

宽度优先搜索的 Web 网页爬行方法

编者：二十一@兰州SEO(WWW.0931seo.cn)

原文地址：http://www.0931seo.cn/seo-book/KuanDuYouXianSouSuoDe_WebWangYePaHangFangFa/

网页爬行器首先从一个由起始的 URL 构成的队列出发，这些 URL 被称为种子，它把队列中的第 1 个 URL 移出队列，然后取得该 URL 所对应的网页 P₀，然后从网页 P₀ 中提取它所包含的所有的 URL，把这些 URL 按照某种策略加进网页爬行器需要爬行的 URL 队列中，网页爬行器再从 URL 队列中取下一个需要爬行的 URL，重复如上所述过程，直到满足要求或 URL 队列为空。可以把网页爬行器爬行过的每个网页看成有向图中的一个节点，网页之间的链接看成是节点之间的有向边，则网页构成的有向图如图 1 所示，网页爬行器在 Web 空间中爬行网页的过程就是对由网页构成的有向图的遍历。网页爬行器爬行网页的策略主要有两种，一种是深度优先搜索策略。另一种是宽度优先搜索策略，Marc Najork 等人的研究证明，爬行器采用宽度优先搜索策略爬行的网页质量要比采用深度优先搜索策略的要好[4~6]，因此，大多数网页爬行器采用宽度优先搜索策略或者是对这种策略的某些改进。其基本算法描述如下。

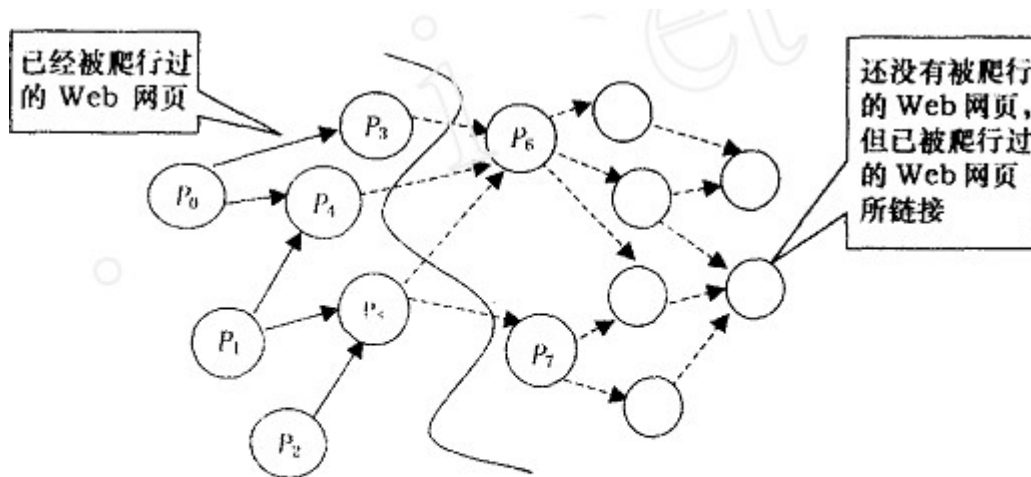


图 1 网页构成的有向图

Fig. 1 The directed graph connected by Web pages

Breadth2 first Crawling Algorithm ()

```
{
·把种子 URLs 加入到爬行器待爬行的队列 URLs - QUEUE 中。
· While (当队列 URLs - QUEUE 不为空和没有满足某种终止条件 3
)
{
·从队列 URLs - QUEUE 中移出一个 URL。
·取得 URL 所对应的网页 P。
·对网页 P 进行存储、索引并解析，取得网页 P 包含的所有 URLs。
·把取得的 URLs 加入到队列 URLs - QUEUE 中。
}
}
```

此处的“没有满足某种终止条件”是指爬行器的爬行过程到目前为止没有满足系统的要求，如爬行的网页数量不够。如果一个爬行器按照如上所述的宽度优先搜索策略在 Web 空间中爬行，它对所有的网页都采取一视同仁的态度，在爬行的过程中，没有考虑网页之间的超链接信息和网页内容，这样盲目爬行的结果就导致了它所爬行回来的网页质量不高。

转载请表明作者和出处：

二十一@兰州seo

<http://www.0931seo.cn>