职达

第九届中国大学生服务外包创新创业大赛

队伍成员：

曾永家右　黄泰北　赵辉

龙兰心　罗雪蕾

基于大数据的人岗匹配系统“职达”

爬虫算法

职达信息技术有限公司

达芬奇队

指导教师：徐进 尹帮旭

目录

一、 图的广度优先遍历

2.1 广度优先搜索介绍

广度优先搜索算法(Breadth First Search)，又称为"宽度优先搜索"或"横向优先搜索"，简称 BFS。

它的思想是：从图中某顶点 v 出发，在访问了 v 之后依次访问 v 的各个未曾访问过的邻接点，然后分别从这些邻接点出发依次访问它们的邻接点，并使得“先被访问的顶点的邻接点先于后被访问的顶点的邻接点被访问，直至图中所有已被访问的顶点的邻接点都被访问到。如果此时图中尚有顶点未被访问，则需要另选一个未曾被访问过的顶点作为新的起始点，重复上述过程，直至图中所有顶点都被访问到为止。

换句话说，广度优先搜索遍历图的过程是以 v 为起点，由近至远，依次访问和 v 有路径相通且路径长度为 1,2...的顶点。

2.2 有向图的广度优先遍历

下面以图 2-1 为例，进行说明

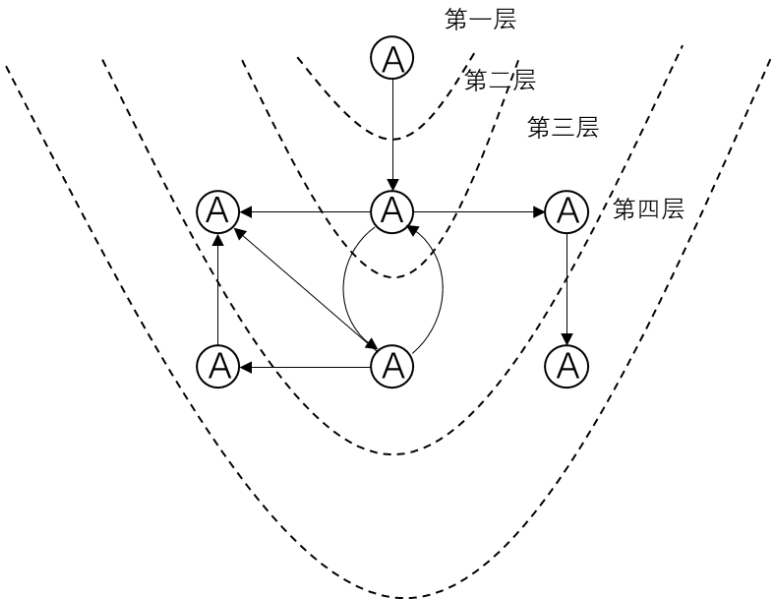

图 2-1

第 1 步：访问 A。

第 2 步：访问 B。

第 3 步：依次访问 C,E,F。

在访问了 B 之后，接下来访问 B 的出边的另一个顶点，即 C,E,F。前面已经说过，在本文实现中，顶点 ABCDEFG 按照顺序存储的，因此会先访问 C，再依次访问 E,F。

第 4 步：依次访问 D,G。

在访问完 C,E,F 之后，再依次访问它们的出边的另一个顶点。还是按照 C,E,F 的顺序访问，C 的已经全部访问过了，那么就只剩下 E,F；先访问 E 的邻接点 D，再访问 F 的邻接点 G。

因此访问顺序是：A -> B -> C -> E -> F -> D -> G

二、爬虫算法

2.1 互联网的抽象

我们将整个互联网抽象成一张有向连接图，网页抽象为定点，定点之间的连接即为 url 之间的相互跳转。

2.2 去重策略

一个原始的想法就是将所有已访问过的所有 url 进行顺序存储。你可以把全部已经下载完成的 url 存储到数据库中。每次有一个爬虫线程得到一个任务 url 开始下载之前，到数据库中检索此 url，如果没有出现过，则将这个此 url 写入数据库。

但是这种去重的的思想是非常直观的。但是占用存储空间不说，查找效率超级低下，因此这个方案行不通。

对 url 进行 hash 运算映射到某个地址，将该 url 和 hash 值当做键值对存放到 hash 表中，只需要对需要检测的 URL 的 hash 的映射进行比对，从而就可以对 url 是否存在进行判断。因此，原来的 url 库就可以简化为 hash 库，这要比 url 简便很多，但是需要考虑 hash 碰撞的问题，在设计中需要对 hash 函数进行考虑，避免因考虑不周造成 hash 碰撞。

所以我们采用使用 url 的 MD5 码去重的方法，MD5 码基于 hash 算法，MD5 算法能够将任何字符串压缩为 128 位整数，并映射为物理地址，同时，与传统的 Hash 去重方法相比，MD5 进行 Hash 映射碰撞概率很低。MD5 经过时间验证，是一种比较好的去重方法。

2.3 爬虫算法

1、 给定初始 url_start，初始化待爬取 url 队列 Todo，初始化已爬取 url 列表 Visited，初始化已爬取 url 的 MD5 码列表 Visited_MD5

2、 将 url_start 加入 Todo

3、 如果 Todo 不为空

Todo 进行出队操作，出队的 url 记作 url_todo

计算 url_to 的 MD5 码，判断其是否在 Visited_MD5 里

如果是：转入 3

如果否：对 url_todo 进行爬取，将应网页中所有的 url 加入到 Todo，将 url_to 加入 Visited，并将 url_todo 的 MD5 码加入 Visited_MD5

4、如果 Todo 为空，爬虫结束。

三、参考文献

[1]成功,李小正,赵全军.一种网络爬虫系统中 URL 去重方法的研究[J].中国新技术新产品,2014(12):23.

[2]王桦. 基于广度优先的主题爬虫的设计与实现[D].复旦大学,2011.

[3]吴小惠.分布式网络爬虫 URL 去重策略的改进[J].平顶山学院学报,2009,24(05):116-119.