Google'PageRank算法

原文：《The PageRank Citation Ranking: Bringing Order to the Web》

一、基本思路

网页上的跳转链接让全部网站构成了一个有向图:

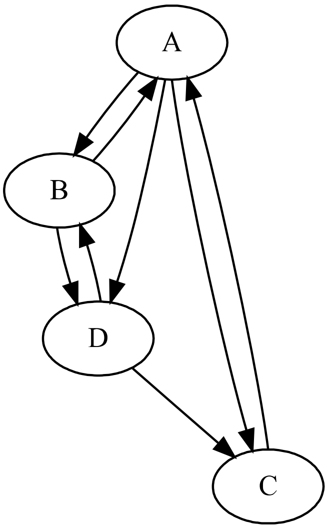

figure 1 网页间跳转

对于一个网站页面而言，指向该页面的入链越多，意味着该页面越重要，或者说该页面影响力越大。那么我们可以定义这样一个网站页面u影响力$PR(u)$:

$$PR(u) = \Sigma_{v \in B_u} \frac{PR(v)}{L(v)}$$

其中$B(u)$是入链集合，$L(v)$是页面v的出链数量，并且假设所有网页影响力总和为1（我们总可以乘以一个常数c以归一化）。在给出该定义后，我们可以考虑这样一个影响力向量:

$$\begin{aligned} R &= [PR(u_1), PR(u_2), ..., PR(u_n)]^T \\ &= [\Sigma_{v \in B_{u_1}} \frac{PR(v)}{L(v)}, \Sigma_{v \in B_{u_2}} \frac{PR(v)}{L(v)}, ..., \Sigma_{v \in B_{u_n}} \frac{PR(v)}{L(v)}]^T \end{aligned}$$

为了更统一的表示上述向量，我们引入一个δ函数:

$$\delta(u,v) = \begin{cases} 1 & u \rightarrow v \\ 0 & u \nrightarrow v \end{cases}$$

那么:

$$\Sigma_{v\in B_{u_k}}\frac{PR(v)}{L(v)} = \Sigma_{i=1}^{n}\delta(u_i,u_k)\frac{PR(u_i)}{L(u_i)}$$
$$= [\frac{\delta(u_1,u_k)}{L(u_1)},\frac{\delta(u_2,u_k)}{L(u_2)},...,\frac{\delta(u_n,u_k)}{L(u_n)}]\cdot[PR(u_1),PR(u_2),...,PR(u_n)]^T$$

对于影响力向量:

$$R = \begin{bmatrix} [\frac{\delta(u_1,u_1)}{L(u_1)},\frac{\delta(u_2,u_1)}{L(u_2)},...,\frac{\delta(u_n,u_1)}{L(u_n)}]\cdot[PR(u_1),PR(u_2),...,PR(u_n)]^T \\ [\frac{\delta(u_1,u_2)}{L(u_1)},\frac{\delta(u_2,u_2)}{L(u_2)},...,\frac{\delta(u_n,u_2)}{L(u_n)}]\cdot[PR(u_1),PR(u_2),...,PR(u_n)]^T \\ ... \\ [\frac{\delta(u_1,u_n)}{L(u_1)},\frac{\delta(u_2,u_n)}{L(u_2)},...,\frac{\delta(u_n,u_n)}{L(u_n)}]\cdot[PR(u_1),PR(u_2),...,PR(u_n)]^T \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\delta(u_1,u_1)}{L(u_1)},\frac{\delta(u_2,u_1)}{L(u_2)},...,\frac{\delta(u_n,u_1)}{L(u_n)} \\ \frac{\delta(u_1,u_2)}{L(u_1)},\frac{\delta(u_2,u_2)}{L(u_2)},...,\frac{\delta(u_n,u_2)}{L(u_n)} \\ ... \\ \frac{\delta(u_1,u_n)}{L(u_1)},\frac{\delta(u_2,u_n)}{L(u_2)},...,\frac{\delta(u_n,u_n)}{L(u_n)} \end{bmatrix}\cdot\begin{bmatrix} PR(u_1) \\ PR(u_2) \\ ... \\ PR(u_n) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\delta(u_1,u_1)}{L(u_1)},\frac{\delta(u_2,u_1)}{L(u_2)},...,\frac{\delta(u_n,u_1)}{L(u_n)} \\ \frac{\delta(u_1,u_2)}{L(u_1)},\frac{\delta(u_2,u_2)}{L(u_2)},...,\frac{\delta(u_n,u_2)}{L(u_n)} \\ ... \\ \frac{\delta(u_1,u_n)}{L(u_1)},\frac{\delta(u_2,u_n)}{L(u_2)},...,\frac{\delta(u_n,u_n)}{L(u_n)} \end{bmatrix}\cdot R$$

注意到这样一个形式正是马尔可夫链随机游走到收敛的形式，那么如果对于一个网站页面有向图的转移矩阵，如果能够收敛，那么我们可以给每个网站页面任意一个影响力初始值，满足总和为1，在足够多此转移后，我们将得到每个页面的影响力。
对于figure 1，我们可以写出其转移矩阵:

$$\begin{bmatrix} & A & B & C & D \\ A & 0 & \frac{1}{2} & 1 & 0 \\ B & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ C & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ D & \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

假设四个页面的初始影响力都是0.25，那么经过足够多次转移后，影响力将收敛到$PR(A)=\frac{1}{3},PR(B)=\frac{2}{9},PR(C)=\frac{2}{9},PR(D)=\frac{2}{9}$。

二、存在的问题

上述模型是PageRank算法的核心，但是仍然存在几个问题:

1.Rank Leak

如果一个网页只有入链，没有出链，那么毫无疑问，在每一轮游走中，该页面的影响力将增大，对于整个有向图，所有的影响力都将最终流到这些只有入链的网页上，其他网页PR值将归于0。

2.Rank Sink

如果一个网页只有出链，没有入链，那么该网页的PR值必将归于0。

三、随机浏览模型

添加一个新的假设：用户并不都按照跳转链接来访问其他网页，还有可能直接访问其他网页（即访问不存在直接连接的网页），采用这种访问方式的概率很小。
通过该假设修正PR的定义：

$$PR(u) = \frac{1-d}{n} + d \cdot \Sigma_{v \in B_u} \frac{PR(v)}{L(v)}$$

基于该修正模型，新的影响力向量R'满足：

$$\begin{aligned}
R' &= d \cdot R + \frac{1-d}{n} \cdot [1, 1, ..., 1]^T \\
&= d \cdot \begin{bmatrix} \frac{\delta(u_1,u_1)}{L(u_1)}, \frac{\delta(u_2,u_1)}{L(u_2)}, ..., \frac{\delta(u_n,u_1)}{L(u_n)} \\ \frac{\delta(u_1,u_2)}{L(u_1)}, \frac{\delta(u_2,u_2)}{L(u_2)}, ..., \frac{\delta(u_n,u_2)}{L(u_n)} \\ ... \\ \frac{\delta(u_1,u_n)}{L(u_1)}, \frac{\delta(u_2,u_n)}{L(u_2)}, ..., \frac{\delta(u_n,u_n)}{L(u_n)} \end{bmatrix} \cdot \begin{bmatrix} PR(u_1) \\ PR(u_2) \\ ... \\ PR(u_n) \end{bmatrix} + \frac{1-d}{n} \cdot [1, 1, ..., 1]^T \\
&= d \cdot \begin{bmatrix} \frac{\delta(u_1,u_1)}{L(u_1)}, \frac{\delta(u_2,u_1)}{L(u_2)}, ..., \frac{\delta(u_n,u_1)}{L(u_n)} \\ \frac{\delta(u_1,u_2)}{L(u_1)}, \frac{\delta(u_2,u_2)}{L(u_2)}, ..., \frac{\delta(u_n,u_2)}{L(u_n)} \\ ... \\ \frac{\delta(u_1,u_n)}{L(u_1)}, \frac{\delta(u_2,u_n)}{L(u_2)}, ..., \frac{\delta(u_n,u_n)}{L(u_n)} \end{bmatrix} \cdot R' + \frac{1-d}{n} \cdot E \cdot R' \\
&= (d \cdot M + \frac{1-d}{n} \cdot E) \cdot R'
\end{aligned}$$

其中E是一个$n \times n$的全1矩阵，那么我们就得到了新模型的转移矩阵M'：

$$M' = d \cdot M + \frac{1-d}{n} \cdot E$$