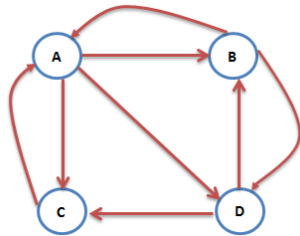


PageRank

一个页面的“得票数”由所有链向它的页面的重要性来决定，到一个页面的超链接相当于对该页投一票。一个页面的 **PageRank** 是由所有链向它的页面（“链入页面”）的重要性经过递归算法得到的。

为了处理那些“没有向外链接的页面”（这些页面就像“黑洞”会吞噬掉用户继续向下浏览的概率）带来的问题，（这里的被称为阻尼系数（**damping factor**）），其意义是，在任意时刻，用户到达某页面后并继续向后浏览的概率，该数值是根据上网者使用浏览器书签的平均频率估算而得。（就是用户停止点击，随机跳到新 URL 的概率）的算法被用到了所有页面上。

互联网中的众多网页可以看作一个有向图。假设一个由 4 个网页组成的群体：A、B、C 和 D。



$$\text{PageRank}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{\text{PageRank}(p_j)}{L(p_j)}$$

PR(pi) pi 页面的 PageRank 值

p1,p2,...pN 是被研究的页面

M(pi) 链入 pi 页面的集合

L(pj) pj 链出页面的数量

N 是所有页面的数量

d:阻尼系数 damping factor 0.85

根据上面的公式，我们可以计算每个网页的 PR 值，在不断迭代趋于平稳的时候，即为最终结果。