**第10版爬虫使用说明**

**一、各脚本的作用**

**A. Prepare脚本**

**下载‘网址排行’页面，并用thisLink3.0提取里面的网站链接**

脚本中整合了以下步骤：

1. 调用get\_web\_list下载‘网址排行’页面，
2. 调用thisLink3.0提取里面的网站链接，作为种子网页

**B. Main脚本**

**总控函数，控制随后的流程：下载种子网页，从种子网页中提取链接，根据链接去下载更多网页。**

脚本中整合了以下步骤：

1. . 用get\_start\_site分片下载种子网站的网页

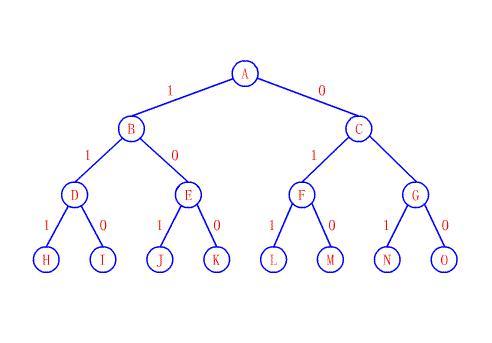
4. 调用thisLink2.1: 爬取网站上的超链接，并保存到URL文件中~

5. 调用File\_web2.1或3.1：去遍历URL文件，开始新一轮网页的下载

6. 第4、5步循环往复，直到片区内所有的网页都下载完为止。

**二、爬虫的访问顺序**

网站的层级：我们先遍历第1层（种子网页），再遍历第2层，然后逐层访问。



1. **爬虫算法**

**从URL文件中读取链接并下载新的网页**

**从网页中提取链接到URL文件**

**下载种子网页**

**爬取算法：**

Prepare:

先用get\_web\_list下载‘网址排行’页面，然后再用thisLink3.0提取里面的网站链接，

这些是种子网站，最后用get\_start\_site 分片下载这些种子网站。

Depth = <你预先设定的值>

//以上的depth是指遍历网站的深度，可以预先设定，当达到设定的深度值时，爬虫不再向下爬取。

对于一个片区来说，

While(还有网页没有被下载或提取链接）{

1. 调用thisLink2.1: 爬取网站上的超链接，并保存到URL文件中~
2. 调用File\_web2.1：去遍历URL文件，根据里面的网址下载新的网页
3. 同时depth -1

}