**网页爬虫抓取策略**

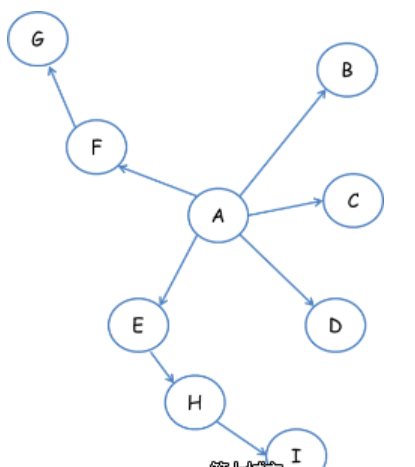
在爬虫系统中，待抓取URL队列是很重要的一部分。待抓取URL队列中的URL以什么样的顺序排列也是一个很重要的问题，因为这涉及到先抓取那个页面，后抓取哪个页面。而决定这些URL排列顺序的方法，叫做抓取策略。

**几种常见的抓取策略：**

**1.深度优先遍历策略：**

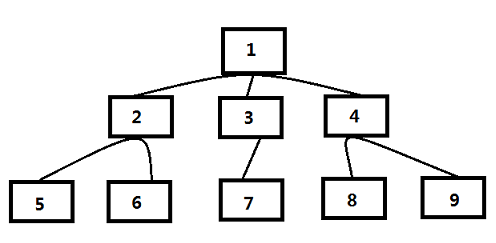
深度优先遍历策略是指网络爬虫会从起始页开始，一个链接一个链接跟踪下去，处理完这条线路之后再转入下一个起始页，继续跟踪链接。

遍历的路径：A-F-G E-H-I B C D



## **2.宽度优先策略（Breath First）**

基本思想：将新下载网页包含的链接直接追加到待抓取URL队列末尾。



上图即为此策略示意图：   
 假设队头的网页是1号网页，从1号网页中抽取出3个链接指向2号、3号和4号网页，于是按照编号顺序依次放入待抓取URL队列，图中网页的编号就是在待抓取URL队列中的顺序编号，之后爬虫以此顺序进行下载。

实验表明，这种策略效果很好，虽然看似机械，但实际上的网页抓取顺序基本上是按照网页的重要性排序。之所以如此，有研究人员认为：如果某个网页包含很多入链，那么更有可能被宽度优先遍历策略早早抓到，入链个数从侧面体现了网页的重要性，即实际上宽度优先遍历策略隐含了一些网页优先级假设。

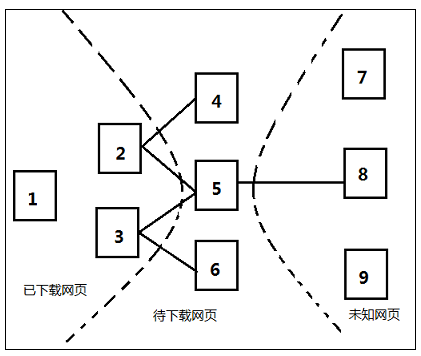
1. **反向链接数策略：**

反向链接数是指一个网页被其他网页链接指向的数量。反向链接数表示的是一个网页的内容受到其他人的推荐的程度。因此，很多时候搜索引擎的抓取系统会使用这个指标来评价网页的重要程度，从而决定不同网页的抓取先后顺序。

在真实的网络环境中，由于广告链接、作弊链接的存在，反向链接数不能完全等他我那个也的重要程度。因此，搜索引擎往往考虑一些可靠的反向链接数。

**4.Partial PageRank策略：**

Partial PageRank算法借鉴了PageRank算法的思想：对于已经下载的网页，连同待抓取URL队列中的URL，形成网页集合，计算每个页面的PageRank值，计算完之后，将待抓取URL队列中的URL按照PageRank值的大小排列，并按照该顺序抓取页面。

  如果每次抓取一个页面，就重新计算PageRank值，一种折中方案是：每抓取K个页面后，重新计算一次PageRank值。但是这种情况还会有一个问题：对于已经下载下来的页面中分析出的链接，也就是我们之前提到的未知网页那一部分，暂时是没有PageRank值的。为了解决这个问题，会给这些页面一个临时的PageRank值：将这个网页所有入链传递进来的PageRank值进行汇总，这样就形成了该未知页面的PageRank值，从而参与排序。   


上图即为非完全PageRank策略示意图：   
设定每下载3个网页进行新的PageRank计算，此时已经有{1,2,3}3个网页下载到本地。这三个网页包含的链接指向{4,5,6}，即待抓取URL队列，如何决定下载顺序？   
将这6个网页形成新的集合，对这个集合计算PageRank的值，这样4、5、6就获得自己对应的PageRank值，由大到小排序，即可得出下载顺序。假设顺序为5,4,6，当下载55页面后抽取出链接，指向页面8，此时赋予8临时PageRank值，如果这个值大于4和6的PageRank，则接下来优先下载页面8，如此不断循环，即形成了非完全PageRank策略的计算思路。

**5.OPIC策略策略：**

该算法实际上也是对页面进行一个重要性打分。在算法开始前，给所有页面一个相同的初始现金（cash）。当下载了某个页面P之后，将P的现金分摊给所有从P中分析出的链接，并且将P的现金清空。对于待抓取URL队列中的所有页面按照现金数进行排序。

## 6.OCIP策略（Online Page Importance Computation）

OCIP字面意思即“在线页面重要性计算”，可以将其看做是一种改进的PageRank算法。

基本思路：   
在算法开始之前，每个互联网页面都给予相同的“现金”，每当下载了某个页面P后，P就将自己拥有的现金平均分配给页面中包含的链接页面，把自己的“现金”清空。而对于待爬取URL队列中的网页，则根据其手头拥有的“现金”金额多少排序，优先下载“现金”最多的网页，OPIC从大的框架上与PageRank思路基本一致。

**与PageRank的区别**在于：PageRank每次需要迭代计算，而OPIC策略不需要迭代过程。所以计算速度远远快与PageRank，适合实时计算使用。同时，PageRank在计算时，存在向无链接关系网页的远程跳转过程，而OPIC没有这一计算因子。实验结果表明，OPIC是较好的重要性衡量策略，效果略优于宽度优化遍历策略。

2.3.6大站优先策略

  对于待抓取URL队列中的所有网页，根据所属的网站进行分类。对于待下载页面数多的网站，优先下载。这个策略也因此叫做大站优先策略。