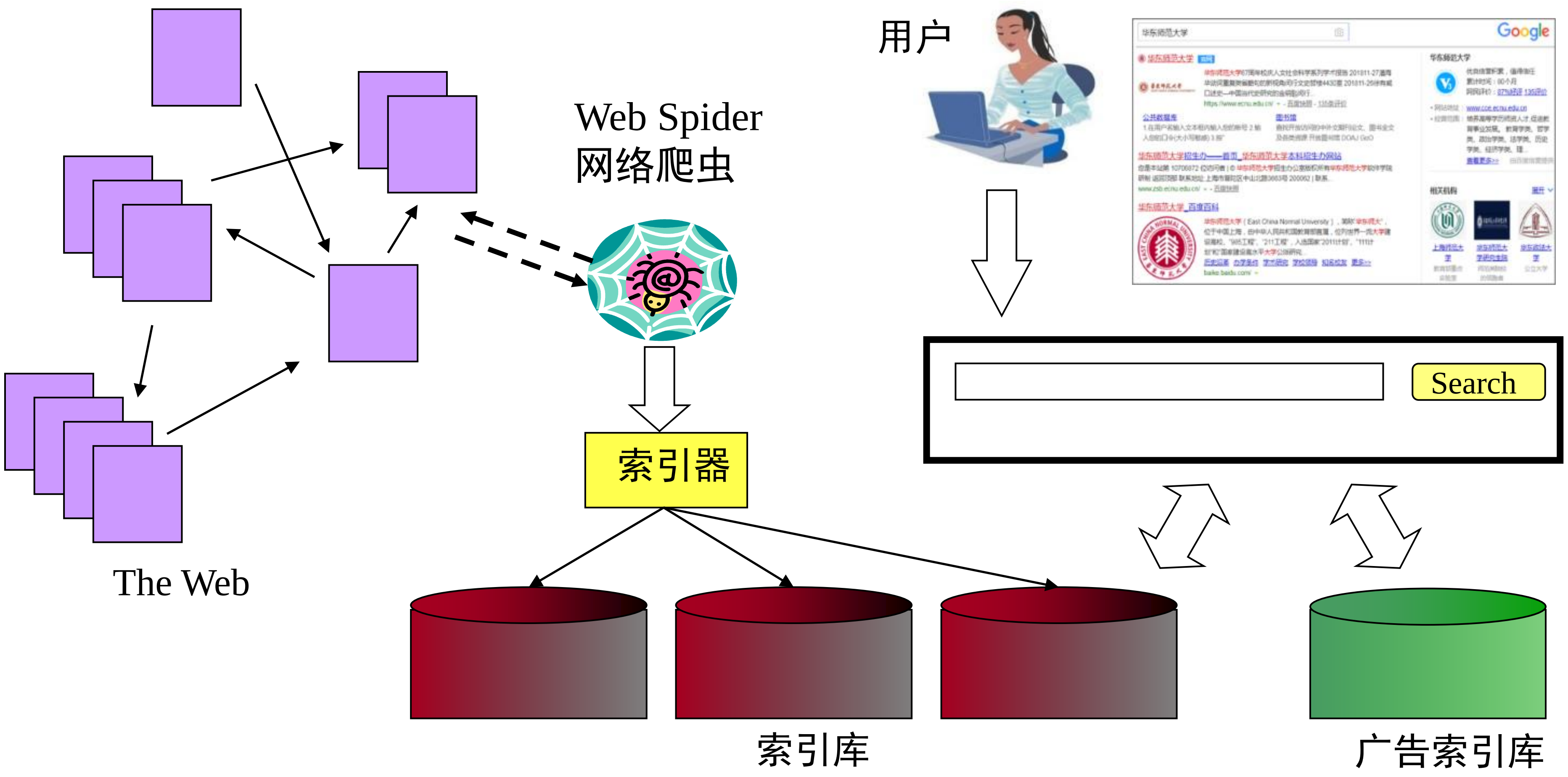
数据科学与工程导论

Introduction to Data Science and Engineering

开篇实例



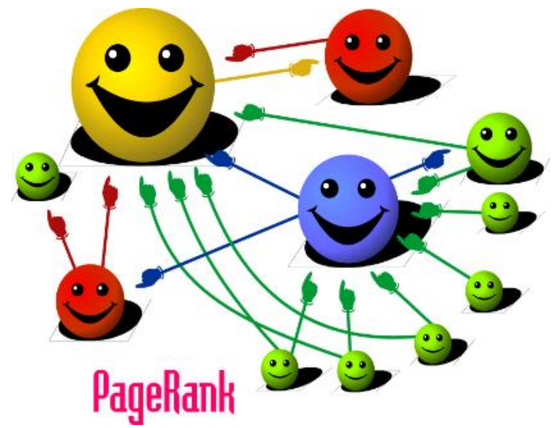
Google - PageRank

开篇实例

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？

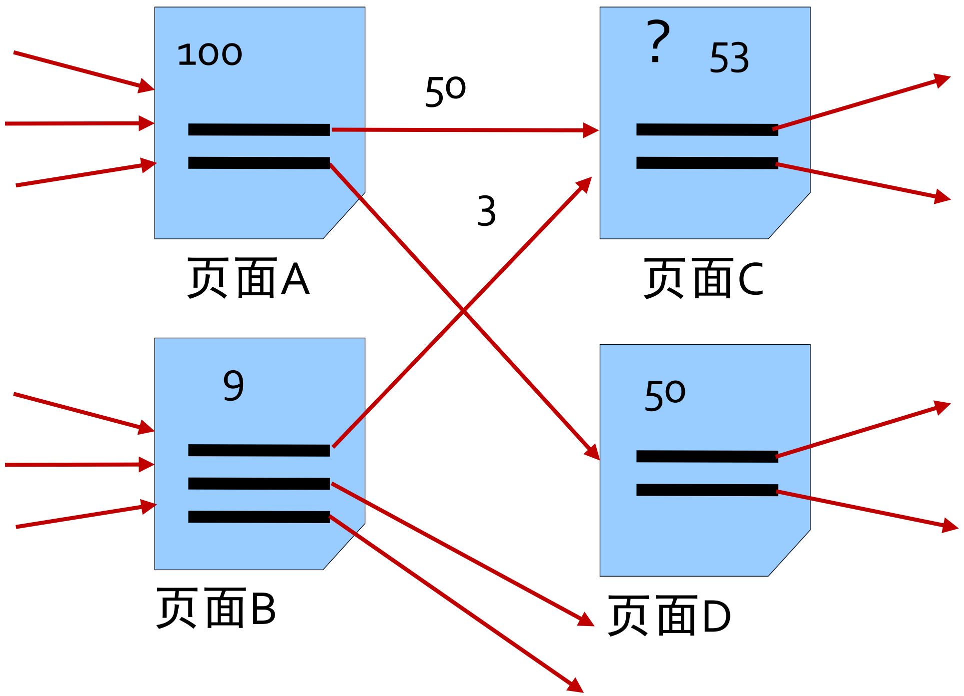
Google - PageRank

- 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

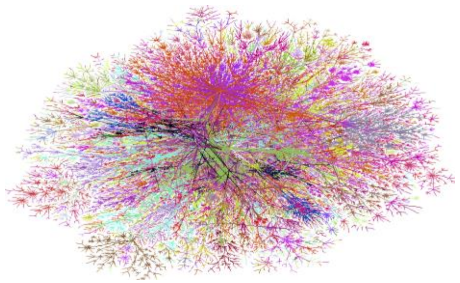


PageRank的决定因素

Google - PageRank



- 问题
  - 先有鸡还是先有蛋？
  - Internet的拓扑结构



可以用数据思维与模型来解决这类问题

尝试计算PageRank值

Google - PageRank

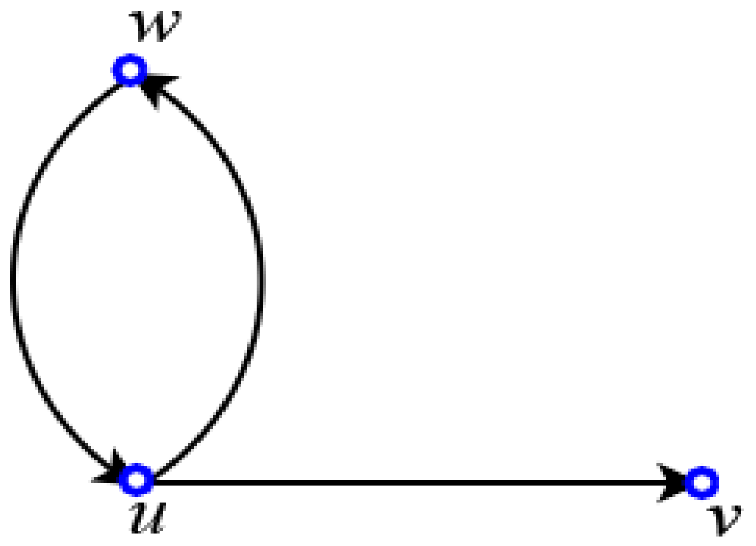
右图为一个有向图，记为 G，G={V,E}

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



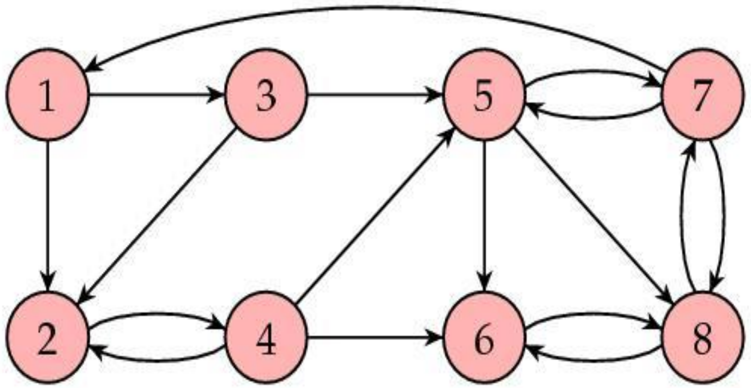
如何表示这个图，以便更好计算PageRank值呢？

有向图

Google - PageRank

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

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

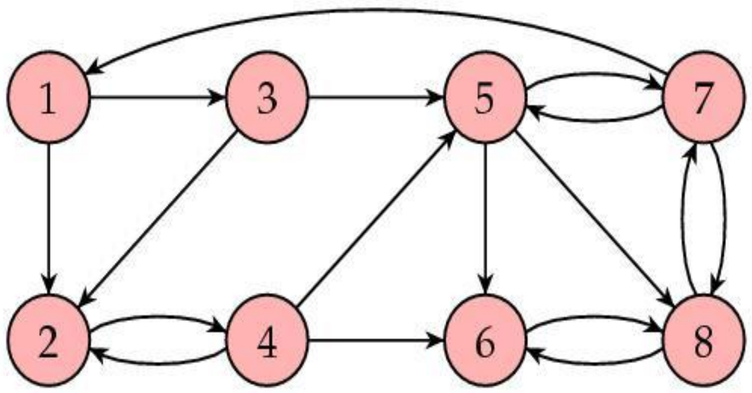


$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

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

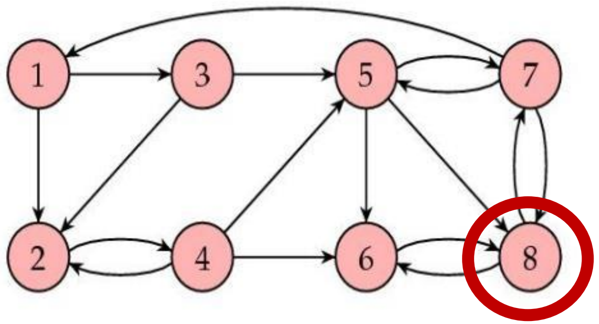
随机矩阵 (Stochastic Matrix)

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 λ=1 的特征向量

数学的奇妙：原来不知如何下手的互联网页的排序问题，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 ⟶ $I^{k+1} = H \cdot I^{k}$

I^0	I^1	I^2	I^3	I^4	…	I^{60}	I^{61}
1	0	0	0	0.0278	…	0.06	**0.0600**
0	0.5	0.25	0.1667	0.0833	…	0.0675	**0.0675**
0	0.5	0	0	0	…	0.03	**0.0300**
0	0	0.5	0.25	0.1667	…	0.0675	**0.0675**
0	0	0.25	0.1667	0.1111	…	0.0975	**0.0975**
0	0	0	0.25	0.1806	…	0.2025	**0.2025**
0	0	0	0.0833	0.0972	…	0.18	**0.1800**
0	0	0	0.0833	0.3333	…	0.295	**0.2950**

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

如何计算PageRank值

Google - PageRank

- 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.

参考文献