# PageRank算法

## pagerank基本概念

### 1.1什么是pagerank

在实际应用中许多数据都以图（graph）的形式存在，比如，互联网、社交网络都可以看作是一个图。图数据上的机器学习具有理论与应用上的重要意义。 PageRank 算 法是图的链接分析（link analysis）的代表性算法，属于图数据上的无监督学习方法。

PageRank算法最初作为互联网网页重要度的计算方法，1996 年由Page和Brin提出，并用于谷歌搜索引擎的网页排序。事实上，PageRank 可以定义在任意有向图上，后来被应用到社会影响力分析、文本摘要等多个问题。

PageRank算法的基本想法是在有向图上定义一个随机游走模型，即一阶马尔可夫链，描述随机游走者沿着有向图随机访问各个结点的行为。在一定条件下，极限情况访问每个结点的概率收敛到平稳分布，这时各个结点的平稳概率值就是其PageRank值，表示结点的重要度。PageRank 是递归定义的，PageRank 的计算可以通过迭代算法进行。

### 1.2 pagerank基本思想

被用户访问越多的网页更可能质量越高，而用户在浏览网页时主要通过超链接进行页面跳转，因此需要通过分析超链接组成的拓扑机构来推算每个网页被访问频率的高低。

## pagerank算法

### 2.1公式定义



* PR(a)表示当前节点a的PR值
* PR(Ti)表示其他各个节点（能够指向a）的PR值
* L(Ti)表示其他各个节点（能够指向a）的出链数
* i代表当前时刻或迭代次数

### 2.2计算演示

接下来以下图为例进行计算演示：

A

B

D

C

1. 将四个节点的初试PR值都设置为1/4
2. 根据每一个节点（a）的入链节点（Ti）的PR值及出链数和自身（a）的PR值
3. 不断进行迭代，直到PR值不再发生变化

**算法基础讲解：**

A有两个入链节点C（出链数为2，PR=1/4）和D（出链数为1，PR=1/4）由计算公式得到i=1时刻A节点的PR值。其余节点计算方法类似，经过计算可得到如下图表。



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 循环次数i  PR值 | PR(A) | PR(B) | PR(C) | PR(D) |
| I=0 PR值初始化=1/N | 1/4 | 1/4 | 1/4 | 1/4 |
| I=1 | 3/8 | 1/8 | 3/8 | 1/8 |
| 排名 | 1 | 2 | 1 | 2 |

**通过矩阵化表达快速计算第一次迭代PR值：**

**PR(a) = M\*V**

** \*  = **

**第一次迭代PR = 0\*1/4 + 0\*1/4 + 1/2\*1/4 + 1\*1/4 = 3/8**

由此可以得到每一步PR值迭代的结果为：MV，MMV，MMMV，最终会收敛为

## 存在问题

### Dead Ends

A

**B**

如上图所示：B没有任何出链，这就是Dead Ends，Dead Ends会导致网站权重变成0

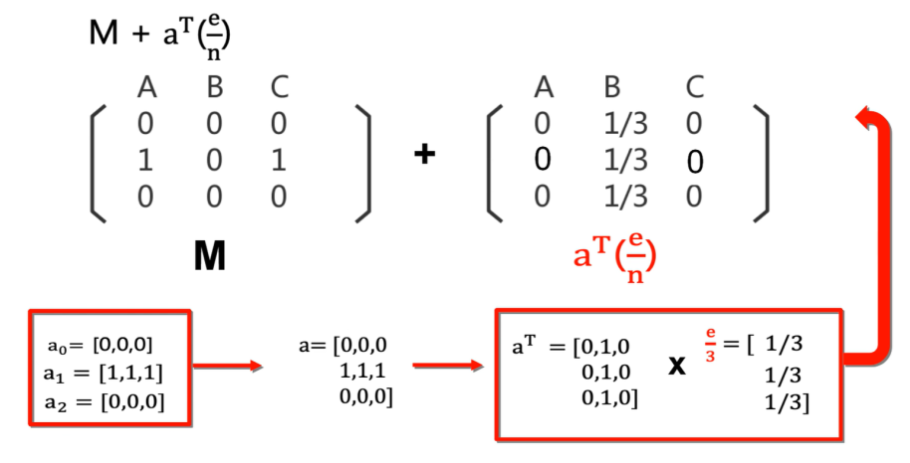
最朴素的想法是：对全是0的列上的每一个元素加上1/n（n为节点个数）对M进行修正

#### 3.1.1 Dead Ends问题修正公式：

****

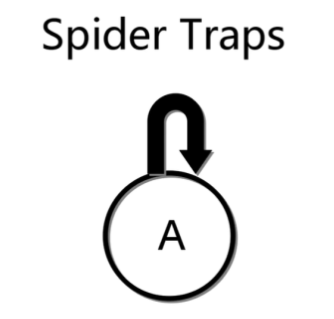
* **a=[a0,a1,...,an]，当有一列全为0时（即该节点无出链），ai=1，其他时候ai=0**
* **e为由1填满的列矩阵**
* **n为M矩阵的行数/列数**
* **V为PR值的矩阵**

#### 3.1.2 矩阵计算



### Spider Traps

#### 3.2.1 基本概念

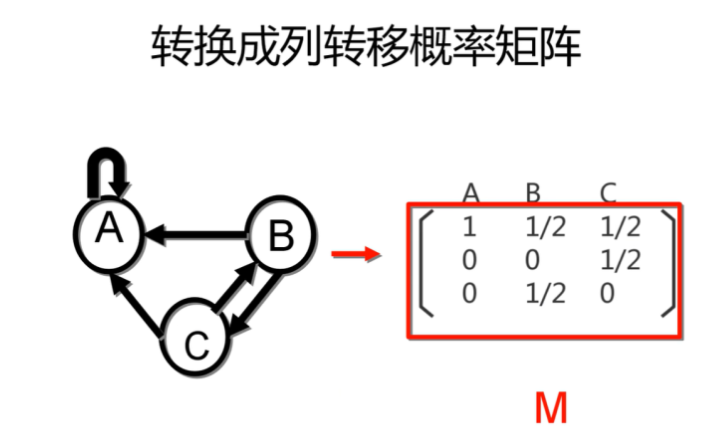


如上图所示，A节点与其他节点之间没有出链，这就是Spider Traps，这将导致网站权重变向一个节点偏移（经过多轮迭代后，A的权重越来越大，趋近于1，而其余节点的权重趋近于0）

#### 3.2.2 解决方案

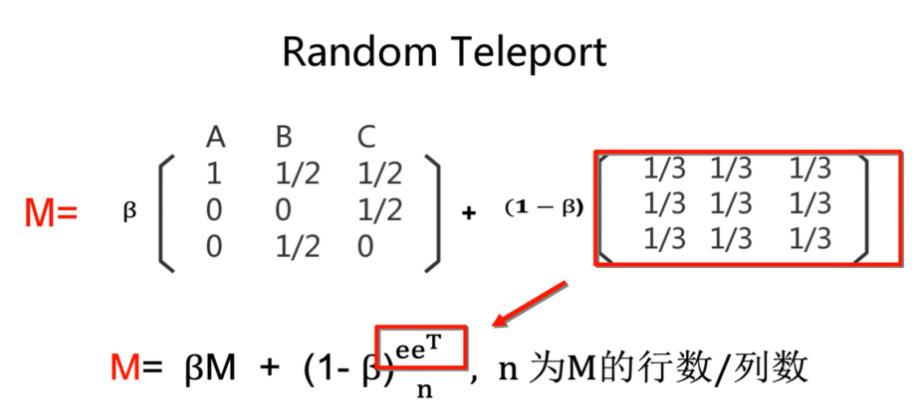
步骤一：将节点图转换为转移概率矩阵

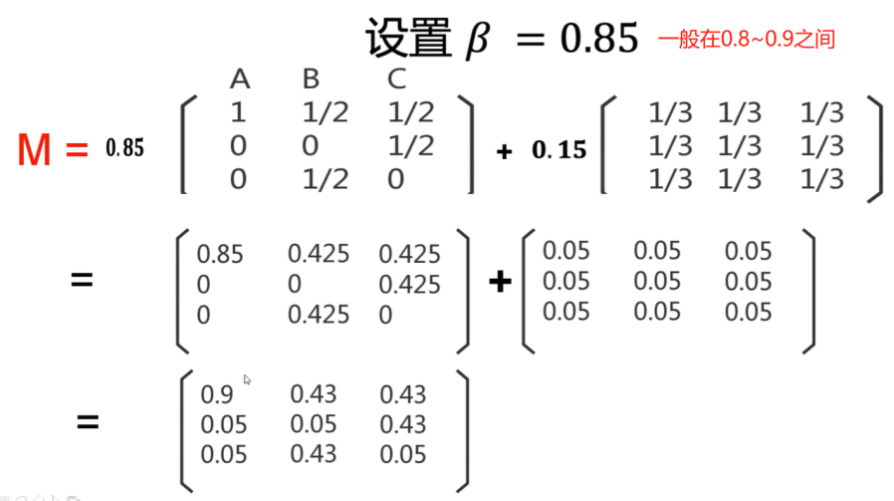
步骤二：修正M

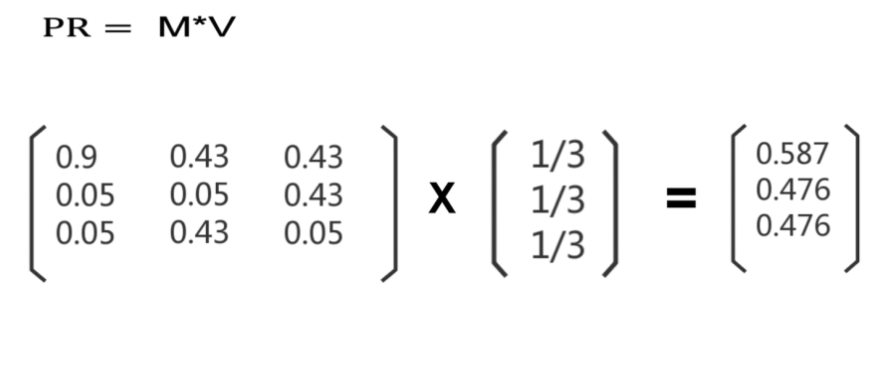


**修正M**

****





**Spider Traps问题修正公式**

****

* β表示跟随出链（out-links）打开网页的概率
* 1-β表示随机跳到其他网页的概率，浏览a的时候有一定概率会打开b或c
* eeT表示由1填满nxn矩阵
* V表示PR值的矩阵

**PageRank最终修正公式**

****

## 优缺点总结

### 4.1 优点

* 通过网页之间的链接来决定网页重要性，一定程度消除了人为对排名结果的影响
* 离线计算PageRank值，而非查找的时候计算，提升了查询的效率

### 4.2 缺点

* 存在时间久的网站，PageRank值会越来越大，而新生的网站，PageRank值增长慢
* 非查询相关的特性，查询结果会偏离搜索内容
* 通过僵尸网站或链接，认为刷PageRank值