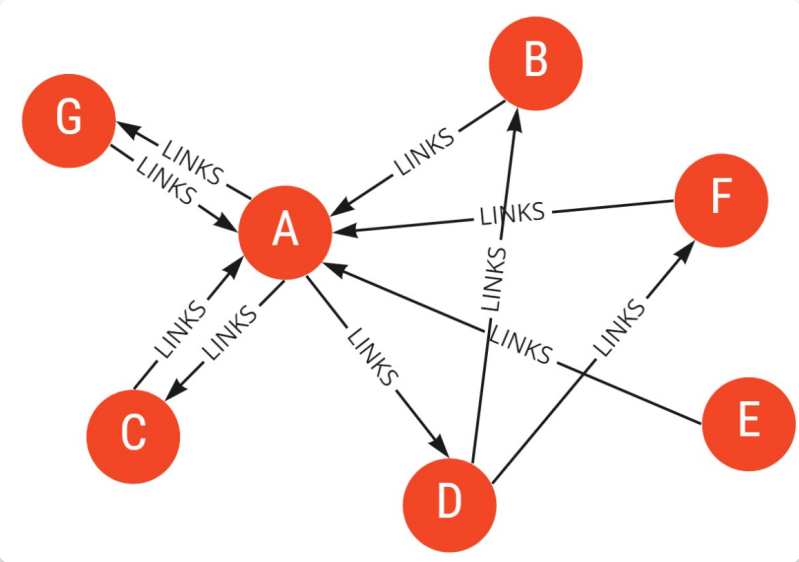
lec18.1 PageRank Algorithm Main Idea

PageRank算法简介

1. 提出时间与提出者：
   - `PageRank` 算法由谷歌的创始人谢尔盖·布林和拉里·佩奇于1998年提出。这一算法是谷歌早期成功的核心技术之一
2. 算法用途：
   - `PageRank` 被用于谷歌搜索引擎，用来对网页进行排序，确定它们在搜索结果中的排名。这是通过评估网页的重要性来实现的
3. 计算原理：
   - `PageRank` 算法使用网络图（web graph），其中节点代表网页，有向边表示网页之间的链接。通过分析这些链接结构，算法可以确定每个网页的重要性
   - 算法会迭代计算每个网页的 `PageRank` 值，直到达到收敛
4. 静态排名：
   - `PageRank` 值是静态的，意味着它在离线计算完成后是固定的。这个值不依赖于具体的搜索查询，而是表示网页在整个网络中的相对重要性



术语解释

入链（In-links）

- 定义：指向页面 a 的超链接
- 详细描述：
  - 入链是从其他页面指向页面 a 的超链接
  - 通常情况下，来自同一个页面的超链接会被忽略（即，一个页面上多次指向同一目标的链接不会重复计算）

出链（Out-links）

- 定义：从页面 a 指向其他页面的超链接。

- 详细描述：
  - 出链是从页面 a 指向其他页面的超链接
  - 通常情况下，指向同一网站内其他页面的超链接会被忽略（即，一个网站内的页面相互链接可能不被计算在内）

PageRank算法的基本思想

定义

`PageRank` 得分定义为：

$$P(a) = \sum_{(x,a)\in E} \frac{P(x)}{O_x}$$

其中：

- $P(a)$：页面 a 的 `PageRank` 得分
- O_a：页面 a 的出链（Out-links）数量
- E：图的有向边集合

这意味着页面 a 的 `PageRank` 得分是所有指向 a 的页面 x 的 PageRank 得分除以页面 x 的出链数量的总和

PageRank 的数学表示

顶点表示

- 设 $1, \ldots, n$ 为图的顶点
- 设 P 是一个 n 维列向量，表示各顶点的 `PageRank` 得分：

$$P = (P(1), \ldots, P(n))^T$$

修改后的邻接矩阵

设 A 为图的修改后的邻接矩阵，有

$$\begin{cases} A_{ij} = \frac{1}{O_i}, & \text{if } (i,j) \in E \\ A_{ij} = 0, & \text{otherwise} \end{cases}$$

矩阵形式的系统

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j},\ i = 1, \ldots, n$$
$$P = A^T P$$

解向量 P 是特征值为 1 的特征向量

迭代

如果 A 满足某些条件，那么 1 是其最大的特征值，解向量 P 可以通过幂迭代法找到

- 初始设置 P_0
- 迭代计算 $P_i = A^T P_{i-1}$

- 直到 $||P_i - P_{i-1}||_1 \leq \epsilon$，其中 ϵ 是预设的阈值

缺陷

对于真实的网页图，条件并不总是满足，因此需要进一步改进算法来处理实际情况