬虫

聚集网络爬虫（Focused Crawler），是一种定向抓取有关联的网页的资源的爬虫。聚集爬虫通过给定的目标，有目的性的访问万维网上相关网页，并且有选择性的抓取网页。聚集爬虫不像通用爬虫，它会先通过算法，对获得的URL进行分析并过滤掉和目标主题无关的链接，将有用的保留到URL队列中。接着它会从队列中通过定好的策略，找到合适的URL，重复上述过程，直到满足一定条件后停止爬行。因此，聚集爬虫相比通用爬虫，具有得到信息更精确，效率更高的特点，当然，聚集爬虫也有其缺陷，如下几个问题：

1. 如何对目标网页进行描述；（2）如何对网页和数据进行分析和过滤；（3）对URL搜索和排序策略。

2.3深层网络爬虫（Deep Web Crawler）

网络数据的快速增加，大量数据因此积累而被埋藏，而传统的搜索方式无法搜索到这部分有价值的页面，所以这部分内容对于大部分来说是隐藏的和不可见的，称之为Deep Web。深层网络爬虫与一般爬虫的区别是，深层网络爬虫在下载完成页面后，没有直接将所有URL进行遍历，而是通过特定的算法策略将其进行分类处理，对于不同的类型URL使用不同的算法进行查询参数，并将参数二次提交给服务器。如果提交的查询参数正确，才会获得到隐藏的页面和URL。

3网络爬虫策略[3]

1. IP地址搜索策略

先给爬虫一个起始IP地址，根据该IP地址以递增的方式搜索本IP地址段后的每个地址中的文档。

1. 深度优先搜索策略

目的是达到被搜索结构的叶节点（就是那些不包含任何超链接的HTML文件）。

1. 宽度优先搜索策略

先搜索完一个Web页面中所有的超链接，然后再继续搜索下一层，直到层底为止。

4网络爬虫的应用[4]

随着信息技术的高速发展，信息化社会逐步走向大数据时代，大数据的最显著的特点就是数据量大，数据形式复杂，数据处理速度快和数据价值高这四个方面，因为万维网成为了一个巨大的数据库，所以如何从海量数据库中采集和提取数据是我们面临的挑战。而网络爬虫的出现为解决这个难题做出了巨大的贡献。网络爬虫在Web信息搜索与数据挖掘中的应用，极大的提高了Web信息的采掘质量，为整个搜索与挖掘系统鉴定了坚实的基础；在档案信息采集中的应用，为从海量数据中采集和提取与档案资源建设相关信息提供巨大方便；在社交网络中，为研究用户偏好，然后为每个用户制定个性化方案及趋向预测提供巨大帮助。

参考文献：

[1]于成龙[1],于洪波[1],网络爬虫技术研究.东莞理工学院学报,2011,18(3)

[2]蔡笑伦[1],网络爬虫技术的发展趋势.科技信息,2010,(12)

[3]于成龙[1],于洪波[1],网络爬虫技术研究.东莞理工学院学报,2011,18(3)

[4]于成龙[1],于洪波[1],网络爬虫技术研究.东莞理工学院学报,2011,18(3)