深度之眼
deepshare.net

第21章 PageRank算法

- 21.1 PageRank算法的定义与幂法计算

导师：Irene

本节学习内容

Learning content in this section

对应书本/课程章节	主要学习内容	学习目标
PageRank算法定义	互联网搜索排序，有向图	理解PageRank算法含义
随机游走模型	马尔科夫链状态转移矩阵	掌握利用有向图计算PageRank值的方法
PageRank一般定义	平稳分布，阻尼因子，PageRank值	掌握PageRank值的表达式
迭代算法与例题	迭代求解PageRank值	掌握PageRank迭代算法步骤
幂法与代数算法	Perron-Frobenius定理，幂法近似计算	掌握幂法近似计算PageRank值的过程

PageRank算法引入

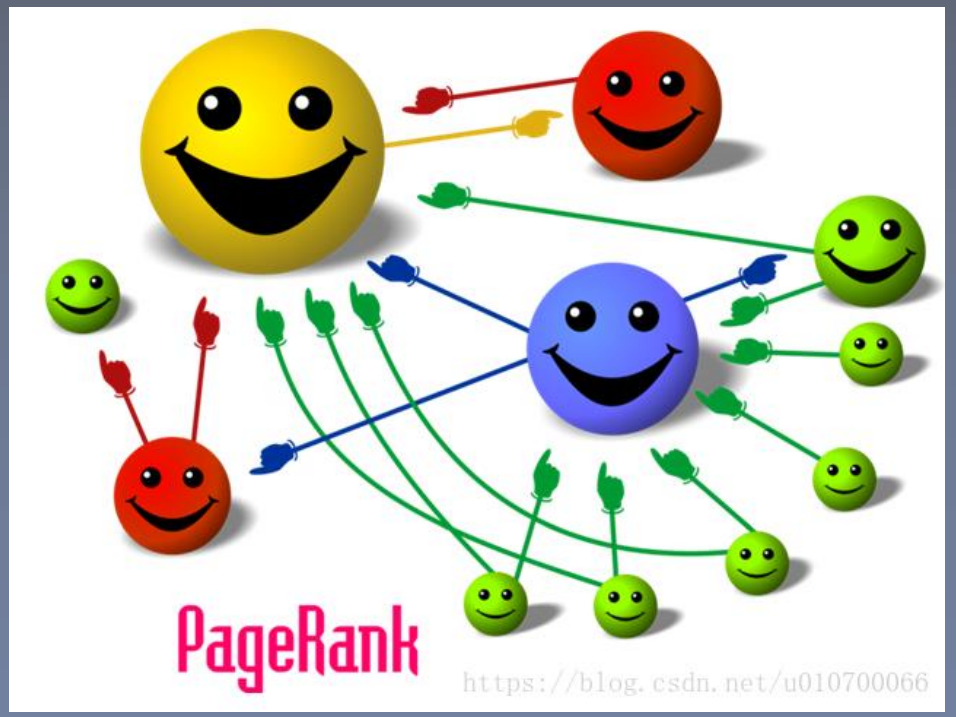
Introduction of PageRank Algorithm

- 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

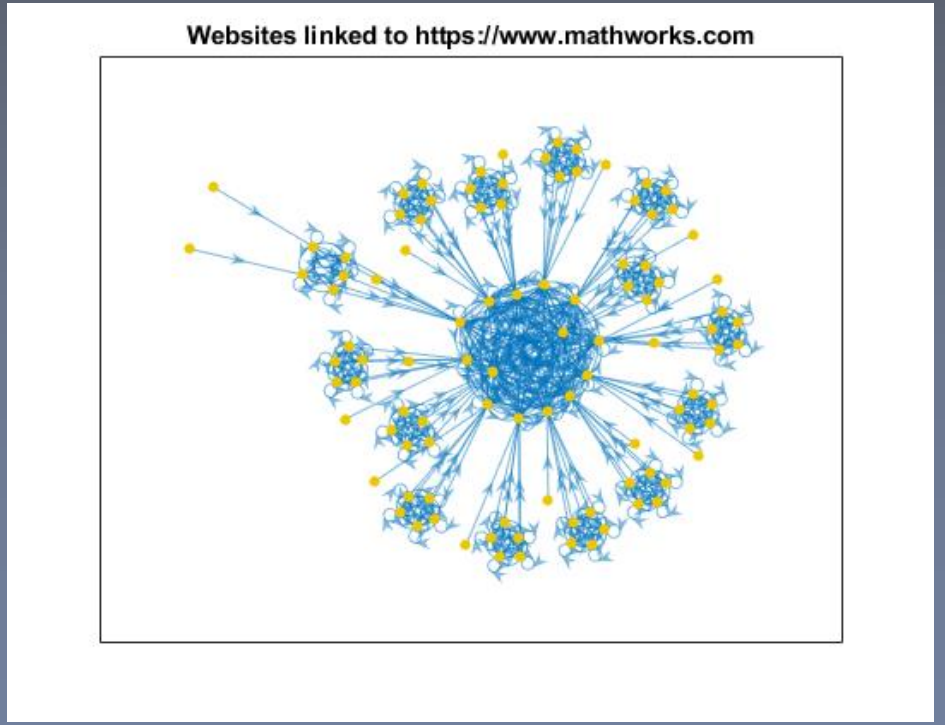
- 在一定条件下，极限情况访问每个结点的概率收敛到平稳分布，这时各个结点的平稳概率值就是其PageRank值，表示结点的重要度。PageRank是递归定义的，PageRank的计算可以 通过迭代算法进行。



PageRank算法基本想法

Basic Idea of PageRank Algorithm

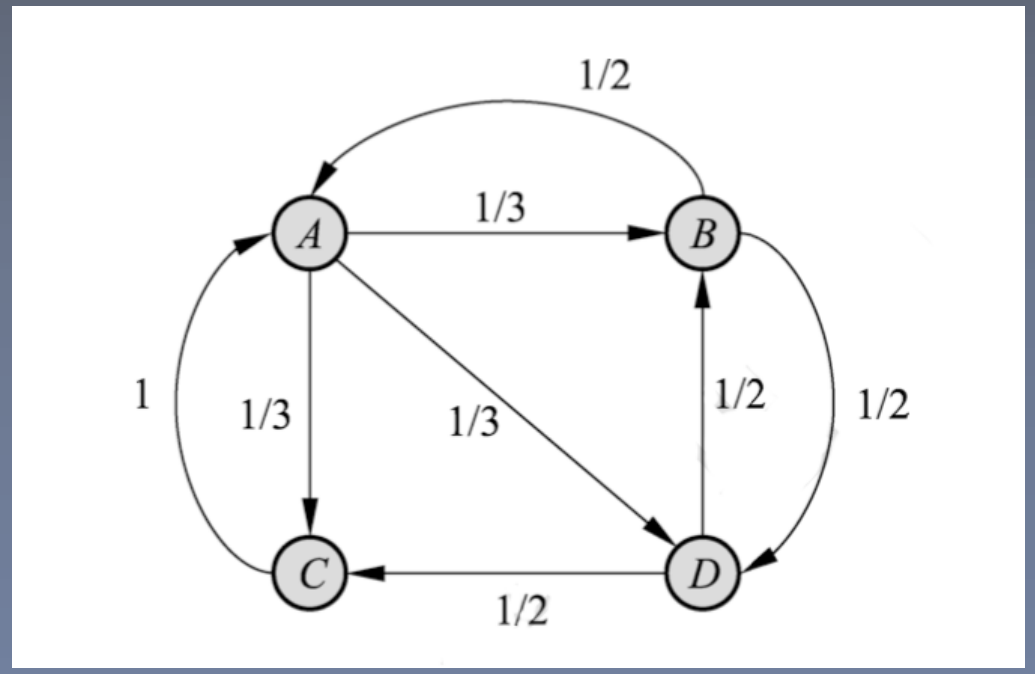
- 历史上，PageRank算法作为计算互联网网页重要度的算法被提出
- PageRank是定义在网页集合上的一个函数，它对每个网页给出一个正实数，表示网页的重要程度， 整体构成一个向量
- PageRank值越高，网页就越重要，在互联网搜索的排序中可能就被排在前面
- 假设互联网是一个有向图，在其基础上定义随机游走模型，即一阶马尔可夫链， 表示网页浏览者在互联网上随机浏览网页的过程
- 假设浏览者在每个网页依照连接出去的超链接以等概率跳转到下一个网页，并在网上持续不断进行这样的随机跳转，这个过程形成一阶马尔可夫链
- PageRank表示这个马尔可夫链的平稳分布。每个网页的PageRank值就是平稳概率。



PageRank算法基本想法

Basic Idea of PageRank Algorithm

- 右图表示一个有向图，假设是简化的互联网例，结点A,B,C和D表示网页， 结点之间的有向边表示网页之间的超链接，边上的权值表示网页之间随机跳转的概率



- 假设有一个浏览者，在网上随机游走
- 如果浏览者在网页A，
  - 则下一步以1/3的概率转移到网页B,C和D
- 如果浏览者在网页B，
  - 则下一步以1/2的概率转移到网页A和D
- 如果浏览者在网页C，
  - 则下一步以概率1转移到网页A
- 如果浏览者在网页D，
  - 则下一步以1/2的概率转移到网页B和C

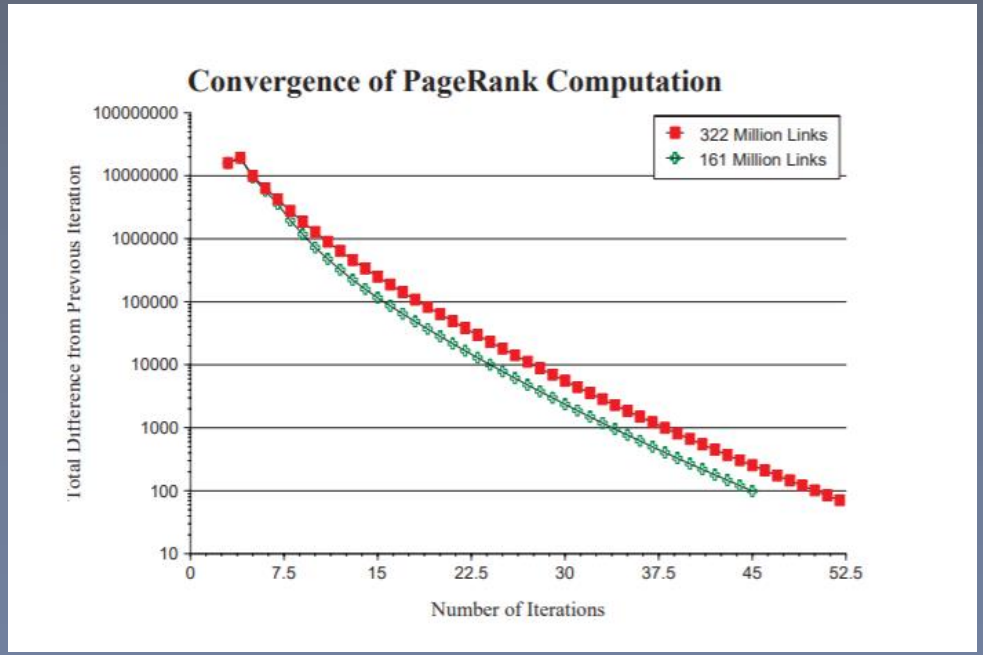
PageRank算法基本想法

Basic Idea of PageRank Algorithm

- 直观上，一个网页，如果指向该网页的超链接越多，随机跳转到该网页的概率也就越高，该网页的PageRank值就越高，这个网页也就越重要

- 一个网页，如果指向该网页的PageRank值越高，随机跳转到该网页的概率也就越高，该网页的PageRank 值就越高，这个网页也就越重要。PageRank值依赖于网络的拓扑结构，一旦网络的拓扑（连接关系）确定，PageRank值就确定

- PageRank的计算可以在互联网的有向图上进行，通常是一个迭代过程。先假设一 个初始分布，通过迭代，不断计算所有网页的PageRank值，直到收敛为止



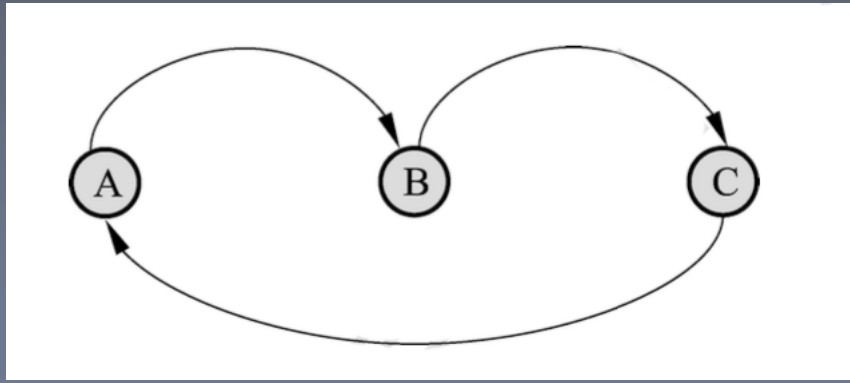
有向图

Directed Graph

定义21.1（有向图）有向图记作G=(V,E)，其中V和E分别表示结点和有向边的集合。

从一个结点出发到达另一个结点，所经过的边的一个序列称为一条路径（(path)，路径上边的个数称为路径的长度。如果一个有向图从其中任何一个结点出发可以到达 其他任何一个结点，就称这个有向图是强连通图（strongly connected graph）

假设k是一个大于1的自然数，如果从有向图的一个结点出发返回到这个结点的路径的长度都是k的倍数，那么称这个结点为周期性结点，如果一个有向图不含有周期性结点，则称这个有向图为非周期性图（aperiodic graph），否则为周期性图



从结点A出发返回到A，必须经过路径 A—B—C—A，所有可能的路径的长度都是3的倍数，所以结点A是周期性结点。 这个有向图是周期性图

随机游走模型

Random Walk Model

- **定义21.2（随机游走模型）**给定一个含有n个结点的有向图，在有向图上定义随机游走模型，即一阶马尔科夫模型，其中结点表示状态，有向边表示状态之间的转移，假设从一个结点到通过有向边相连的所有结点的转移概率相等。具体地，转移矩阵是一个n阶矩阵M
- $$M=[m_{ij}]_{n\times n} \tag{21.1}$$
- 第i行第j列的元素mij取值规则如下：如果结点j有k个有向边连出，并且结点i是其连出的一个结点，则$m_{ij}=1/k$, 否则$m_{ij}=0$, i，j=1, 2, . . . , n。
- 注意转移矩阵具有性质
- $$m_{ij}\geqslant 0$$
- $$\sum_{i=1,\ldots,n} m_{ij}=1$$
- 即每个元素非负，每列元素之和为1，即矩阵M为随机矩阵（stochastic matrix)。

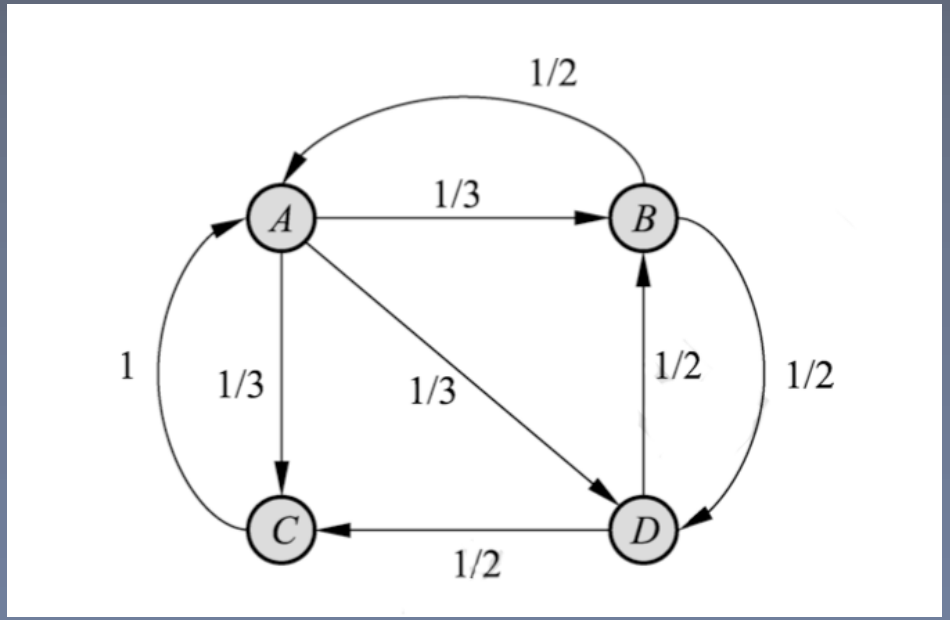
随机游走模型

Random Walk Model

- 在有向图上的随机游走形成马尔可夫链。也就是说，随机游走者每经一个单位时间转移一个状态。如果当前时刻在第j个结点（状态），那么下一个时刻在第i个结 点（状态）的概率是m_{ij}

- 这一概率只依赖于当前的状态，与过去无关，具有马尔可夫性。在右图的有向图上可以定义随机游走模型

- 结点A到结点B, C和D存在有向边，可以以概率1/3从A分别转移到B, C和D，并以概率0转移到A，于是可以写出转移矩阵的第1列。结点B到结点A和D存在有向边，可以以概率1/2从B分别转移到A和D，并以概率0分别转移到B和C，于是可以写出矩阵的第2列



随机游走模型

Random Walk Model

- 于是得到转移矩阵

$$M=\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- 随机游走在某个时刻t访问各个结点的概率分布就是马尔可夫链在时刻t的状态 分布，可以用一个n维列向量 R_t 表示，那么在时刻t+1访问各个结点的概率分布 R_{t+1}满足

- $$R_{t+1}=MR_t$$

PageRank基本定义

Basic Definition of PageRank

- 给定一个包含n个结点的强连通且非周期性的有向图，在其基础上定义随机游走模型
- 假设转移矩阵为M，在时刻0,1,2,...,t,…访问各个结点的概率分布为$R_0, MR_0, M^2R_0, ..., M^tR_0, ...$
- 则极限 $\lim_{t\to\infty} M^tR_0$ 存在
- 极限向量R表示马尔可夫链的平稳分布，满足 MR=R

- **定义21.3（PageRank基本定义）**给定一个包含n个结点v1,v2,...,vn的强连通且非周期性的有向图，在有向图上定义随机游走模型，即一阶马尔科夫链，随机游走的特点是从一个结点到有有向边连出的所有结点的转移概率相等，转移矩阵为M，这个马尔科夫链具有平稳分布R,MR=R （21.6）
- 平稳分布R称为这个有向图的PageRank.R的各个分量称为各个结点的PageRank值。
- $R=[PR(v_1)\ PR(v_2)\ ...\ PR(v_n)]^T$
- 其中PR(vi),i=1,2,...,n，表示结点vi的PageRank值。
-

PageRank基本定义

Basic Definition of PageRank

- 显然有

$$PR(v_i) \geqslant 0, \quad i = 1, 2, \cdots, n$$

$$\sum_{i=1}^{n} PR(v_i) = 1$$

$$PR(v_i) = \sum_{v_j \in M(v_i)} \frac{PR(v_j)}{L(v_j)}, \quad i = 1, 2, \cdots, n$$

- $M(v_i)$表示指向结点v_i的结点集合
- $L(v_j)$表示结点v_j连出的有向边的个数
- PageRank的基本定义是理想化的，在这中情况下，PageRank存在，而且可以通过不断迭代求得PageRank值

PageRank基本定义

Basic Definition of PageRank

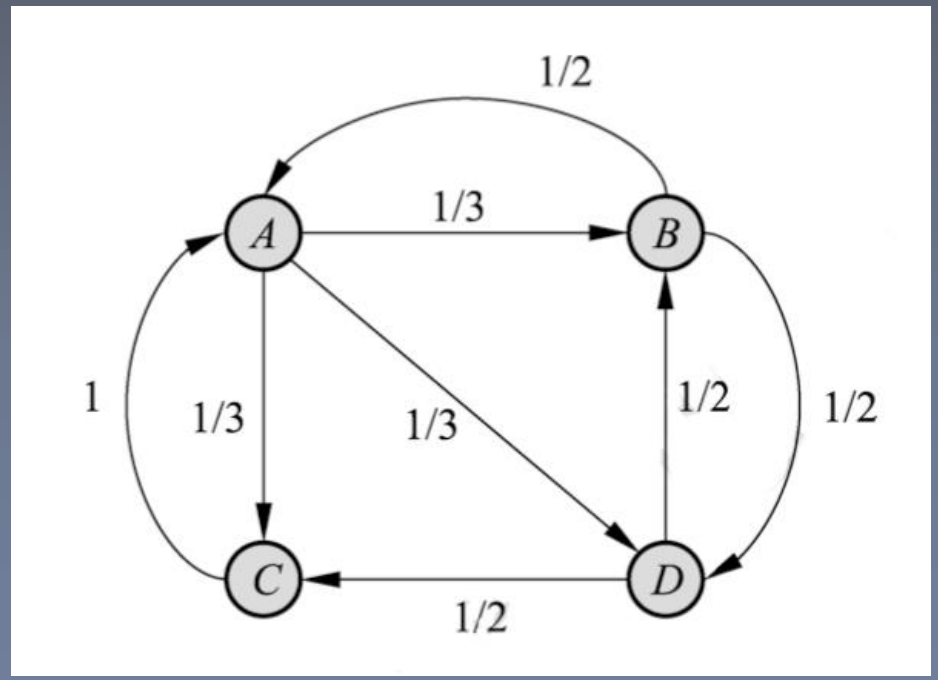
- **定理21.1（马尔科夫链平稳分布定理）** 不可约且非周期的有限状态马尔科夫链，有唯一平稳分布存在，并且当时间趋于无穷时状态分布收敛于唯一的平稳分布
- 根据马尔科夫链平稳分布定理，强连通且非周期的有向图上定义的随机游走模型（马尔科夫链），在图上的随机游走当时间趋于无穷时状态分布收敛于唯一的平稳分布。

PageRank例题

Example of PageRank

- 转移矩阵

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$



- 取初始分布向量 R_0 为$R_0=[1/4\ \ 1/4\ \ 1/4\ \ 1/4]^T$
- 以转移矩阵M连乘初始向量 R_0 得到向量序列

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}, \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}, \begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix}, \cdots, \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

- 最后得到极限向量$R=[3/9\ \ 2/9\ \ 2/9\ \ 2/9]^T$ 即有向图的PageRank值

- 一般的有向图未必满足强连通且非周期性的条件。所以PageRank 的基本定义不适用。

PageRank的一般定义

General Definition of PageRank

- PageRank一般定义的想法是在基本定义的基础上导入平滑项
- 给定一个含有n个结点v_i， i=1, 2, … , n，的任意有向图. 假设考虑一个在图上 随机游走模型，即一阶马尔可夫链，其转移矩阵是M，从一个结点到其连出的所有结点的转移概率相等。
- 这个马尔可夫链未必具有平稳分布. 假设考虑另一个完全随机游走的模型，其转移矩阵的元素全部为1/n，也就是说从任意一个结点到任意一个结点的转移概率都是1/n
- 两个转移矩阵的线性组合又构成一个新的转移矩阵，在其上可以定义一个新的马尔可夫链。

PageRank的一般定义

General Definition of PageRank

- 容易证明这个马尔可夫链一定具有平稳分布，且平稳分布满足R=dMR+（1-d）/n *1 (21.10)
- 式中d(0≤d ≤ 1）是系数，称为阻尼因子（damping factor).R是n维向量，1是所有分量为1的n维向量
- R表示的就是有向图的一般PageRank,R=[PR(v_1) PR(v_2) . . . PR(v_n)]^T。PR(v_i), i=1,2,. . . ,n，表示结点v_i的PageRank值
- 式（21.10）中第一项表示（状态分布是平稳分布时）依照转移矩阵M访问各个结点的概率，第二项表示完全随机访问各个结点的概率
- 阻尼因子d取值由经验决定。例如d=0.85。当d接近1时，随机游走主要依照转移矩阵M进行。当d接近0时， 随机游走主要以等概率随机访问各个结点。
- 可以由式（21.10)写出每个结点的PageRank，这是一般PageRank的定义
- $PR(v_i)=d\left(\sum_{vj\in M(vi)}PR(v_i)/L(v_j)\right)+(1-d)/n, \quad i=1,2,\ldots,n$
- 第二项称为平滑项，由于采用平滑项，所有结点的PageRank值都不会为0，具有以下性质：
  PR(vi)>0, i=1,2,. . . ,n, $\sum_{i=1,\ldots,n}PR(v_i)=1$

PageRank的一般定义

General Definition of PageRank

- **定义21.4（PageRank的一般定义）**给定一个含有n个结点的任意有向图，在有向图上定义一个一般的随机游走模型，即一阶马尔科夫链，一般的随机游走模型的转移矩阵由两部分的线性组合组成，一部分是有向图的基本转移矩阵M，表示从一个结点到其连出的所有结点的转移概率相等，另一部分是完全随机的转移矩阵，表示从任意一个结点到任意一个结点的转移概率都是1/n，线性组合系数为阻尼因子d(0≤d≤1)，这个一般随机游走的马尔科夫链存在平稳分布，记作R，定义平稳分布向量R为这个有向图的一般PageRank. R由公式 R=dMR+（1-d）/n *1　决定，其中1是所有分量为1的n维向量

- 一般PageRank的定义意味着互联网浏览者，按照以下方法在网上随机游走：，在任意一个网页上，浏览者或者以概率d决定按照超链接随机跳转，这时以等概率从连接出去的超链接跳转到下一个网页，或者以概率（1-d)决定完全随机跳转，这时以等概率1/n跳转到任意一个网页
- 第二个机制保证从没有连接出去的超链接的网页也可以跳转出。这样可以保证平稳分布，即一般PageRank的存在，因而一般PageRank适用于任何结构的网络。
-

迭代算法

Iteration Algorithm

- 给定一个含有n个结点的有向图，转移矩阵为M，有向图的一般PageRank由迭代公式
- $R_{t+1}=dMR_t+(1-d)/n \ *1$ 的极限向量R确定
- PageRank的迭代算法，就是按照这个一般定义进行迭代，直至收敛

算法 21.1（PageRank 的迭代算法）

输入：含有 n 个结点的有向图，转移矩阵 M，阻尼因子 d，初始向量 R_0；

输出：有向图的 PageRank 向量 R。

（1）令 $t=0$

（2）计算

$$R_{t+1}=dMR_t+\frac{1-d}{n}\mathbf{1}$$

（3）如果 R_{t+1} 与 R_t 充分接近，令 $R=R_{t+1}$，停止迭代。

（4）否则，令 $t=t+1$，执行步（2）。 ■

PageRank算法正确性

Correctness of PageRank Algorithm

取e为所有分量都为 1 的列向量,接着定义矩阵:

$$A = \alpha S + \frac{(1-\alpha)}{N} e e^T$$

则PR值的计算如下，其中P_n为第n次迭代时各网页PR值组成的列向量:

$$P_{n+1} = AP_n$$

于是计算PR值的过程就变成了一个 Markov 过程，那么PageRank算法的证明也就转为证明 Markov 过程的收敛性证明：如果这个 Markov 过程收敛，那么$\lim_{n\to\infty} P_n$存在，且与P_0的选取无关。

若一个 Markov 过程收敛，那么它的状态转移矩阵A需要满足[6]:

1. A为随机矩阵。
2. A是不可约的。
3. A是非周期的。

先看第一点，随机矩阵又叫概率矩阵或 Markov 矩阵，满足以下条件:

令a_{ij}为矩阵A中第i行第j列的元素，则$\forall i = 1 \dots n, j = 1 \dots n, a_{ij} \geq 0$, 且$\forall i = 1 \dots n, \sum_{j=1}^{n} a_{ij} = 1$

显然我们的A矩阵所有元素都大于等于0，并且每一列的元素和都为1。

第二点，不可约矩阵：方针A是不可约的当且仅当与A对应的有向图是强联通的。有向图$G = (V, E)$是强联通的当且仅当对每一对节点对$u, v \in V$，存在从u到v的路径。因为我们在之前设定用户在浏览页面的时候有确定概率通过输入网址的方式访问一个随机网页，所以A矩阵同样满足不可约的要求。

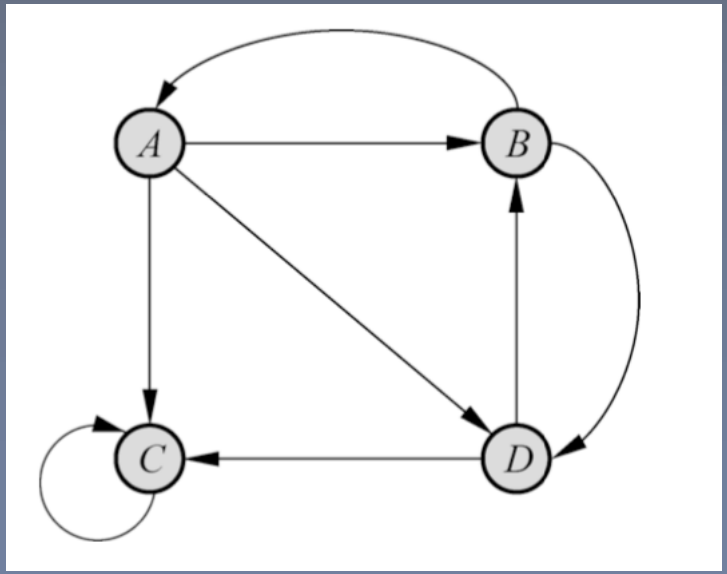
第三点，要求A是非周期的。所谓周期性，体现在Markov链的周期性上。即若A是周期性的，那么这个Markov链的状态就是周期性变化的。因为A是素矩阵（素矩阵指自身的某个次幂为正矩阵的矩阵），所以A是非周期的。

至此，我们证明了PageRank算法的正确性。

迭代算法例题

Example of Iteration Algorithm

- 图中所示的有向图，取d = 0.8，求图的PageRank



迭代算法例题

Example of Iteration Algorithm

- 可得转移矩阵为

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

按照式（21.15）计算

$$dM = \frac{4}{5} \times \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix}$$

$$\frac{1-d}{n}\mathbf{1} = \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

迭代算法例题

Example of Iteration Algorithm

- 迭代公式为

$$R_{t+1} = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} R_t + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

- 令初始向量
- $R_0=[1/4\ \ 1/4\ \ 1/4\ \ 1/4]^T$
- 进行迭代

$$R_1 = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix} = \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix} + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix} = \begin{bmatrix} 41/300 \\ 53/300 \\ 153/300 \\ 53/300 \end{bmatrix}$$

迭代算法例题

Example of Iteration Algorithm

- 最后得到

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}, \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix}, \begin{bmatrix} 41/300 \\ 53/300 \\ 153/300 \\ 53/300 \end{bmatrix}, \begin{bmatrix} 543/4500 \\ 707/4500 \\ 2543/4500 \\ 707/4500 \end{bmatrix}, \cdots, \begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix}$$

- 计算结果表明，结点C的PageRank值超过一半，其他结点也有相应的 PageRank值。

幂法

Power Method

- 幂法（(power method)是一个常用的PageRank计算方法，通过近似计算矩阵的主特征值和主特征向量求得有向图的一般PageRank

- 幂法主要用于近似计算矩阵的主特征值（dominant eigenvalue）和 主特征向量（dominant eigenvector）

- 主特征值是指绝对值最大的特征值,主特征向量是其对应的特征向量
- 注意特征向量不是唯一的，只是其方向是确定的，乘上任意系数还是特征向量

幂法

Power Method

- 假设要求n阶矩阵A的主特征值和主特征向量，采用下面的步骤。
- 首先，任取一个初始向量x_0，构造如下的一个n维向量序列
- $$x_0, x_1=Ax_0, x_2=Ax_1, \ldots, x_k=Ax_{k-1}$$
- 然后，假设矩阵A有n个特征值，按照绝对值大小排列
- $$|\lambda_1| \geqslant \ldots \geqslant |\lambda_n|$$
- 对应的n个线性无关的特征向量为$u_1, u_2, \ldots, u_n$
- 这n个特征向量构成n维空间的一组基

幂法

Power Method

- 于是，可以将初始向量 x_0 表示为$u_1, u_2, \ldots, u_n$的线性组合
- $$x_0=a_1u_1+a_2u_2+\ldots+a_nu_n$$
- 得到

$$
\begin{aligned}
x_1 &= Ax_0 = a_1Au_1 + a_2Au_2 + \cdots + a_nAu_n \\
&\vdots \\
x_k &= A^kx_0 = a_1A^ku_1 + a_2A^ku_2 + \cdots + a_nA^ku_n \\
&= a_1\lambda_1^ku_1 + a_2\lambda_2^ku_2 + \cdots + a_n\lambda_n^ku_n
\end{aligned}
$$

幂法

Power Method

- 接着，假设矩阵A的主特征值λ_1是特征方程的单根，由上式得

$$x_k = a_1\lambda_1^k\left[u_1 + \frac{a_2}{a_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k u_2 + \cdots + \frac{a_n}{a_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k u_n\right]$$

- 由于 $|\lambda_1| \geqslant |\lambda_j|$, j=2, 3, . . . , n， 当k充分大时有
- $x_k = a_1\lambda_1^k[u_1+\varepsilon_k]$
- 这里ε_k是当$k\to\infty$时的无穷小量。即$x_k \to a_1\lambda_1^k u_1 (k\to\infty)$

幂法

Power Method

- 说明当k充分大时向量 x_k 与特征向量 v_1 只相差一个系数。由式（21.18)知，
- $$x_k \approx a_1 \lambda_1^k u_1$$
- $$x_{k+1} \approx a_1 \lambda_1^{k+1} u_1$$
- 于是主特征值 λ_1 可表示为
- $$\lambda_1 \approx x_{k+1,\,j} / x_{k,\,j}$$
-
- 其中 $x_{k,\,j}$ 和 $x_{k+1,\,j}$ 分别是 x_k 和 x_{k+1} 的第j个分量

幂法

Power Method

- 在实际计算时，为了避免出现绝对值过大或过小的情况，通常在每步迭代后即进 行规范化，将向量除以其范数，即 $y_{t+1}=Ax_t$ ，$x_{t+1}=y_{t+1}/||y_{t+1}||$
- 这里的范数是向量的无穷范数，即向量各分量的绝对值的最大值
- $$||x||_{\infty}=\max\{|x_1|,|x_2|,\ldots,|x_n|\}$$
- 现在回到计算一般PageRank。转移矩阵可以写作

$$R=\left(dM+\frac{1-d}{n}\mathbf{E}\right)R=AR$$

- 其中d是阻尼因子
- E是所有元素为1的n阶方阵
- 根据Perron-Frobenius定理， 一般PageRank的向量R是矩阵A的主特征向量，主特征值是1
- 所以可以使用幂法 近似计算一般PageRank

计算一般PageRank的幂法

Power Method for Calculating General PageRank

输入：含有 n 个结点的有向图，有向图的转移矩阵 M，系数 d，初始向量 x_0，计算精度 ε；

输出：有向图的 PageRank$R$。

（1）令 $t=0$，选择初始向量 x_0

（2）计算有向图的一般转移矩阵 A

$$A = dM + \frac{1-d}{n}\mathbf{E}$$

（3）迭代并规范化结果向量

$$y_{t+1} = Ax_t$$

$$x_{t+1} = \frac{y_{t+1}}{\|y_{t+1}\|}$$

（4）当 $\|x_{t+1} - x_t\| < \varepsilon$ 时，令 $R = x_t$，停止迭代。

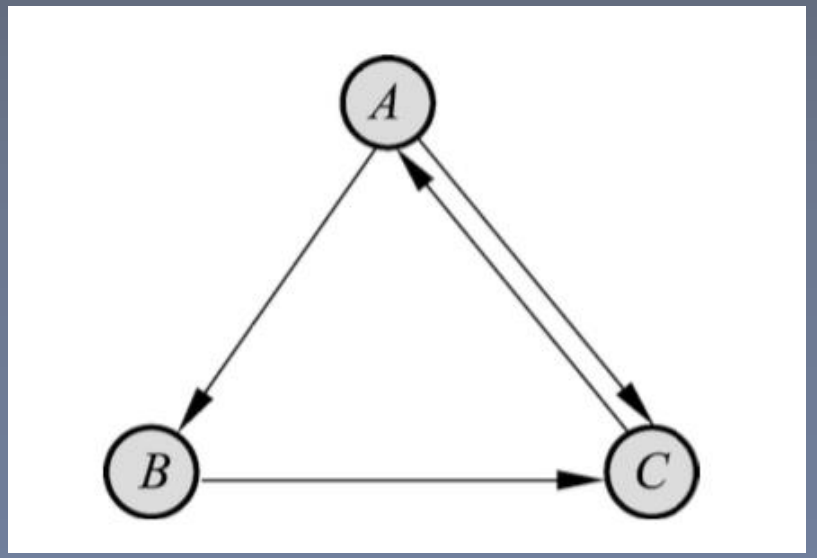
（5）否则，令 $t = t+1$，执行步（3）。

（6）对 R 进行规范化处理，使其表示概率分布。 ■

计算一般PageRank的幂法

Example of Calculating General PageRank

- 给定一个如图所示的有向图，取 d = 0.85，求有向图的一般PageRank。
- 利用幂法，按照算法 21.2，计算有向图的一般 PageRank



- 由图可知转移矩阵

$$M = \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{bmatrix}$$

- （1）令 t = 0 , $x_0=[1, 1, 1]^T$

计算一般PageRank的幂法

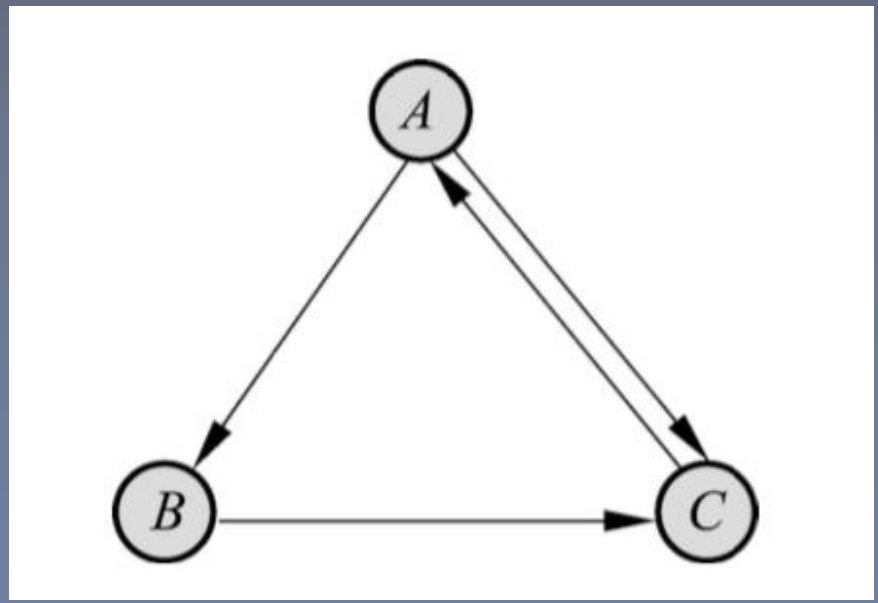
Example of Calculating General PageRank

- （2）计算有向图的一般转移矩阵 A

$$A = dM + \frac{1-d}{n}\mathbf{E}$$

$$= 0.85 \times \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{bmatrix} + \frac{0.15}{3} \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.05 & 0.05 & 0.9 \\ 0.475 & 0.05 & 0.05 \\ 0.475 & 0.9 & 0.05 \end{bmatrix}$$



计算一般PageRank的幂法

Example of Calculating General PageRank

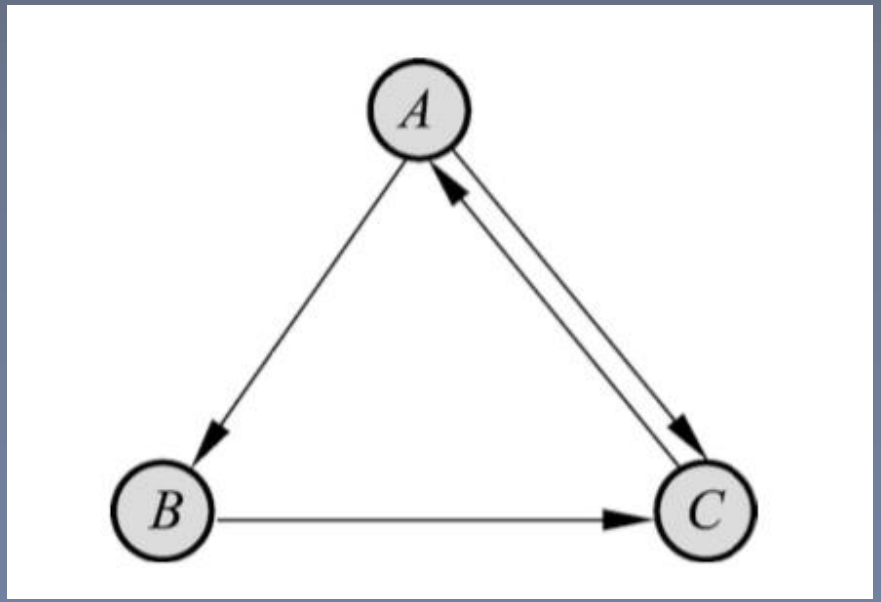
- （3）迭代并规范化

$$y_1 = Ax_0 = \begin{bmatrix} 1 \\ 0.575 \\ 1.425 \end{bmatrix}$$

$$x_1 = \frac{1}{1.425}\begin{bmatrix} 1 \\ 0.575 \\ 1.425 \end{bmatrix} = \begin{bmatrix} 0.7018 \\ 0.4035 \\ 1 \end{bmatrix}$$

$$y_2 = Ax_1 = \begin{bmatrix} 0.05 & 0.05 & 0.9 \\ 0.475 & 0.05 & 0.05 \\ 0.475 & 0.9 & 0.05 \end{bmatrix}\begin{bmatrix} 0.7018 \\ 0.4035 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.9553 \\ 0.4035 \\ 0.7465 \end{bmatrix}$$

$$x_2 = \frac{1}{0.9553}\begin{bmatrix} 0.9553 \\ 0.4035 \\ 0.7465 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.4224 \\ 0.7814 \end{bmatrix}$$



计算一般PageRank的幂法

Example of Calculating General PageRank

- （3）如此继续迭代规范化，得到x_t, t=0, 1, 2, . . . , 21, 22的向量序列

$$\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}0.7018\\0.4035\\1\end{bmatrix}, \begin{bmatrix}1\\0.4224\\0.7814\end{bmatrix}, \begin{bmatrix}0.8659\\0.5985\\1\end{bmatrix}, \begin{bmatrix}0.9732\\0.4912\\1\end{bmatrix}, \begin{bmatrix}1\\0.5516\\0.9807\end{bmatrix},$$

$$\begin{bmatrix}0.9409\\0.5405\\1\end{bmatrix}, \cdots, \begin{bmatrix}0.9760\\0.5408\\1\end{bmatrix}, \begin{bmatrix}0.9755\\0.5404\\1\end{bmatrix}, \begin{bmatrix}0.9761\\0.5406\\1\end{bmatrix}, \begin{bmatrix}0.9756\\0.5406\\1\end{bmatrix}, \begin{bmatrix}0.9758\\0.5404\\1\end{bmatrix}$$

- 假设后面得到的两个向量已满足计算精度要求，那么取$R=[0.9756, 0.5406, 1]^T$即得所求的一般PageRank作为一个概率分布，进行规范化，使各分量之和为1，那么相应的一般PageRank可以写作$R=[0.3877, 0.2149, 0.3974]^T$。

代数算法

Algebra Algorithm

- 代数算法通过一般转移矩阵的逆矩阵计算求有向图的一般PageRank
- 按照一般PageRank的定义式（21.14）
- $$R=dMR+(1-d)/n \ *1$$
- 于是
- $$(I-dM)R=(1-d)/n \ *1$$
- $$R=(I-dM)^{-1}(1-d)/n \ *1$$

- 这里I是单位矩阵。当0<d<1时，线性方程组（21.23）的解存在且唯一
- 这样，可以通过求逆矩阵$(I-dM)^{-1}$得到有向图的一般PageRank

PageRank算法的优缺点

Strength and Weakness of PageRank Algorithm

- **优点：**是一个与查询无关的静态算法，所有网页的PageRank值通过离线计算获得；有效减少在线查询时的计算量，极大降低了查询响应时间。
- **缺点：** 第一，没有区分站内导航链接，忽略了主题相关性。很多网站的首页都有很多对站内其他页面的链接，称为站内导航链接。这些链接与不同网站之间的链接相比，肯定是后者更能体现PageRank值的传递关系。
- 第二，没有过滤广告链接和功能链接（例如常见的“分享到微博”）。这些链接通常没有什么实际价值，前者链接到广告页面，后者常常链接到某个社交网站首页。
- 第三，对新网页不友好。一个新网页的一般入链相对较少，除非它是某个站点的子站点。即使它的内容的质量很高，要成为一个高PR值的页面仍需要很长时间的推广。
- 针对PageRank算法的缺点，有人提出了TrustRank算法。其最初来自于2004年斯坦福大学和雅虎的一项联合研究，用来检测垃圾网站。TrustRank算法的工作原理：先人工去识别高质量的页面(即“种子”页面)，那么由“种子”页面指向的页面也可能是高质量页面，即其TR值也高，与“种子”页面的链接越远，页面的TR值越低。“种子”页面可选出链数较多的网页，也可选PR值较高的网站。
- TrustRank算法给出每个网页的TR值。将PR值与TR值结合起来，可以更准确地判断网页的重要性。

结 语

在这次课程中我们了解到了PageRank算法一般定义、随机游走模型、迭代算法与幂法求解PageRank值等

希望在课下，大家都能

掌握本节知识并且完成作业

deepshare.net

深度之眼

联系我们：

电话：18001992849

邮箱：service@deepshare.net

Q Q： 2677693114

公众号

客服微信

关注公众号深度之眼，后台回复 统计学 ， 获取统计学习方法第二版电子书及其他AI必学书籍