PageRank算法浅析

2013-06-04

tajpure

制作与演示

PageRank算法

PageRank算法的作用是评估网页的重要性，以此来作为搜索结果的重要排序依据之一。 PageRank算法是由Larry Page等人提出的一种基于超链分析的排序算法。

基本原理：借鉴传统情报检索理论中的引文分析方法—根据引文数量来确定文献权威性。

PageRank利用网络自身的超链接结构给所有网页确定一个重要性的等级数。

由上面的基本原理可以得到的直观的公式：

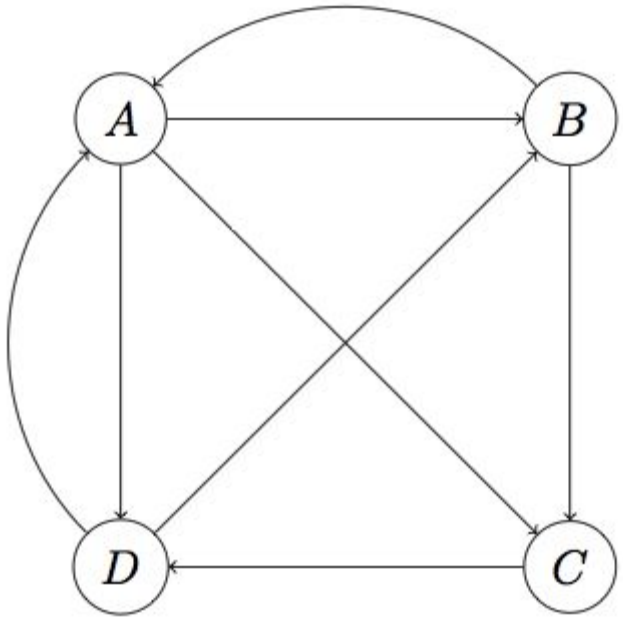
$$R(i) = C \sum_{j \in B(i)} \frac{R(j)}{N(j)}$$

其中R(x)表示页面x的rank值，B(i)表示所有指向i的网页，N(j)表示页面j所含有的超链接数。

1、将每个网页抽象成一个节点；

2、如果一个页面A有链接直接链向B，则存在一条有向边从A到B（多个相同链接不重复计算边）。

因此，整个Web被抽象为一张有向图。现在假设世界上只有四张网页：A、B、C、D，其抽象结构如下图：



设一共有N个网页，则可以组织这样一个N维矩阵：其中i行j列的值表示用户从页面j转到页面i的概率。那么可以建立如下转移矩阵：

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

之后，设初始时每个页面的rank值为1/N，此时这里的rank值就是1/4，按A-D的顺序将页面的rank值表示为向量v:

$$v = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

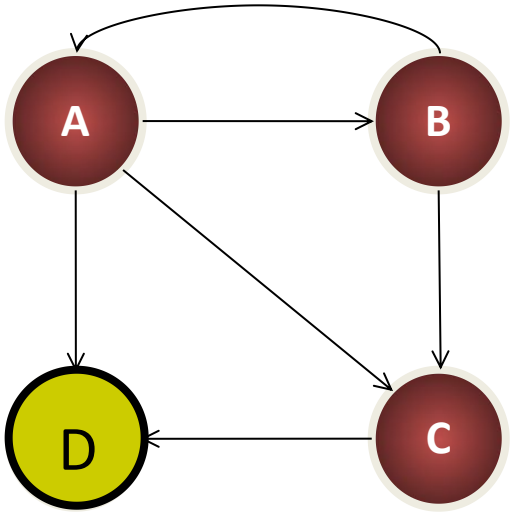
不断迭代，最终收敛

$$\mathrm{M} \cdot \mathrm{v} \longrightarrow Mv = \begin{bmatrix} 1/4 \\ 5/24 \\ 5/24 \\ 1/3 \end{bmatrix}$$

PageRank可能会遇见的问题

1，处理Dead Ends

所谓Dead Ends，就是这样一类节点：它们不存在外链。看下面的图：



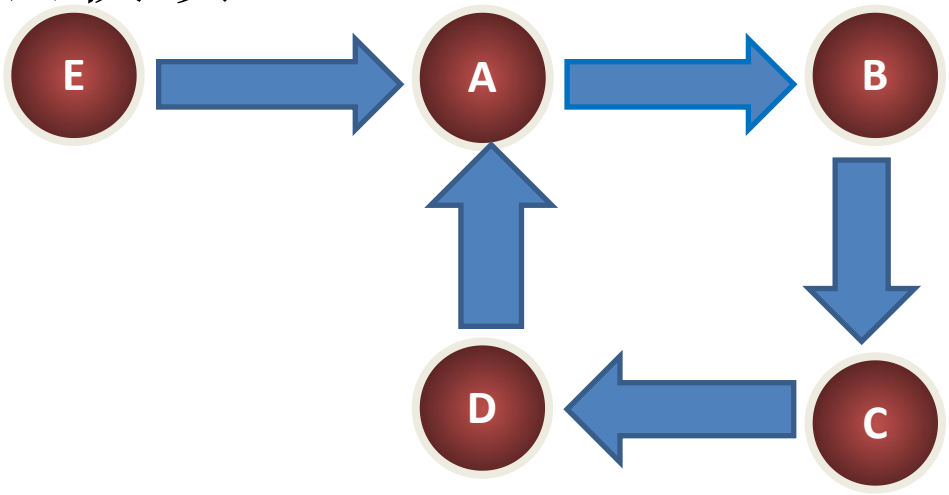
之所以能够成功收敛到非零值，是因为转移矩阵的这样一个性质：每列的加和为1。

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

解决方法：拿掉图中的Dead Ends节点及其相关的边，直到图中没有Dead Ends。对剩余部分计算rank,然后以拿掉Dead Ends逆向顺序反推Dead Ends的rank。

2，Rank Sink问题的解决

Rank Sink问题:在互联网的超链接结构中，一旦出现封闭的情况，就会使得网页的rank值无法收敛。



上图中例子就是封闭情况。

解决方法：

如果沿着网页的链接一直点下去，发现老是在同样的几个网页中徘徊，怎么办？

我们此时会将当前的页面关掉，再打开一个新的网页。

解决Rank Sink问题，具体的做法是加入“逃脱因子”或者称为“心灵转移”（teleporting)。

返回

此时，

向量迭代公式变为：

$$v' = (1-\beta)Mv + e\frac{\beta}{N}$$

β 被设定为一个比较小的参数（0.2或更小），e为N维单位向量。这样每一个页面就会拥有一个合理的rank值。

总结

通过以上的简述，以及例子的分析，我们将PageRank算法的核心思想总结如下：

“被越多优质的网页所指向的网页，它是优质的的概率就越大”。

实际上的PageRank算法没有这么简单，目前几乎所有现代搜索引擎页面权重的计算方法都基于PageRank及其变种。这里，我们只是简要地分析一下PageRank算法。