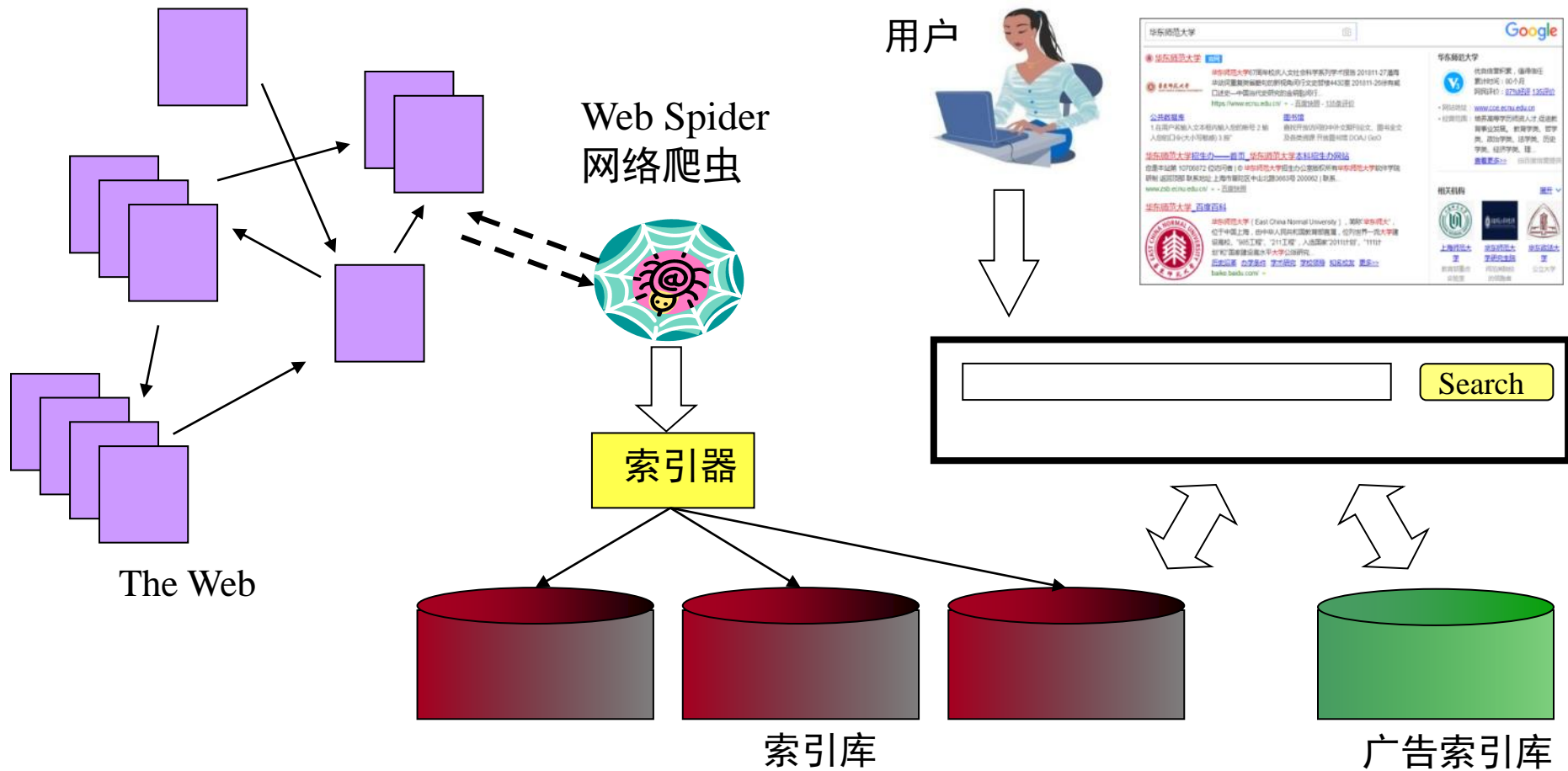




# 数据科学导论

---

Introduction to Data Science





华东师范大学



登录

[全部](#) [图片](#) [地图](#) [新闻](#) [视频](#) [更多](#)

[设置](#) [工具](#)

已启用安全

找到约 12,500,000 条结果 (用时 0.70 秒)

[www.ecnu.edu.cn](http://www.ecnu.edu.cn) ▾

华东师范大学

华东师范大学等四校“组团式”援建丽江师专. RightLeft. 热点新闻. >> » 跨学科合作 华东师大青年学者在超分子光化学领域取得重要进展. 12345 ...

[院系设置](#) · [华东师范大学留学生办公室](#) · [大夏学堂](#) · [校情简介](#)

[baike.baidu.com](http://baike.baidu.com) > item > 华东师范大学 ▾

华东师范大学\_百度百科

华东师范大学 (East China Normal University), 简称“华东师大”, 位于上海市, 由中华人民共和国教育部直属, 是教育部和上海市人民政府重点共建的综合性研究 ...

本科专业: 83个

地址: 中山北路校区: 上海市普陀区中山...

院系设置: 3个学部; 30个全日制学院; 4个...

院校代码: 10269

[改革发展](#) · [院系专业](#) · [教学建设](#) · [科研机构](#)

[zh.wikipedia.org](http://zh.wikipedia.org) > zh-hans > 华东师范大学 ▾

华东师范大学- 维基百科, 自由的百科全书

网页排序结果

其他  
相关  
信息



华东师范大学

网站

路线

保存

中国上海市的大学

华东师范大学, 简称华师大或华东师大, 是一所位于中国上海市的公立研究型大学, 1951年由大夏大学和光华大学合并而成, 是中华人民共和国创办的第一所师范大学。在1952年院系调整中, 圣约翰大学主要院系并入。1972年更名为上海师范大学, 1980年恢复原校

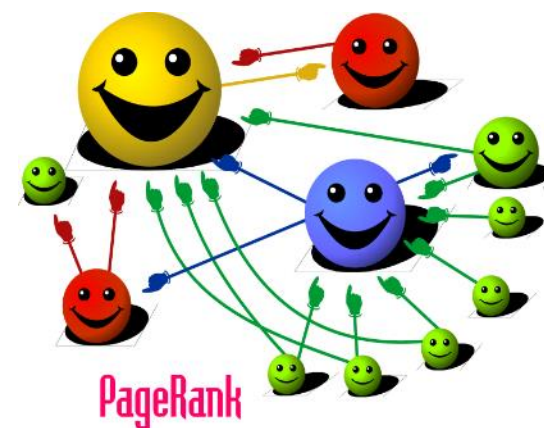
问题: 搜索引擎怎么知道哪个网页排在前面, 哪个排在后面呢? 即如何衡量网页的重要性?

Google - PageRank

- Google 的 PageRank 是基于这样一个理论：



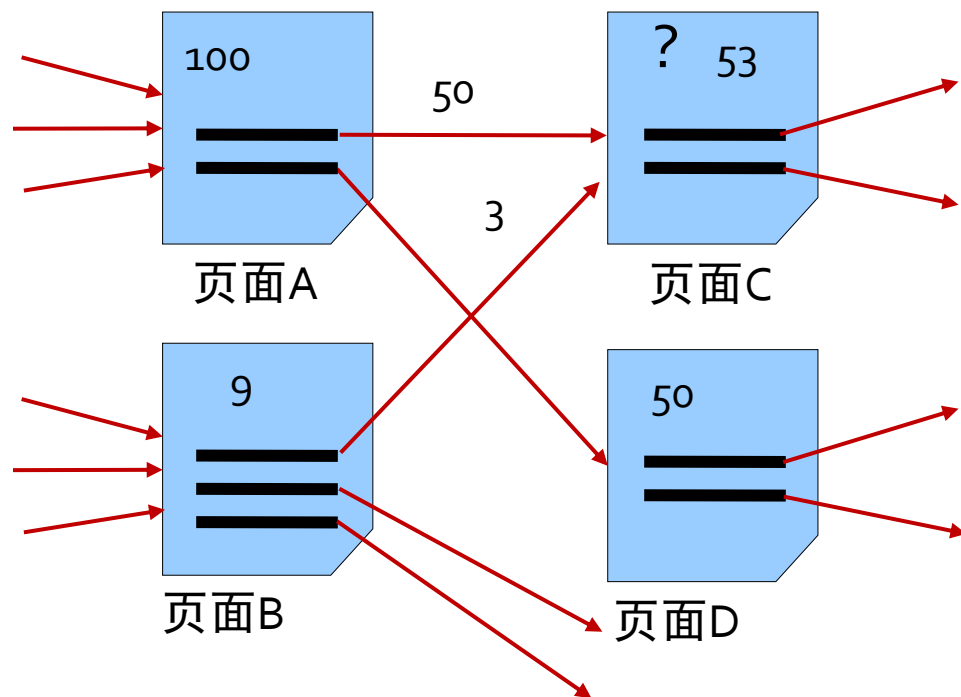
- 若 B 网页上有连接到 A 网页的链接，说明 B 认为 A 有链接价值，是一个“重要”的网页
- 一个网页的重要性大致由下面两个因素决定：
  - 该网页的导入链接的数
  - 这些导入链接的重要性



PageRank的决定因素

Google - PageRank



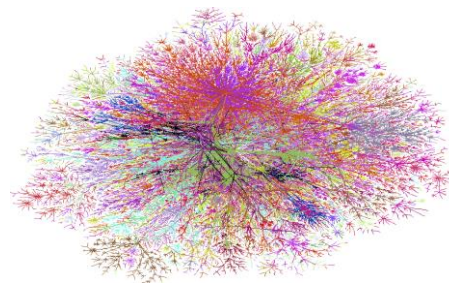


可以用数据思维与模型来解决这类问题

尝试计算PageRank值

### • 问题

- 先有鸡还是先有蛋?
- Internet的拓扑结构



Google - PageRank

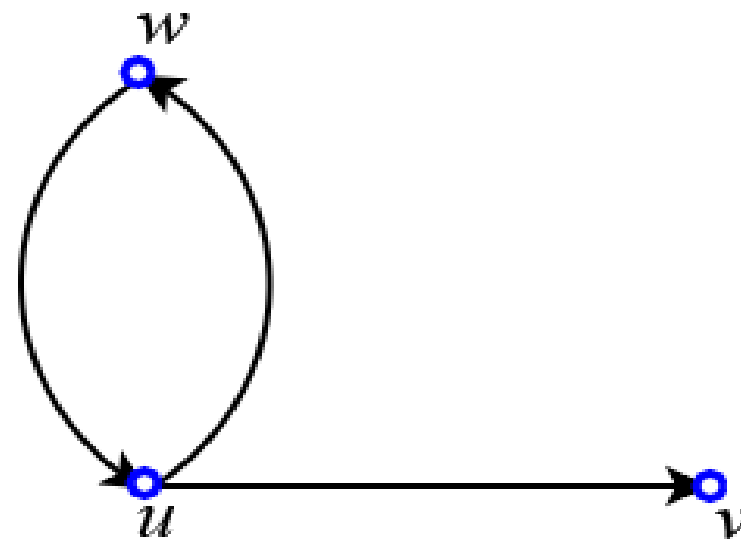
右图为一个有向图，记为  $G$ ,  $G = \{V, E\}$

顶点组成的集合:  $V = \{u, v, w\}$

弧组成的集合:  $E = \{(u, w), (w, u), (u, v)\}$

顶点  $u$  的出度:  $od(u) = 2$

顶点  $u$  的入度:  $id(u) = 1$



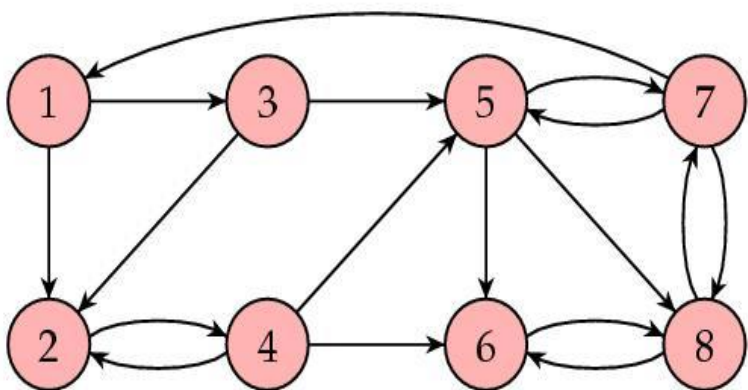
如何表示这个图，以便更好计算PageRank值呢？

有向图

Google - PageRank

$$A = (a_{ij})$$

$$a_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{otherwise} \end{cases}$$

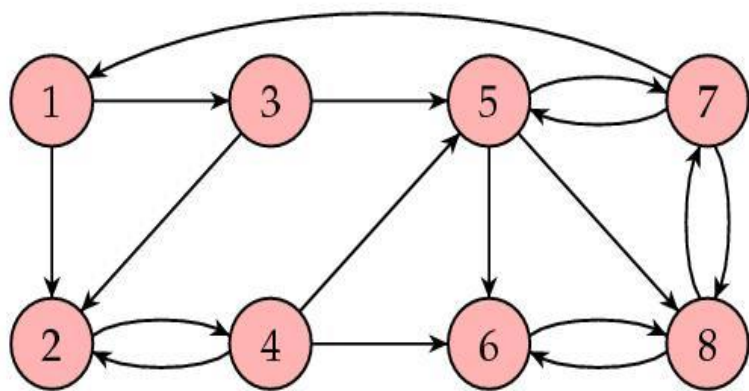


$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

邻接矩阵

Google - PageRank

进一步，如果将邻接矩阵中的元素除以对应节点的出度，可以得到该图的**超链接矩阵**



$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{3} & 0 \end{bmatrix}$$

• 超链接矩阵的特点：

- 所有元素非负
- 每列元素的总和为1

随机矩阵 (Stochastic Matrix)

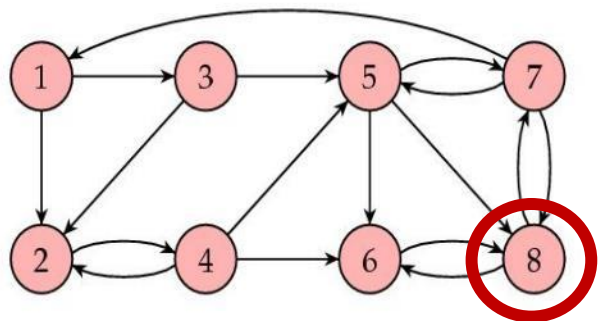
马尔可夫矩阵

超链接矩阵

Google - PageRank



定理：超链接矩阵H的最大特征向量即为该矩阵的PageRank值



$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{3} & 0 \end{bmatrix}$$

$$I = \begin{bmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ \boxed{0.2950} \end{bmatrix}$$

$I$  是  $H$  的对应于特征值  $\lambda=1$  的特征向量

数学的奇妙：原来不知如何下手的互联网网页的排序问题，现在已经轻而易举地变成了求解矩阵H的特征向量问题

矩阵的特征向量与特征值

Google - PageRank

幂迭代方法  $\longrightarrow I^{k+1} = H \cdot I^k$

$I^0$	$I^1$	$I^2$	$I^3$	$I^4$	...	$I^{60}$	$I^{61}$	$I =$	<b>0.0600</b>
1	0	0	0	0.0278	...	0.06	<b>0.0600</b>		<b>0.0675</b>
0	0.5	0.25	0.1667	0.0833	...	0.0675	<b>0.0675</b>		<b>0.0300</b>
0	0.5	0	0	0	...	0.03	<b>0.0300</b>		<b>0.0675</b>
0	0	0.5	0.25	0.1667	...	0.0675	<b>0.0675</b>		<b>0.0975</b>
0	0	0.25	0.1667	0.1111	...	0.0975	<b>0.0975</b>		<b>0.2025</b>
0	0	0	0.25	0.1806	...	0.2025	<b>0.2025</b>		<b>0.1800</b>
0	0	0	0.0833	0.0972	...	0.18	<b>0.1800</b>		<b>0.2950</b>
0	0	0	0.0833	0.3333	...	0.295	<b>0.2950</b>		

如何计算PageRank值

- 第一步：将互联网作为一个有向图，并用邻接矩阵进行表示；
- 第二步：将该邻接矩阵转换为超链接矩阵；
- 第三步：求解该超链接矩阵的最大特征向量（如幂迭代法）；
- 第四步：求得的特征向量中的值即为对应网页的PageRank值。



PageRank 算法

Google - PageRank



- ❑ 这一漂亮的想法出自于Stanford大学1998年  
在读博士研究生 *Larry Page* 和 *Sergey Brin*
- ❑ 第七次国际World Wide Web会议(WWW'98)上的论文  
*The PageRank citation ranking : Bringing order to the Web*
- ❑ PageRank 算法中使用的数学知识包括：矩阵  
的性质、特征值和特征向量、幂迭代方法等

PageRank 算法

Google - PageRank

- L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, *Technical Report*, Stanford University, 1998.
- K. Bryan, T. Leise, The \$25,000,000,000 eigenvector: The linear algebra behind Google, *SIAM Review*, 48 (3), 569-81, 2006.
- P. Berkin, A survey on PageRank computing, *Internet Mathematics*, 2:73–120, 2005.