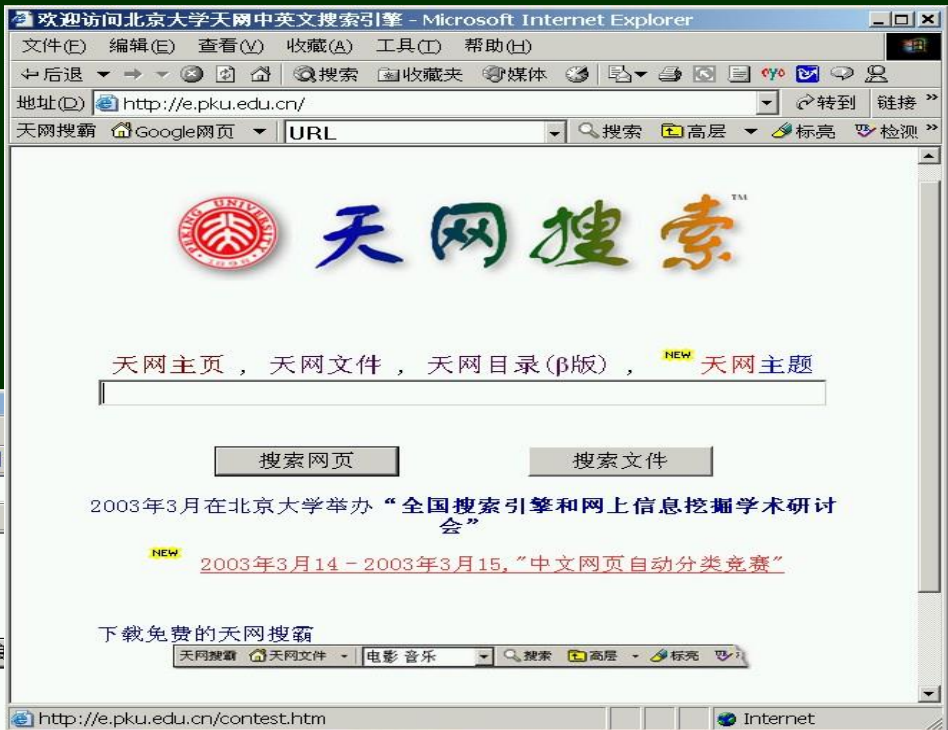


中枢与权威



搜索引擎关心的基本问题

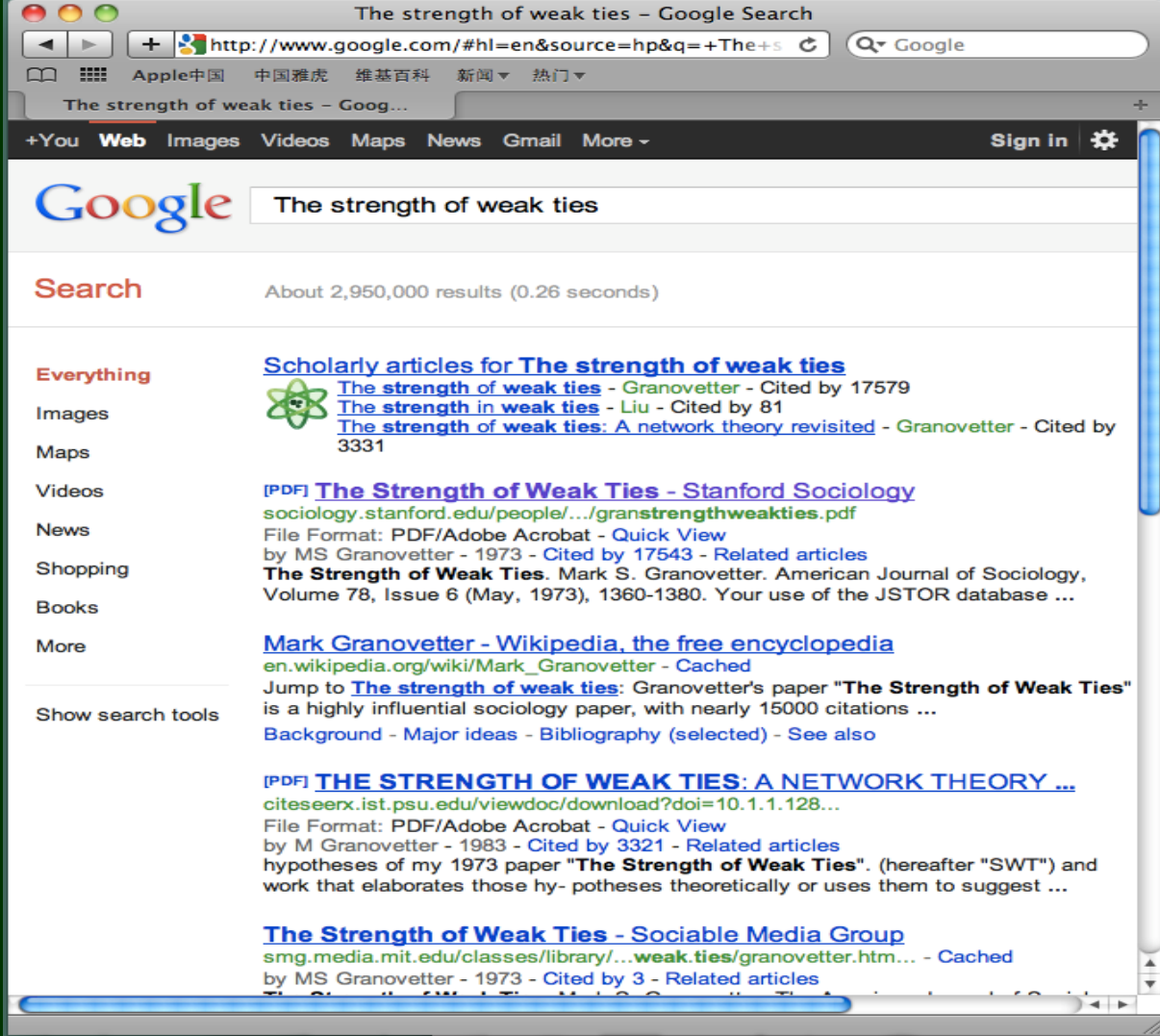
- 计算机显示屏一次只能显示5-6个结果，典型搜索引擎掌握的网页超过10亿
- 对用户提交的一个查询，如何从这种海量网页集合中将最可能满足用户需求的少数几个结果找出来，展现在计算机显示屏上？

传统信息检索（IR）技术的要点

- 基于词语之间的相关性（relevance）
- 传统应用背景
 - 文档集合：图书，规范的文献
 - 查 询：主题词，关键词
 - 查询意图：获取与查询词有关的书籍和文章
 - 用 户：图书管理人员
- “查询目标包含查询词”是一个合理假设
 - 在形成查询词的时候就有这样的潜意识

现在查找学术文献有类似预期

- 但人们在网络上不光是要找“文献”，而是多方面意义的“信息”
- 例如，人们给出“北京大学”查询词，多数会有什么预期？
- 查询“大学”呢？（意图会相当多样化）



查找某些非文献信息呢？

- 主页放在最前面，一定不是因为其中包含许多“北京大学”字样
- 很可能是由于许多包含“北京大学”字样的网页指向它
 - 利用链接中隐含的信息



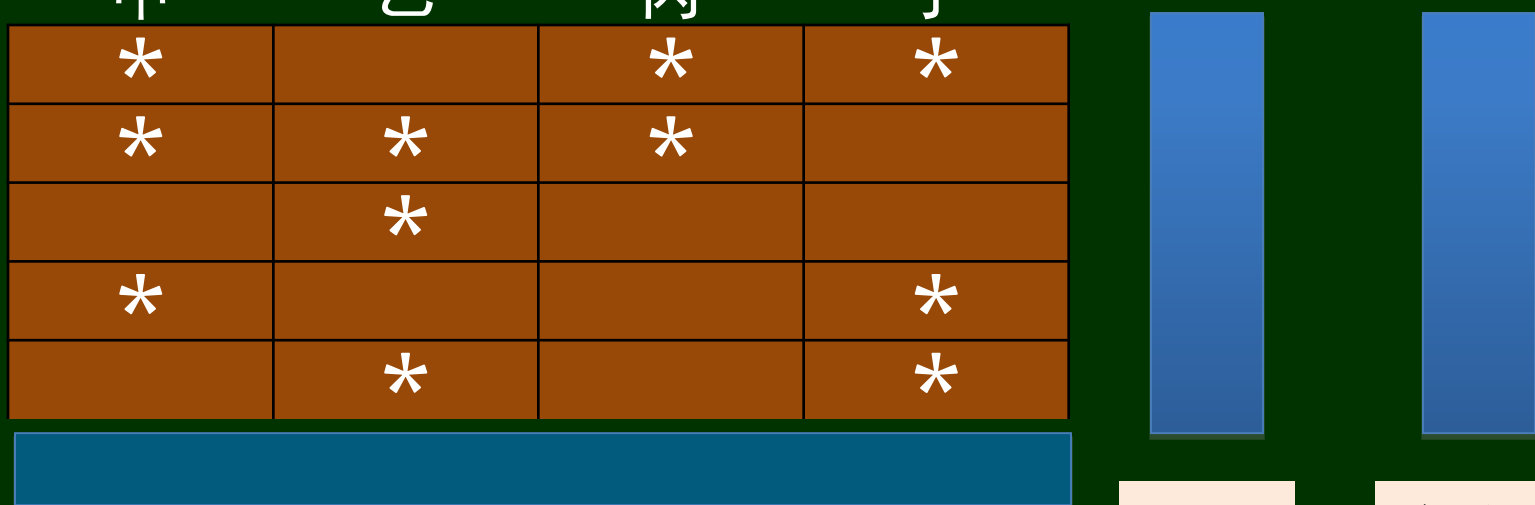
有效利用链接关系蕴含的信息，是搜索引擎超越传统信息检索系统、技术进步的最重要标志

- Web page之间的链接有两层含义：关系，描述

餐馆推荐问题

新辣道
海底捞
麦当劳
五方院
俏江南

甲	乙	丙	丁
*		*	*
*	*	*	
	*		
*			*
	*		*



看推荐人的“水平”

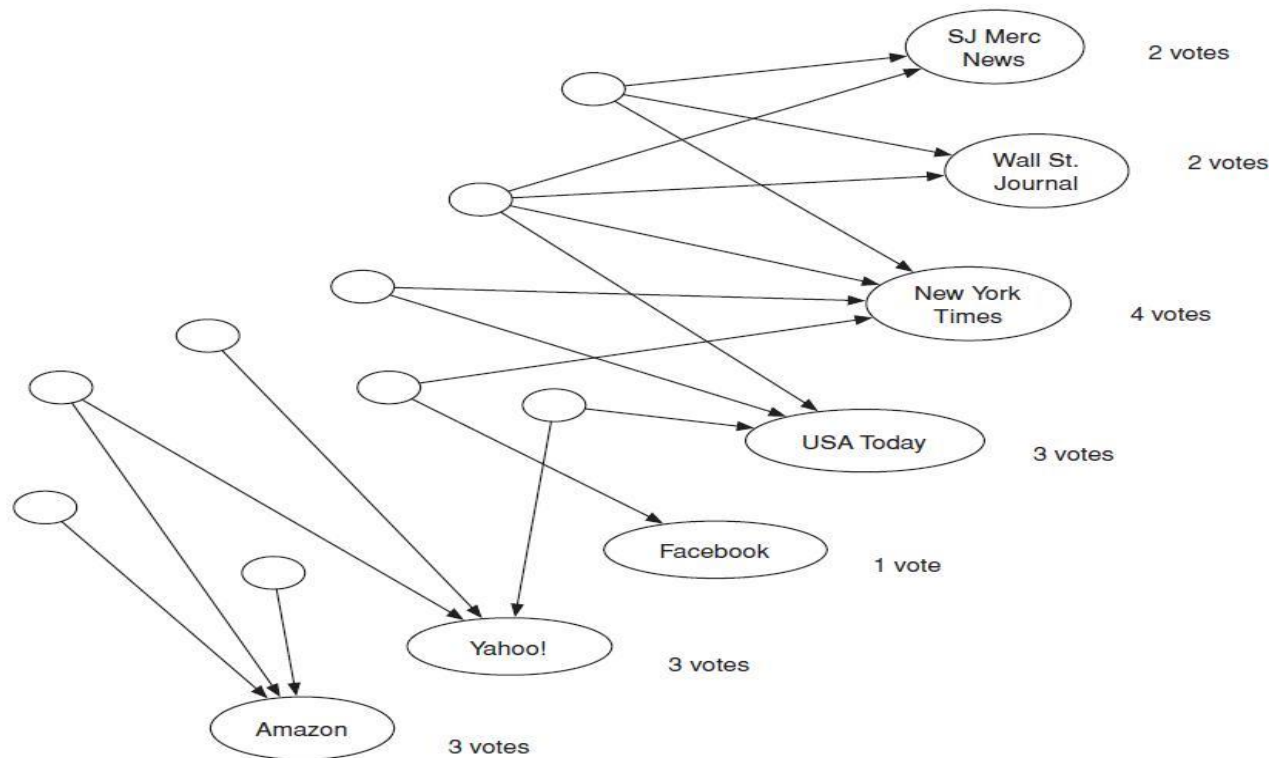
不能
完全
区分

完全
区分
开来

反复改进原理

假设查询词 “newspaper”

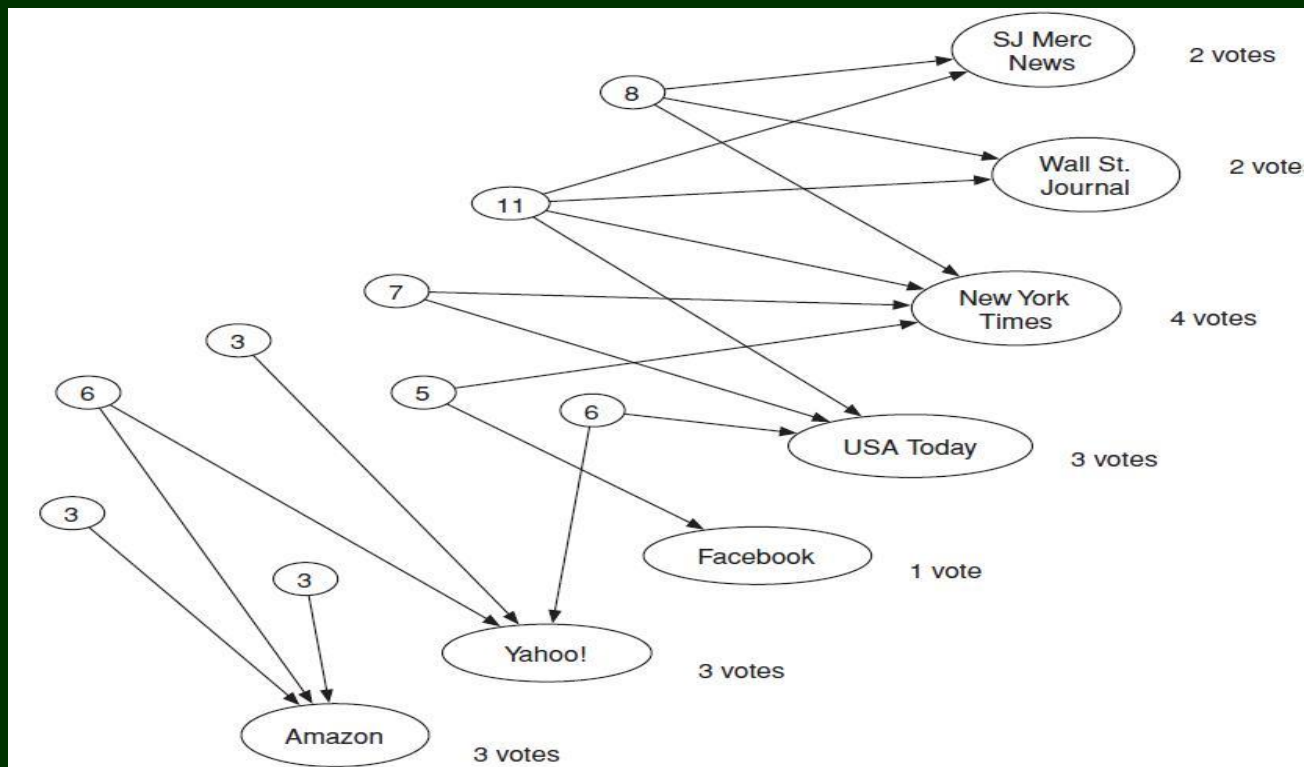
- 左边是与“newspaper”字面上相关的网页。
- 右边是它们所指向的网页，得到的“票数”表示一定的认可度



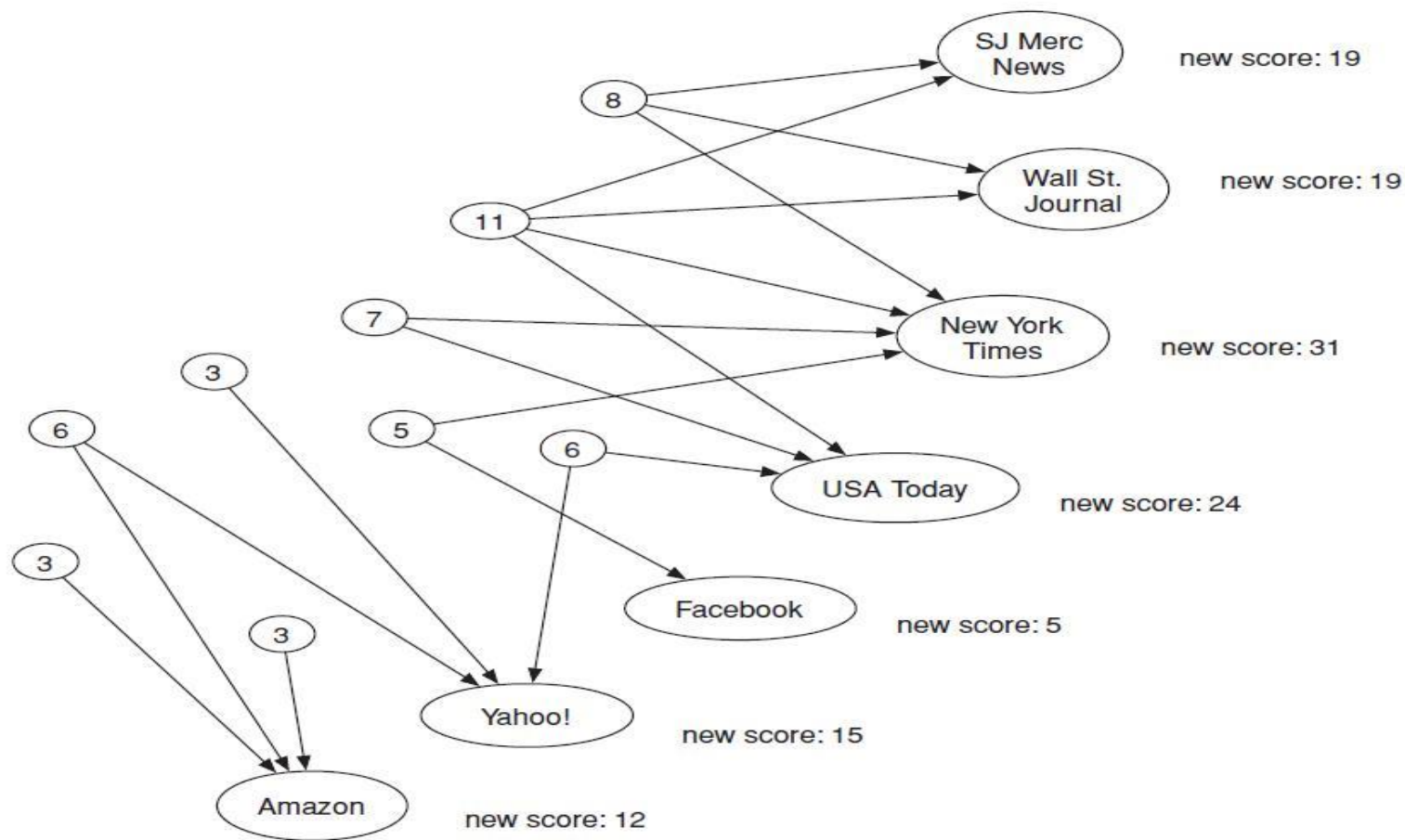
反复改进原理（续）

（principle of repeated improvement）

- 也可以反过来评估“推荐者”的分量
- 然后可以在考虑推荐者分量的情况下重新评估网站相对于“newspaper”的重要性（相当于加权评分）



反复改进原理



- 这个过程可以反复进行下去

网页的“中枢”与“权威”性

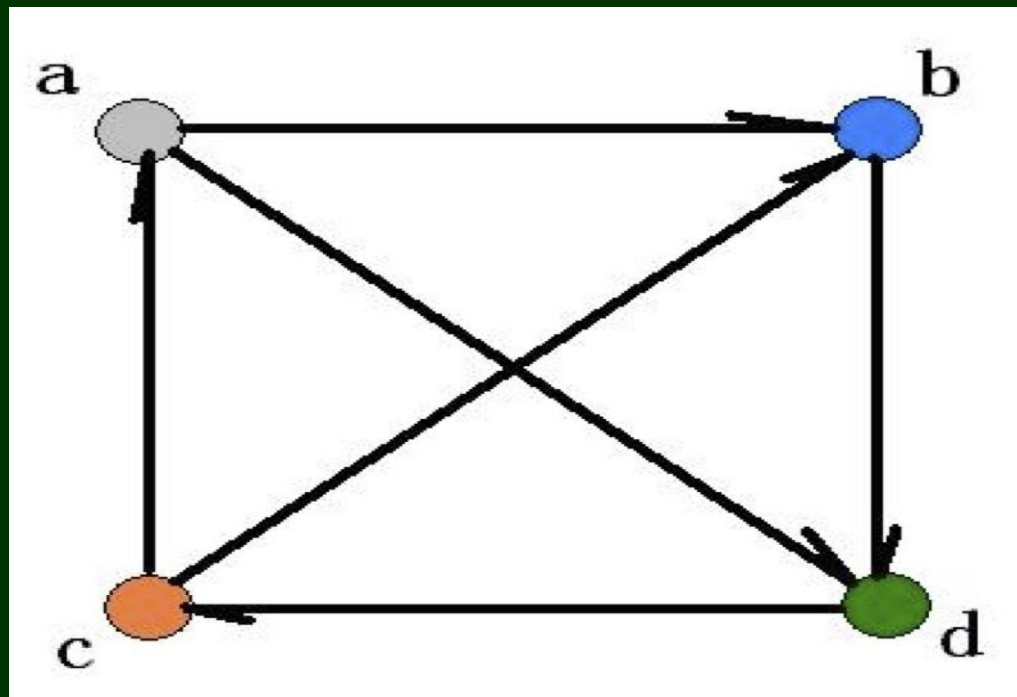
- 万维网中一篇网页的两面属性。观念：
 - 被很多网页指向：权威性高，认可度高
 - 指向很多网页：中枢性强
- HITS算法：计算网页的权威值（auth）和中枢值（hub）
 - Hyperlink-Induced Topic Search

auth(p) 和 hub(p) 的计算方法

- 输入：一个有向图
- 初始化：对于每一个节点p， $\text{auth}(p)=1$ ， $\text{hub}(p)=1$
- 利用中枢值更新权威值
 - 对于每一个节点p，让 $\text{auth}(p)$ 等于指向p的所有节点q的 $\text{hub}(q)$ 之和
- 利用权威值更新中枢值
 - 对于每一个节点p，让 $\text{hub}(p)$ 等于p指向的所有节点q的 $\text{auth}(q)$ 之和
- 重复上述两步若干（k）次

在搜索引擎领域，auth值或hub值高的网页，有时分别称为“权威网页”和“中枢网页”。一篇网页可以兼具二者。

例子：求下图各节点的auth和hub值
(算法运行3轮即可)



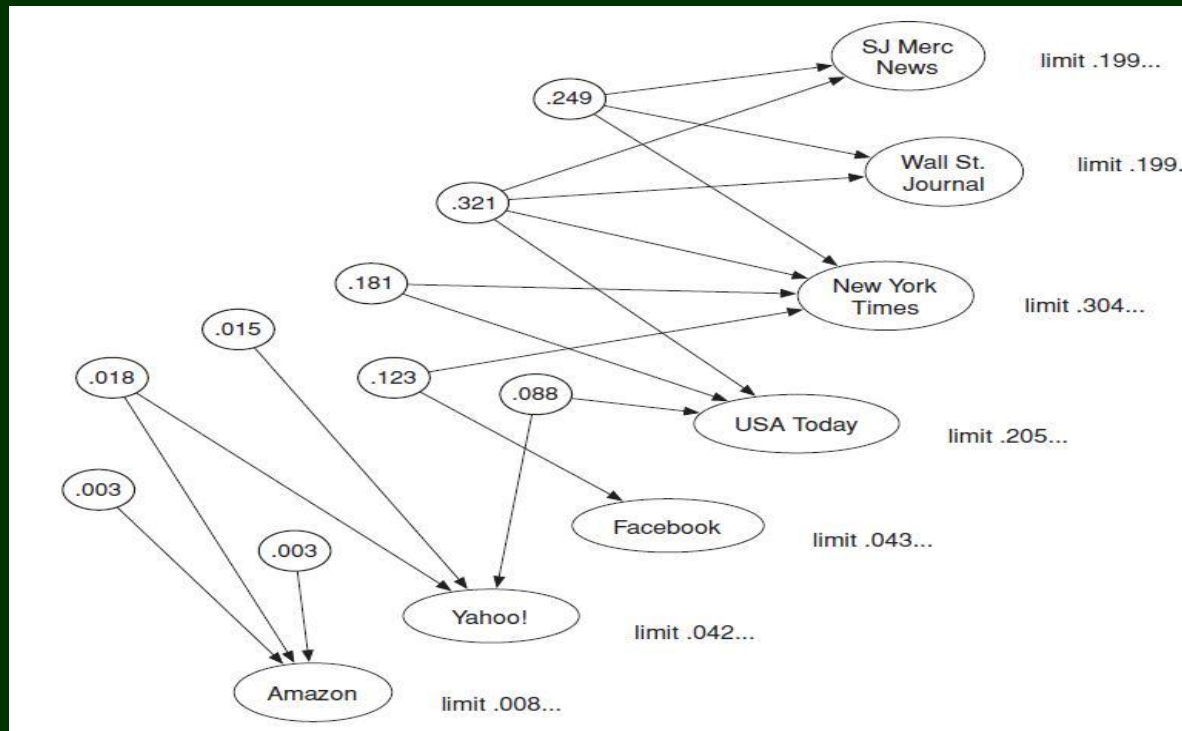
例子：中枢与权威值的迭代改进

Auth				Hub			
a	b	c	d	a	b	c	d
$a=H[c]$	$b=H[a]+H[c]$	$c=H[d]$	$d=H[a]+H[b]$	$a=A[b]+A[d]$	$b=A[d]$	$c=A[a]+A[b]$	$d=A[c]$
1	1	1	1	1	1	1	1
1	2	1	2				
				4	2	3	1
3	7	1	6				
				13	6	10	1
10	23	1	19				
				42	19	33	1
33	75	1	61				
				136	61	108	1

- 越来越大，什么时候算完？收敛？

归一化与极限

- 数值随迭代次数递增
- Auth和hub值的意义在于相对大小
- 在每一轮结束后做归一化：值 / 总和
- 归一化结果随迭代次数趋向于一个极限
 - 相继两次迭代的值不变
 - 极限与初值无关，即存在“均衡”



小结

- 在一个由“引用”或者“推荐”关系构成的信息网络中，每个节点有两种自然的作用：“权威”与“枢纽”（中枢）
- 这样的作用可以通过“HITS算法”得到量化
- HITS算法的基本精神是基于信息网络的结构，在两个量之间交叉进行“反复改进”