PageRank排名思想

链入数大的节点就一定更重要吗？

- 被北大主页指向 VS 被你的闺蜜指向
- 一言九鼎 VS 攒鸡毛凑掸子

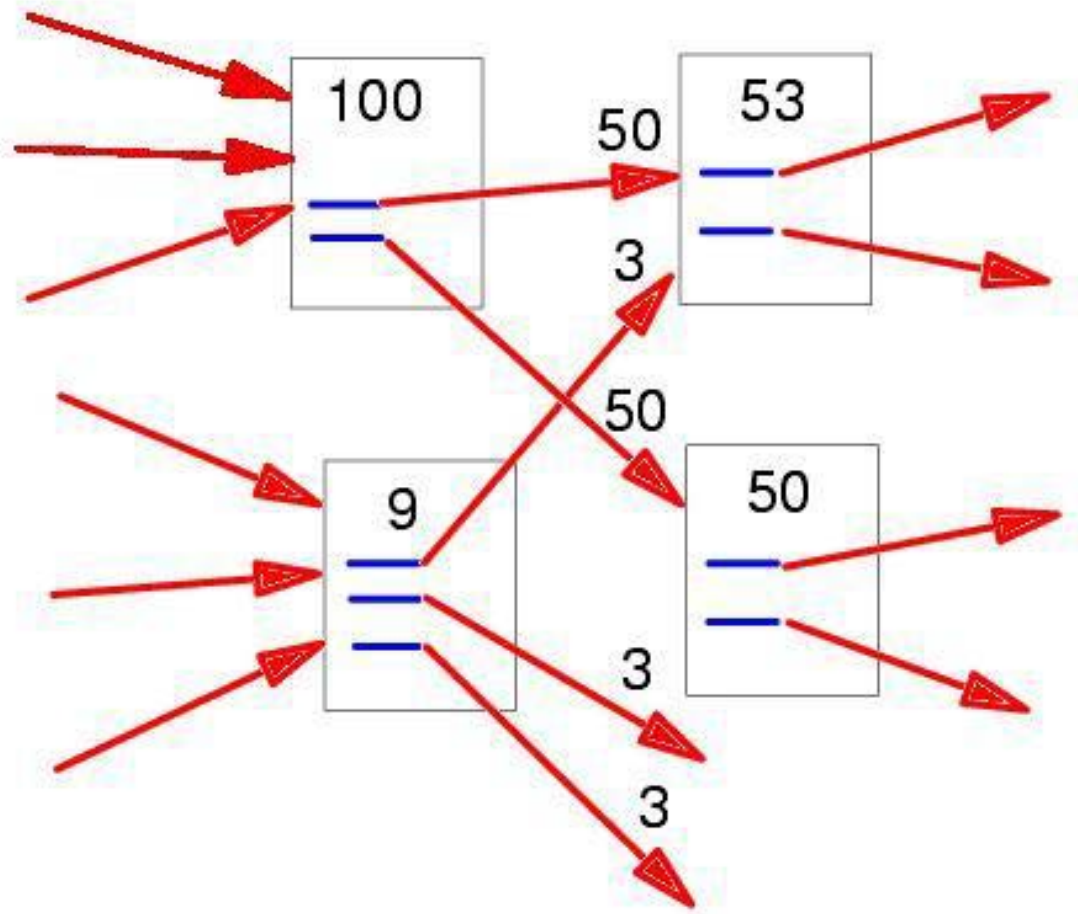
一个网页要想拥有较高的PR值

① 有很多网页链接到它

② 有高质量网页链接到它

PageRank：被越多优质网页所指的网页是优质网页的概率越大

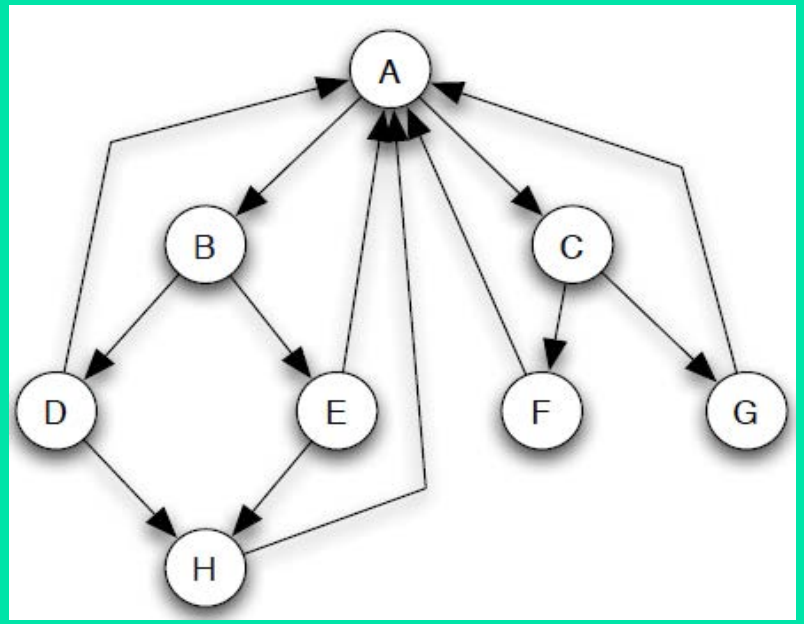
PageRank的计算过程



$$R(P_i) = \sum_{P_j \in B_i} \frac{R(P_j)}{L_j}$$

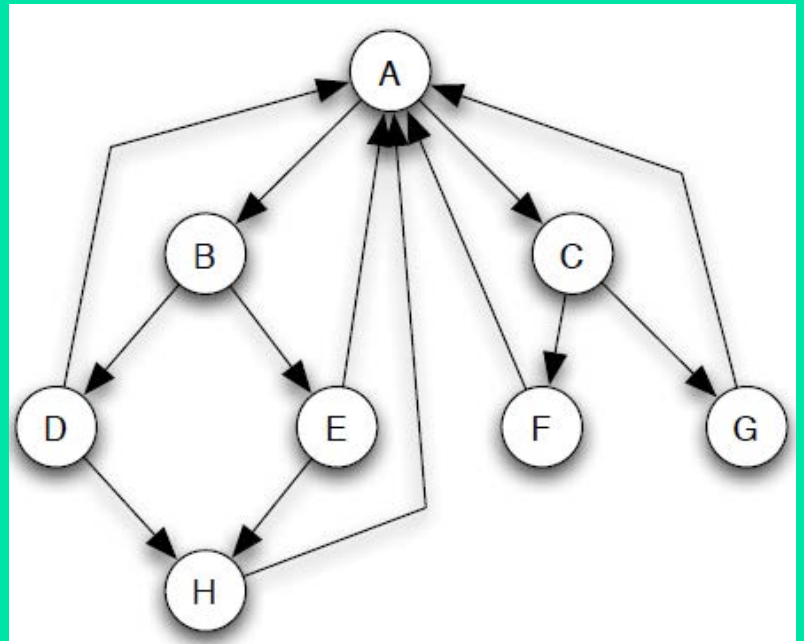
PageRank的计算过程

		A	B	C	D	E	F	G	H
$A:BC$	$\frac{1}{8}$								
$B:DE$	$\frac{1}{8}$								
$C:FG$	$\frac{1}{8}$								
$D:AH$	$\frac{1}{8}$								
$E:AH$	$\frac{1}{8}$								
$F:A$	$\frac{1}{8}$								
$G:A$	$\frac{1}{8}$								
$H:A$	$\frac{1}{8}$								



PageRank的计算过程

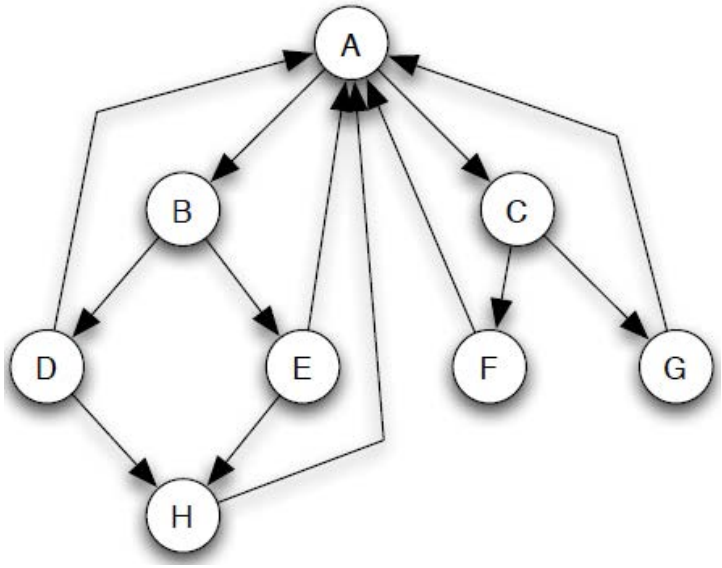
		A	B	C	D	E	F	G	H
$A:BC$	$\frac{1}{2}$								
$B:DE$	$\frac{1}{16}$								
$C:FG$	$\frac{1}{16}$								
$D:AH$	$\frac{1}{16}$								
$E:AH$	$\frac{1}{16}$								
$F:A$	$\frac{1}{16}$								
$G:A$	$\frac{1}{16}$								
$H:A$	$\frac{1}{8}$								



使用SQL计算PageRank

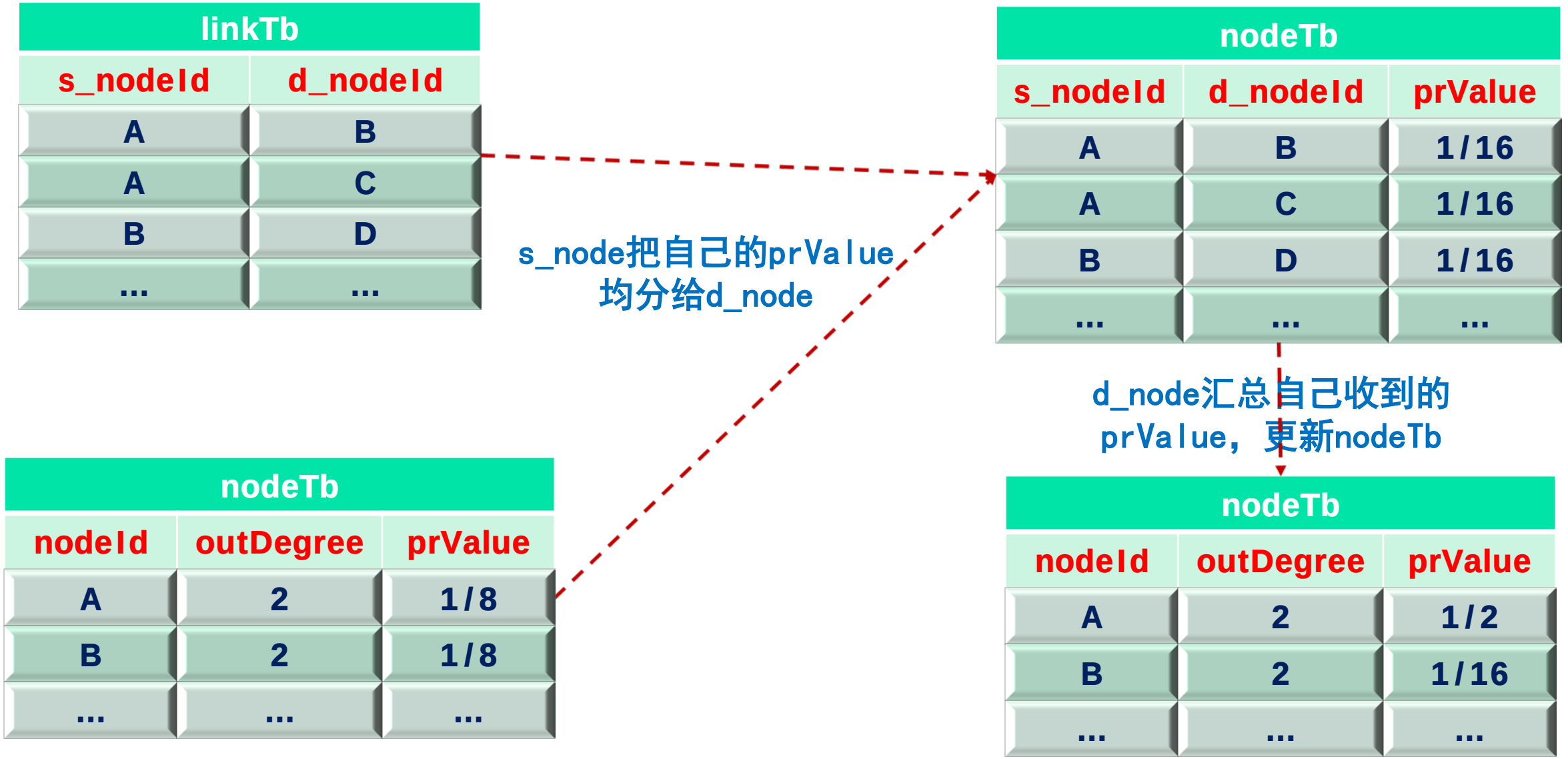
linkTb	
s_nodeId	d_nodeId
A	B
A	C
B	D
B	E
C	F
C	G
D	A
D	H
E	A
E	H
F	A
G	A
H	A

初始化

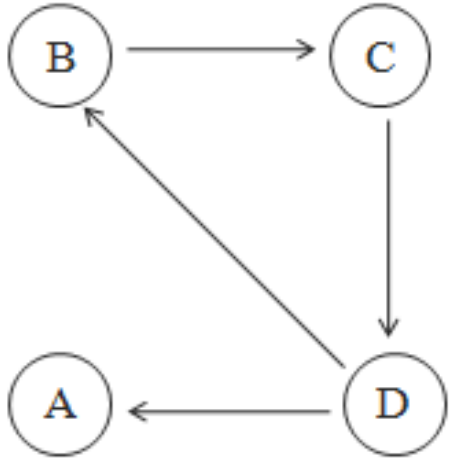


nodeTb		
nodeId	outDegree	prValue
A	2	1/8
B	2	1/8
C	2	1/8
D	2	1/8
E	2	1/8
F	1	1/8
G	1	1/8
H	1	1/8

使用SQL计算PageRank的一次迭代过程



Rank leak



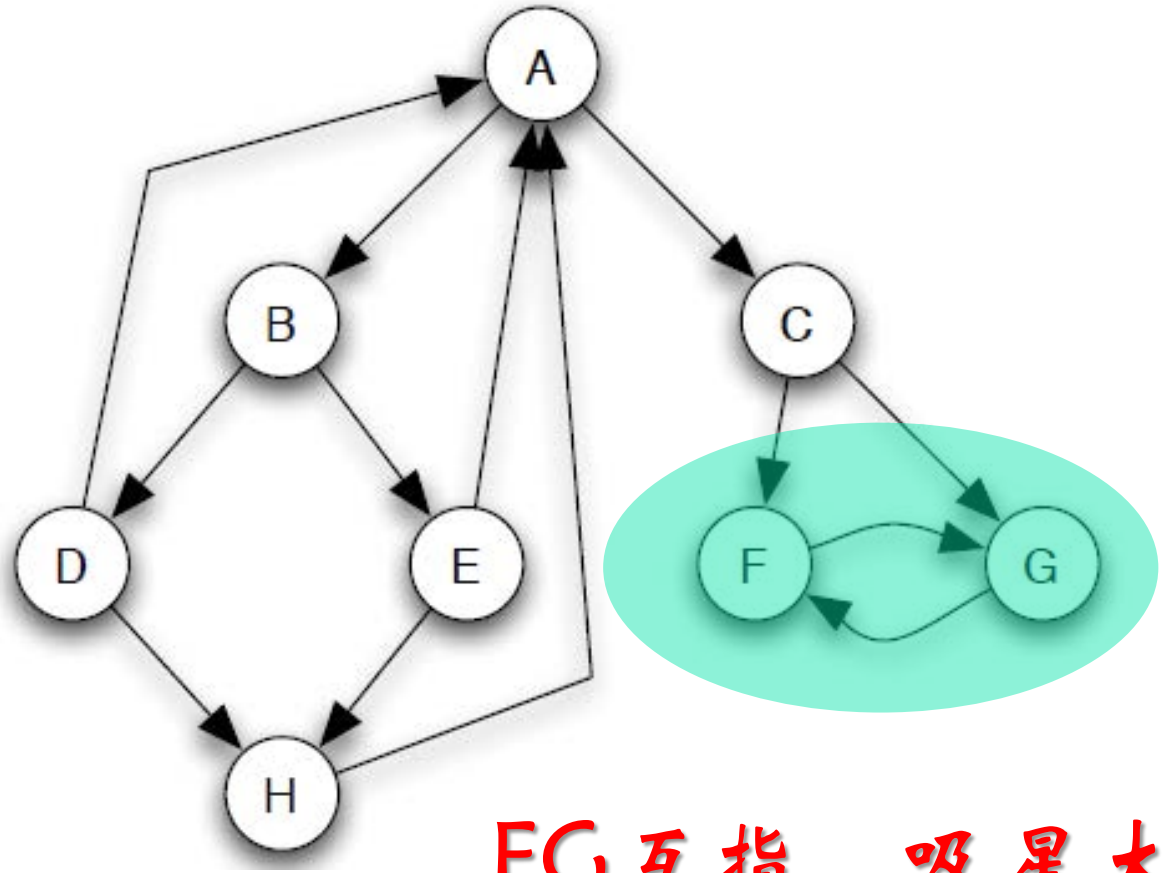
	PR(A)	PR(B)	PR(C)	PR(D)
初始	0.25	0.25	0.25	0.25
一次迭代	0.125	0.125	0.25	0.25
二次迭代	0.125	0.125	0.125	0.25
三次迭代	0.125	0.125	0.125	0.125
…	…	…	…	…
n次迭代	0	0	0	0

Rank leak：一个独立的网页如果没有外出的链接就产生等级泄漏

解决办法：

1. 将无出度节点递归的从图中去掉，待其他节点计算完后再添加
2. 对无出度的节点添加一条边，指向那些指向它的顶点

排名算法中的北冥神功



Rank sink：

整个网页图中的一组紧密链接成环的网页如果没有外出的链接就产生Rank sink

引入随机浏览模型

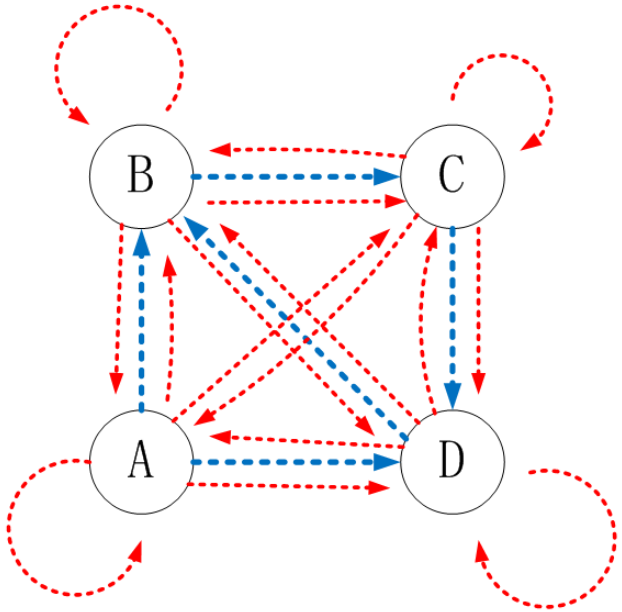
FG互指，吸星大法

排名值慢泄露

PageRank的随机浏览模型

冲浪模型：

随机上网者访问一个新网页的概率等于这个网页的PageRank值



在每个顶点处以概率d按原来蓝色方向转移，以概率1-d按红色方向转移

$$PR(P_i) = \frac{1-d}{N} + d\sum_{P_j \in M(P_i)} \frac{PR(P_j)}{L_j}$$