**pagerank 基本思想**

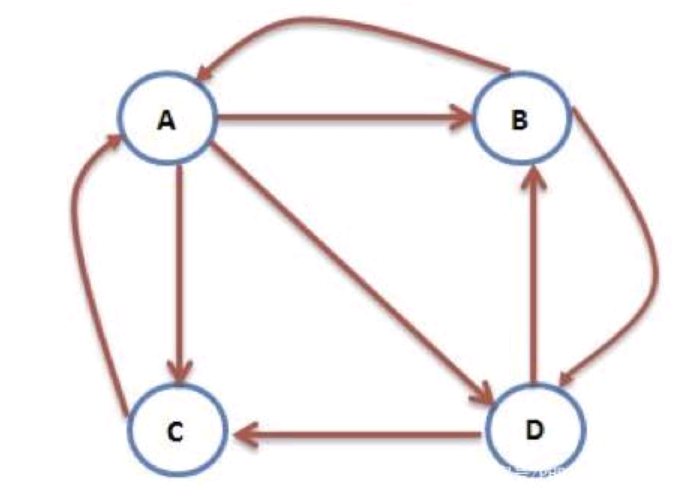
数量假设：一个页面越被其他页面链接，说明他越重要

质量假设：越是被高质量页面链接，说明该页面越重要

**pagerank 模型**

pagerank 模型模拟的是一个用户在互联网上浏览到每个网页的概率。

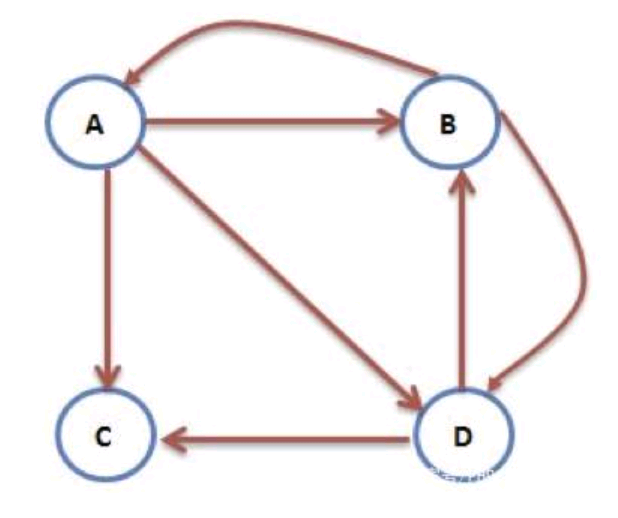
下面来一个简单计算，更好的理解下



pg（A） = pg（C）/1 + pg(B) /2

A网页的pagerank值由网页B和网页C的pagerank贡献而来，因为B网页有两个外链，假设等概率贡献，则贡献给A的值为自身的一半。

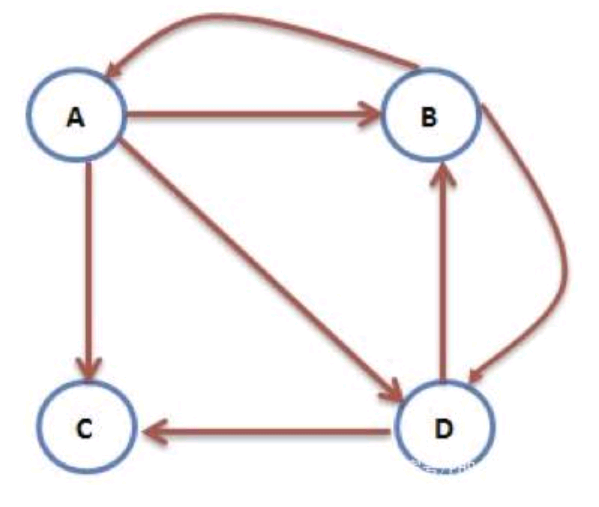
再看一个例子



此处C因为没有外链，所以我们假设他给其他所有网页都贡献了pg，于是：

pg(A) = pg（C）/4 + pg(B) /2

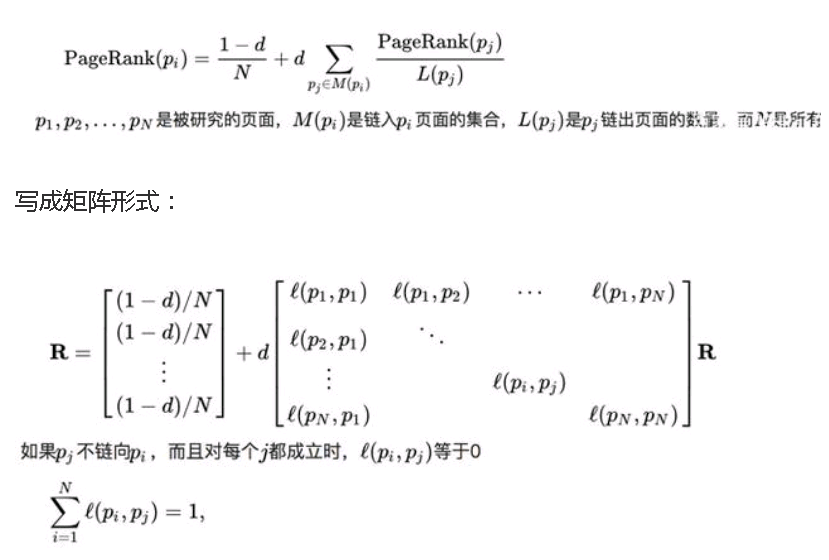
在考虑一个情况，即用户在浏览网页的过程中，直接输入新网址进行浏览，即一个网页都有可能跳转到任意其他网页，于是针对下图：



pg(A) = （a \* pg(B) /2） + (1-a) / 4) 上面公式的含义是：

在任意时刻，用户到达某页面后并继续向后浏览的概率为a，则用户停止浏览的概率为(1-a)，此时用户停止浏览后，可能会直接通过输入浏览器地址进行浏览网页，此时跳转到任意网址的概率都一样，于是上面的 a \* pg(B) /2 表示从b跳转过来的概率，(1-a) / 4 表示直接输入网址跳转过来的概率。

下面对上面公式进行推广：



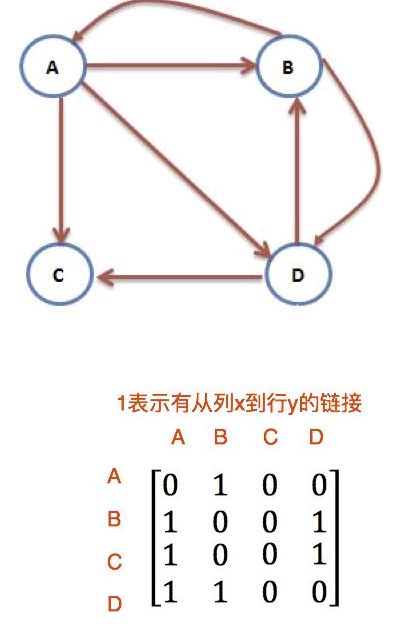
上面式子中l(pi,pj)表示网页j指向网页i,其值为:

l(pi,pj) = 1 / L(pj)

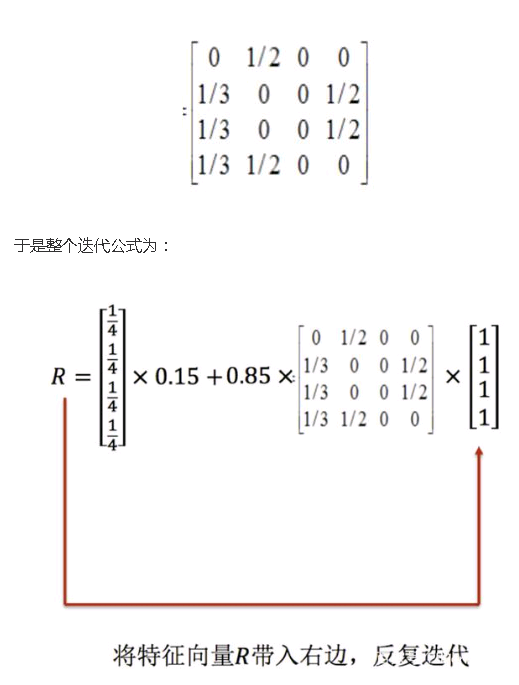
其中L(pj)表示pj外链的数量。

**计算示例**

下面我们来看一个具体的例子，假设有如下图：



根据上面的矩阵，我们可以得到下面的矩阵：



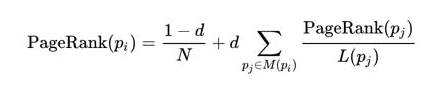
注意：pagerank如果只考虑从其他网页转移过来，会出现下面的两个问题：

终止点问题：只入不出，则所有节点的概率最后会变为0陷阱问题：网页只有指向自己的链接，则最后所有的概率都会转移到这些网站

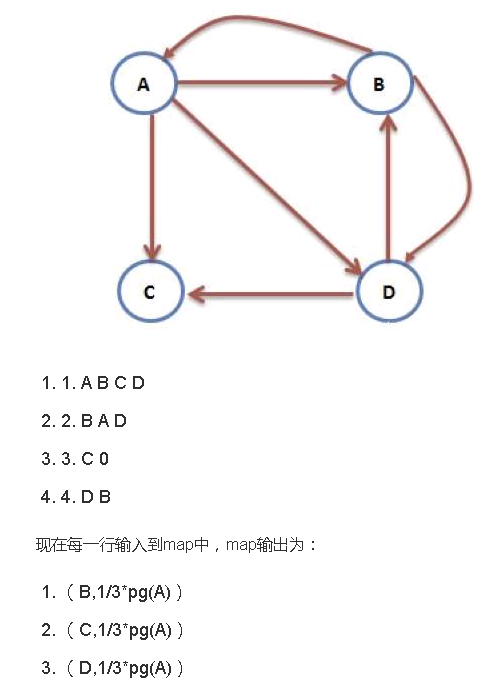
更具体的解释可以看PageRank算法简介及Map-Reduce实现

**map reduce 实现**

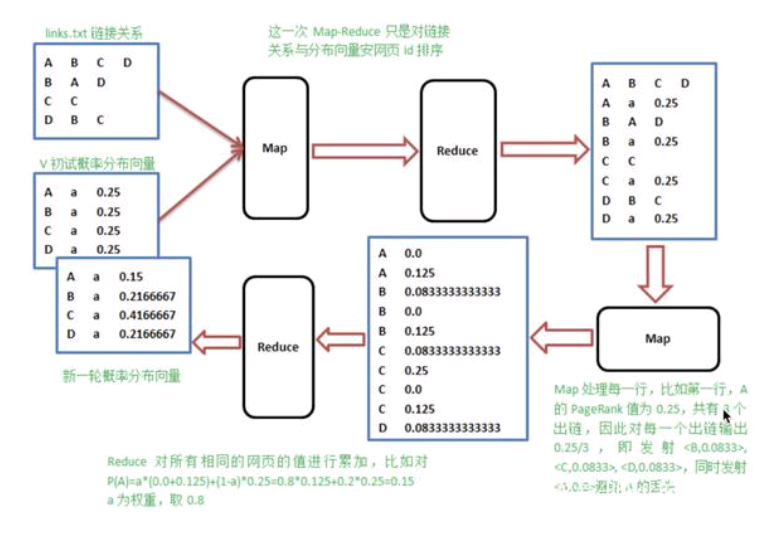
mapreduce关键在于设计map和reduce，先来看pr值的计算公式：



所以每个pr值都是由其他网页pr值贡献得到，因此我们还是以下图为例子：



此处我们发现缺少的为每个网页的当前pg值，另外每个网页的pg值是单独存放的，因此我们需要先来一个map-reduce，将pg值映射到每个对应的行，具体流程如下图：



参考资料：

<https://www.letiantian.me/2014-06-10-pagerank/>浅入浅出：PageRank算法

**深入探讨PageRank（一）：PageRank算法原理入门**

**深入探讨PageRank（二）：PageRank原理剖析**

**pagerank 算法 快速入门**