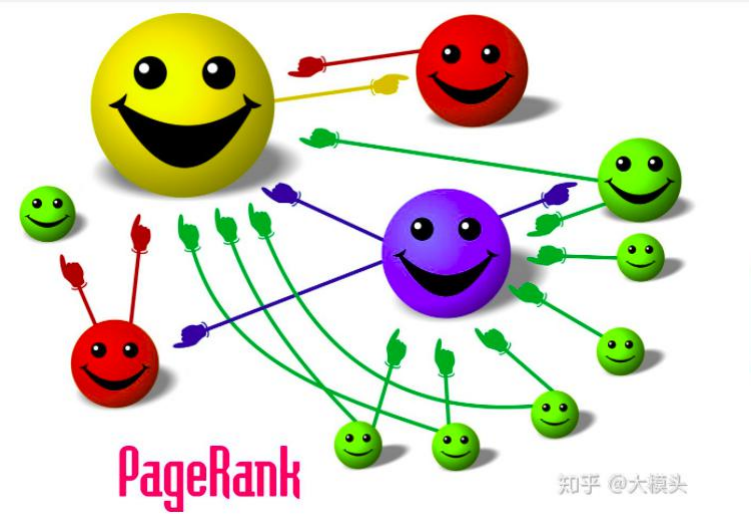
PageRank算法分析

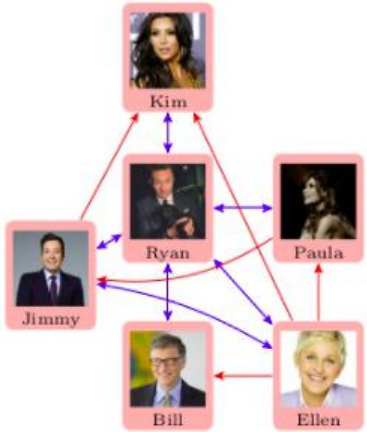
王文卿
2021.08

关于PageRank算法

网页推荐



社交推荐



交易关系数据都以网络图（graph）的形式存在，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。在互联网、社交网络等领域也有广泛应用。

• PageRank算法基本思想

PageRank算法通过分析交易数据，对每个交易节点（个人、公司、小微商户等）给出一个正实数，表示交易节点的重要程度，整体构成一个向量，PageRank值越高，交易节点就越重要，在产品推荐的排序中可能就被排在前面。

图1表示一个有向图，假设是简化的交易网络，结点A，B，C和D表示交易节点，结点之间的有向边表示交易节点之间的转账等交易，边上的权值表示节点之间随机交易的概率。假设有一个交易节点，在网络中随机游走。如果节点A想要发起一笔转账交易，则以 1/3的概率转账给B，C和D。同理，B节点想要发起一笔转账交易，则以1/2的概率转账给A和D。

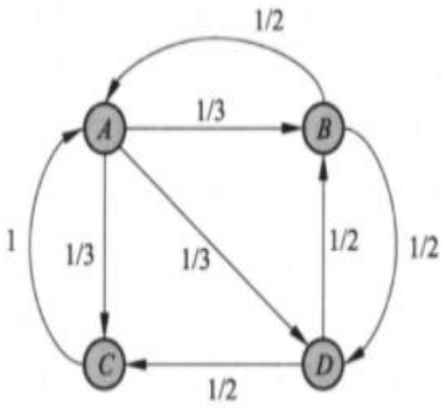


图1 有向图

直观上，一个交易节点，如果指向该节点的交易越多，随机与该节点交易的概率也就越高，该节点的PageRank值就越高，这个节点也就越重要。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 的计算可以在交易网络的有向图上进行，通常是一个迭代过程。先假设一 个初始分布，通过迭代，不断计算所有节点的PageRank值，直到收敛为止。

• 随机游走

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（random walk）模型，即一阶马尔可夫链，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

$$M = [m_{ij}]_{n \times n} \tag{1}$$

第i行第j列的元素 m_{ij} 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 $m_{ij} = 0$ ，$i, j = 1, 2, \cdots\cdots, n$

在图1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。同理，得到转移矩阵M:

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

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满足:

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

• 随机游走

t = 0 , 则:

$$R_{t+1} = MR_t = \begin{pmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{pmatrix}$$

当t→+∞，且概率转移矩阵M满足以下3个条件时，lim t→+∞，最终收敛与R，保持在一个稳定值附近。

M为随机矩阵。即所有M[i][j]≥0 ，且的所有列向量的元素加和为1，∑M[i][j]=1
M是不可约的。比如：0矩阵，导致R为0
M**是非周期的。**

• PageRank基本定义

给定一个包含n个结点的有向图，在其基础上定义随机游走。假设转移矩阵为M，在时刻 $0,1,2,\cdots,t,\cdots$ 访问各个结点的概率分布为

$$R_0, MR_0, M^2R_0, \cdots, M^tR_0, \cdots$$

则极限为（无限节点一直随机游走）：

$$\lim_{t\to\infty} M^tR_0 = R$$

存在，极限向量R表示马尔可夫链的平稳分布，满足

$$MR = R$$

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。

$$R = \begin{bmatrix} PR(v_1) \\ PR(v_2) \\ \vdots \\ PR(v_n) \end{bmatrix}$$

计算

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

这里 $M(v_i)$ 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计算

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，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 $d(0 \leq d \leq 1)$ 。 这个一般随机游走的马尔可夫链存在平稳分布，记作R。定义平稳分布向量R为这个有向图的一般 PageRank。 由公式

$$R = dMR + \frac{1-d}{n}1$$

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

$$R = dMR + \frac{1-d}{n}1$$

(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 $t = t + 1$,执行 (2)

谢谢！