《数据挖掘技术》

⭐ CH21 PageRank算法

➡️ Create by *Wang JingHui*

➡️ Last Revision Time: *2021.05.02*

章节目录

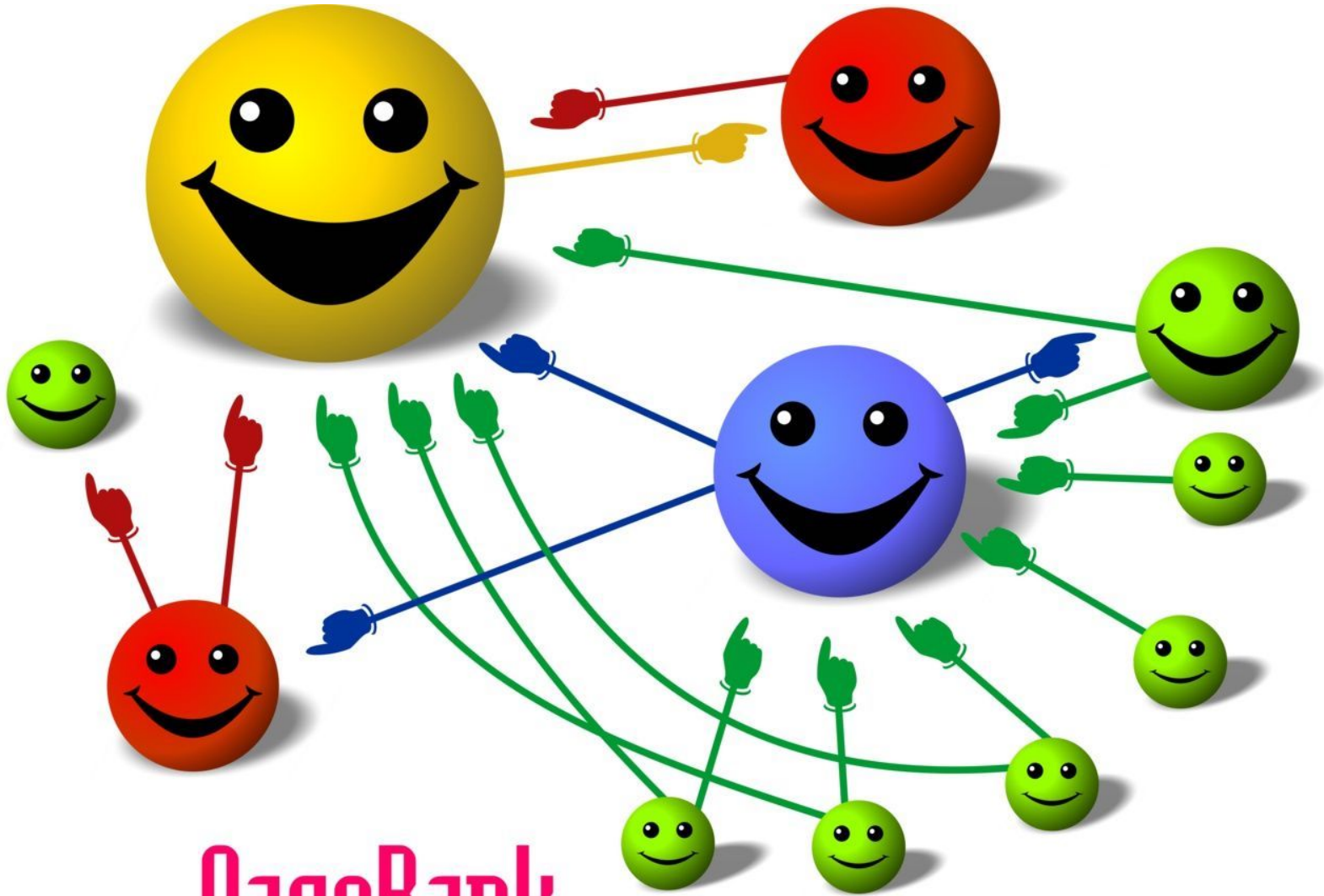
1. PageRank的定义
    i. 基本想法
    ii. 有向图和随机游走模型
    iii. PageRank的基本定义
    iv. PageRank的一般定义
2. PageRank的计算
    i. 迭代算法
    ii. 幂法
    iii. 代数算法

引论

Google早已成为全球最成功的互联网搜索引擎，但这个当前的搜索引擎巨无霸却不是最早的互联网搜索引擎，在Google出现之前，曾出现过许多通用或专业领域搜索引擎。Google最终能击败所有竞争对手，很大程度上是因为它解决了困扰前辈们的最大难题：对搜索结果按重要性排序。而解决这个问题的算法就是PageRank。毫不夸张的说，是PageRank算法成就了Google今天的地位。

- 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

PageRank的定义

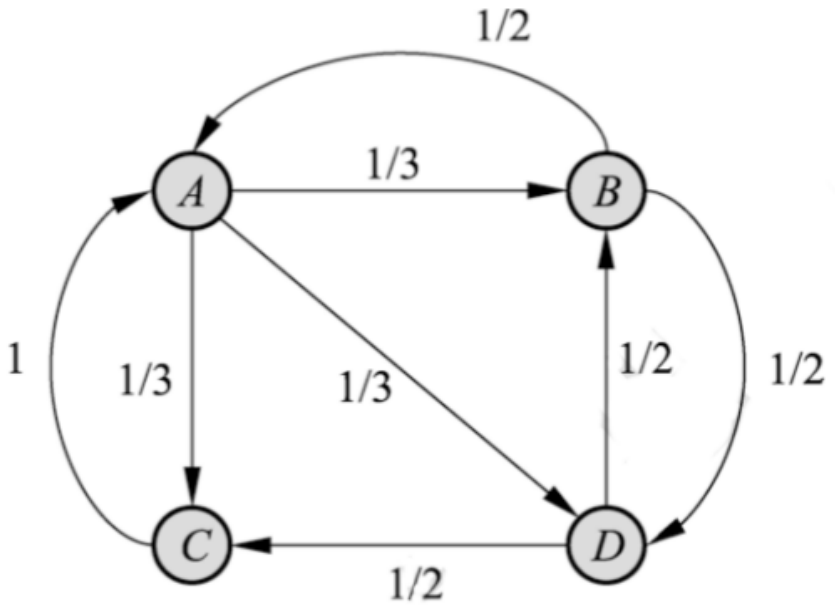
基本思想

- 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。

例：

如图表示一个有向图，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



分析

- 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。

有向图和随机游走模型

(Def)有向图

设V是一个非空集合，A是一个由V中元素的有序对构成的多重集，有序对 $D = < V, E >$ 称为一个有向图，其中，V称为顶点集，其中的元素称为顶点或点；E称为弧集，其中的元素是弧。

- 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

(Def)强连通图

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)

(Def)非周期性有向图

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。

随机游走模型

- 注意转移矩阵具有性质，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)。
- 给定一个含有n个结点的有向图，在有向图上的随机游走模型，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，具体的转移概率是一个n阶矩阵M

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

- 也就是说，随机游走者 每经一个单位时间转移一个状态。
- 第i行第j列元素m_{ij}取值规则如下：如果结点j有k的有向边连出，并且结点i是其连出的一个结点，则$m_{ij} = \frac{1}{k}$，否则 $m_{ij} = 0, i, j = 1, 2, ..., n.$
- 其中$m_{ij} \geq 0, \sum_{i=1}^{n} m_{ij} = 1$

PageRank的基本定义

(Def) 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$0, 1, 2, \ldots, t, \ldots$访问各个结点的概率分布为

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

则极限

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

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PageRank，R的各个分量称为各个结点的PageRank值。

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

其中$PR(v_i)$，$i = 1, 2, ..., n$，表示结点v_i的PageRank值。
显然有，$PR(v_i) \geq 0, i = 1, 2, ..., n$

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

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

$M(v_i)$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。

针对这个网页权重的计算模型还不够完善，主要有**两个问题**：

- **等级泄露（Rank Leak）**：如果一个网页没有出链，就像是一个黑洞一样，吸收了其他网页的影响力而不释放，最终会导致其他网页的 PR 值为 0。可以按照上面的计算公式，初始设置为1/4，经过一次更新后就变为了原来的一半，而C一直没有贡献，所以对应的那一列为0，多次迭代后也会导致其它页面为0 。
- **等级沉没（Rank Sink）**：如果一个网页只有出链，没有入链，计算的过程迭代下来，会导致这个网页的 PR 值为 0（也就是不存在公式中的 V）。这种情况看上去感觉影响小一点，只是自己页面最后变为0。

- 为了解决简化模型中存在的等级泄露和等级沉没的问题，拉里·佩奇提出了 PageRank 的随机浏览模型。他假设了这样一个场景：用户并不都是按照跳转链接的方式来上网，还有一种可能是不论当前处于哪个页面，都有概率访问到其他任意的页面，比如说用户就是要直接输入网址访问其他页面，虽然这个概率比较小。所以他定义了阻尼因子 d，这个因子代表了用户按照跳转链接来上网的概率，通常可以取一个固定值 0.85，而 $1-d=0.15$ 则代表了用户不是通过跳转链接的方式来访问网页的，比如直接输入网址。

PageRank的一般定义

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，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

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

- 式中$d(0 \leq d \leq 1)$是系数，称为阻尼因子(damping factor) $\mathbf{1}$是n维向量是所有分量为1的n维向量，R表示的就是有向图的一般PageRank，表示结点vi的PageRank值。

PageRank的一般定义

- 式中第一项表示(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。

PageRank的计算

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.

算法：

- (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

幂法

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

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x_0，计算精度ϵ;

输出：有向图的PageRank向量R。

算法：

- (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

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

- (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$\left|\left|x_{t+1} - x_t\right|\right| \leq \epsilon$，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。

代数算法

- 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

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

- 得到

$$(I - dM)R = \frac{1-d}{n}\mathbf{1}$$

$$R = (I - dm)^{-1}\frac{1-d}{n}\mathbf{1}$$

- 这里I是单位矩阵。当$0 < d < 1$时，线性方程组的解存在且唯一解， 这样，可以通过求逆矩阵$(I - dM)^{-1}$得到有向图的一般PageRank。

参考

https://networkx.org

https://www.osgeo.cn/networkx/

Enjoy your machine learning!

https://github.com/wjssx/Statistical-Learning-Slides-Code

E-mail: csr_dsp@sina.com