



# 链接分析与网络搜索

# 提要

- 网页排序（**ranking**、排名、排位）
  - 搜索服务的基本问题，传统信息检索技术的不足
- 中枢与权威
  - 一篇网页的两面性；有向图的启示
- 中枢值与权威值的计算（**HITS**算法）
- **PageRank**（含义）
- **PageRank**（计算）
- 退化图结构带来的问题
- 随机游走及其与**PageRank**定义的等价关系

# 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？
  - “最可能满足”的多义性
    - 同一个查询，不同的需求（苹果，病毒等）；
    - 不同的查询，相同的需求（电脑，计算机等）

# 传统信息检索技术的要点

(information retrieval, IR)

- 基于词语之间的相关性 (relevance)
  - $\text{similarity}(q, d) \approx \sum \text{score}(d, \text{term})$
- 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）

The strength of weak ties - Google Search

http://www.google.com/#hl=en&source=hp&q=+The+s

Apple中国 中国雅虎 维基百科 新闻 热门

The strength of weak ties - Goog...

+You Web Images Videos Maps News Gmail More Sign in

Google The strength of weak ties

Search About 2,950,000 results (0.26 seconds)

Everything

Images

Maps

Videos

News


Shopping

Books

More

Show search tools

**Scholarly articles for The strength of weak ties**

 [The strength of weak ties](#) - Granovetter - Cited by 17579

[The strength in weak ties](#) - Liu - Cited by 81

[The strength of weak ties: A network theory revisited](#) - Granovetter - Cited by 3331

**[PDF] The Strength of Weak Ties - Stanford Sociology**

[sociology.stanford.edu/people/.../granstrengthweakties.pdf](http://sociology.stanford.edu/people/.../granstrengthweakties.pdf)

File Format: PDF/Adobe Acrobat - Quick View

by MS Granovetter - 1973 - Cited by 17543 - Related articles

**The Strength of Weak Ties.** Mark S. Granovetter. American Journal of Sociology, Volume 78, Issue 6 (May, 1973), 1360-1380. Your use of the JSTOR database ...

**Mark Granovetter - Wikipedia, the free encyclopedia**

[en.wikipedia.org/wiki/Mark\\_Granovetter](http://en.wikipedia.org/wiki/Mark_Granovetter) - Cached

Jump to [The strength of weak ties](#): Granovetter's paper "The Strength of Weak Ties" is a highly influential sociology paper, with nearly 15000 citations ...

[Background](#) - [Major ideas](#) - [Bibliography \(selected\)](#) - [See also](#)

**[PDF] THE STRENGTH OF WEAK TIES: A NETWORK THEORY ...**

[citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128...](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128...)

File Format: PDF/Adobe Acrobat - Quick View

by M Granovetter - 1983 - Cited by 3321 - Related articles

hypotheses of my 1973 paper "The Strength of Weak Ties". (hereafter "SWT") and work that elaborates those hypotheses theoretically or uses them to suggest ...

**The Strength of Weak Ties - Sociable Media Group**

[smg.media.mit.edu/classes/library/...weak.ties/granovetter.htm](http://smg.media.mit.edu/classes/library/...weak.ties/granovetter.htm) - Cached

by MS Granovetter - 1973 - Cited by 3 - Related articles

# 为什么能恰到好处？

- 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





# 这个两个结果哪一个较好？

百度搜索\_大学

http://www.baidu.com/s?wd=%B4%F3%D1%A7&rsv\_bp=0&rsv\_spt=3&in

Apple中国 中国雅虎 维基百科 新闻 热门

百度搜索\_大学

**Baidu 百度** 新闻 网页 贴吧 知道 MP3 图片 视频 地图 更多

大学

[大学 百度百科](#)

大学，指提供教学和研究条件和授权颁发学位的高等教育机关。分为综合大学、专科大学或学院。它选拔具有高中以上学历者进行教育和培训，并以考试考核的方式检验其所学知识和技能。另有，儒家基本经典之一《大学》，也指聚集在特定地...共226次编辑  
baike.baidu.com/view/4410.htm 2011-11-1

[欢迎访问北京大学主页](#)

北京大学作为国内前茅的文理医工综合性大学，在培养高素质创新型人才、取得突破性科研进展，以及为国民经济和社会发展提供智力支持等方面都发挥着极其重要的作用。  
www.pku.edu.cn/ 2011-10-12 - 百度快照

[北京大学 百度百科](#)

北京大学，简称北大，创办于1898年，初名京师大学堂，是中国近代第一所国立大学，被公认为中国的最高学府，也是亚洲和世界最重要的大学之一。在中国现代史上，...  
baike.baidu.com/view/1471.htm 2011-10-20 - 百度快照

[hao123网址之家 -- 大学](#)

把hao123设为首页 网友反馈 首页>大学 分数线查询 地区批次线正在加载，请稍候...211高校名单 985 高校名单 校园社区 大学排名(仅供参考) 阳光高考 大学专业介绍...  
www.hao123.com/edu.htm 2011-10-21 - 百度快照

[在北京市搜索大学\\_百度地图](#)



A. [北京大学](#) - (010)62752114  
海淀区颐和园路5号

B. [清华大学](#)  
北京市海淀区

C. [清华大学](#) - (010)62782165  
清华园

D. [西南财经大学](#) - (010)82380594  
北京市海淀区学院路30号北京科技大学会...

大学 - Google Search

http://www.google.com/#hl=en&newwindow=1&q=大学

Apple中国 中国雅虎 维基百科 新闻 热门

大学 - Google Search

**Google** 大学

Search About 5,030,000,000 results (0.11 seconds)

**Everything**

Images 30+ items - 把hao123设为首页 · 网友反馈 · 首页>大学. 分数线查询. 地区批次 ...

Maps • 地区 - 普通本科院校(820所 - 高职院校(1228所 - 独立学院(311所  
• 北京 - 58所[点击查看 - 25所[点击查看 - 5所[点击查看  
• 天津 - 19所[点击查看 - 26所[点击查看 - 10所[点击查看

Videos 211高校名单 - 北京普通本科 - 山东普通本科 - 江西普通本科

News [搜狐教育高校信息库-高校、专业一站查询](#)  
daxue.learning.sohu.com/ - Cached - Translate this page  
中国大学信息查询系统所有信息、数据均来源于高校网站或相关出版物，仅供考生参考，请以官方公布信息为准。建议考生综合考虑各校公布的信息及各种因素填报 ...

Shopping

Books

Blogs [大学-精品学习网](#)  
www.51edu.com/daxue/ - Cached - Translate this page  
大学排行：精品学习网大学频道为广大网友搜集了2011三本大学排名汇总，供各位考试参考。[详细] · 2011中国三星级私立大学学科专业名单（港澳台） · 2011年全国二 ...

More

**All results**

Sites with images

More search tools

**News for 大学**



[寒门子弟如何上大学?](#)

新民网 - 5 hours ago  
网上舆情要览：在大学教育阶段再去争取教育平权，其实为时已晚，教育的公平首先应当是起跑线上的公平，中小学的招生同样应当注重教育平权，向农村孩子作出一定倾斜， ...

1225 related articles

[郎朗哈佛大学开课 与刘翔相约明年奥运\(图\)](#)

[腾讯网 - 57 related articles](#)

[张尧学任中南大学校长 黄伯云不再任校长职务](#)

[红网 - 47 related articles](#)

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



# 餐馆推荐问题

	甲	乙	丙	丁		
新辣道	*		*	*	3	21
海底捞	*	*	*		3	20
麦当劳		*			1	6
五方院	*			*	2	15
俏江南		*		*	2	13
	8	6	6	7		

看推荐人的“水平”

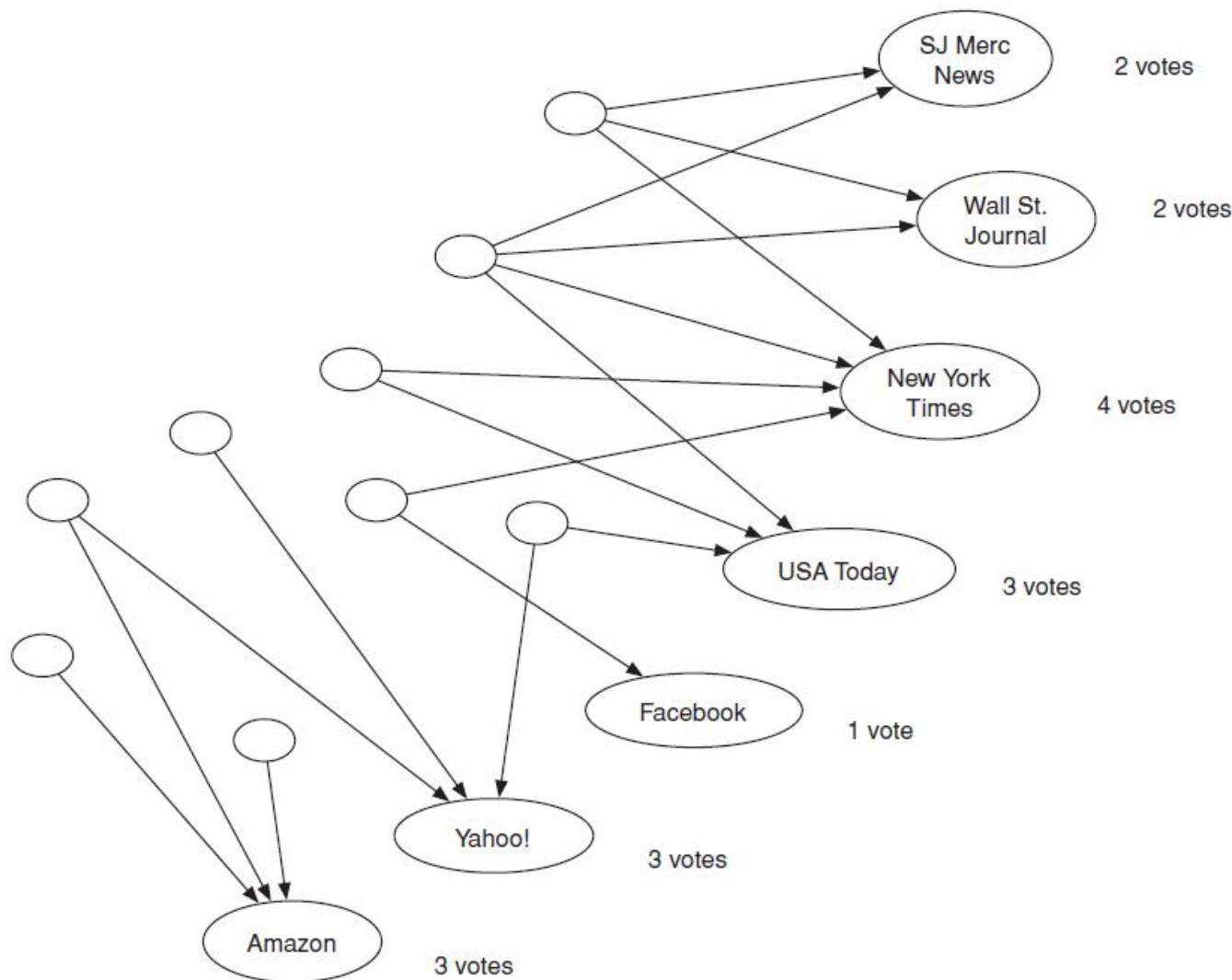
不能  
完全  
区分

完全  
区分  
开来

# 反复改进原理（例）

假设查询词 “newspaper”

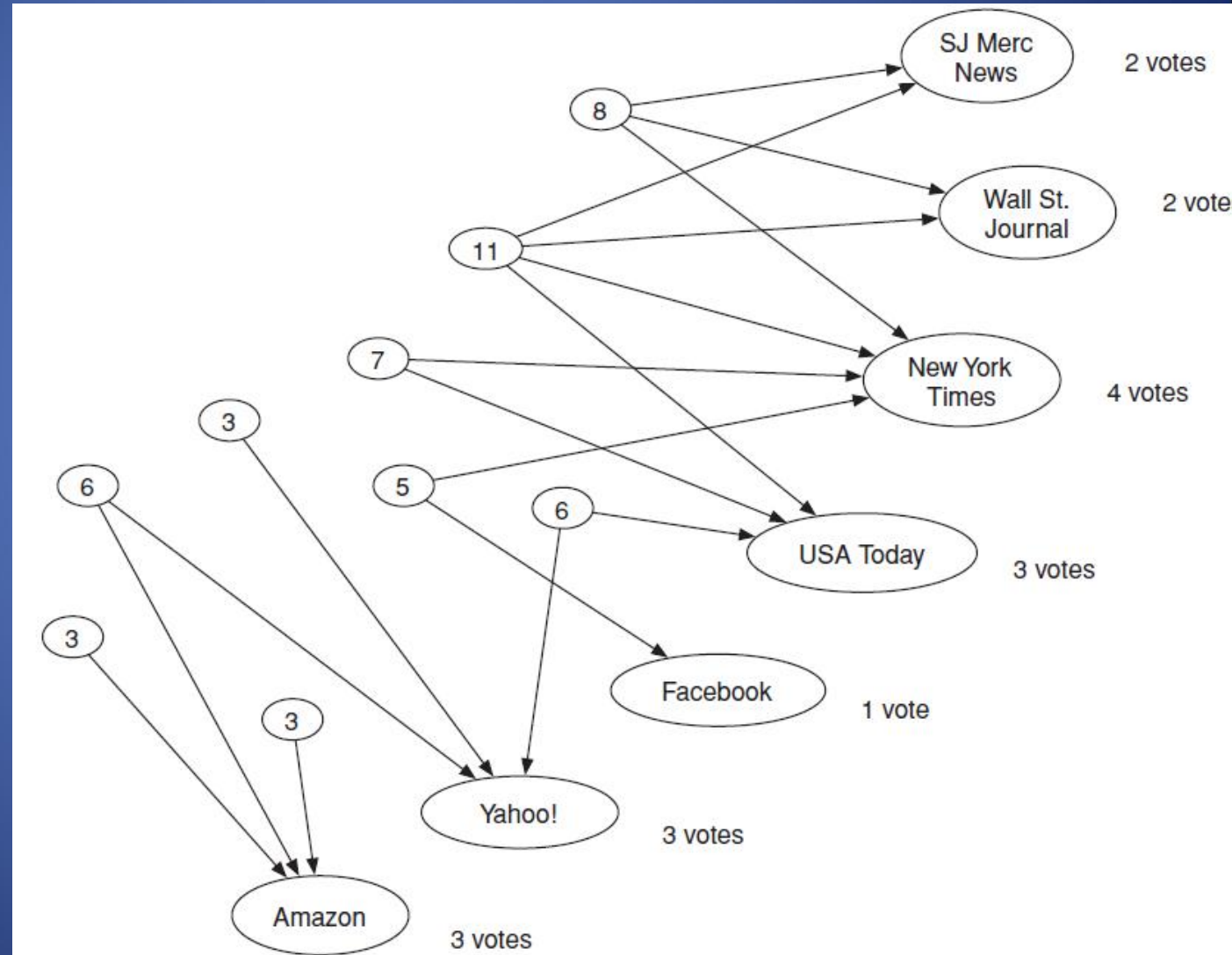
- 左边是与“newspaper”字面上相关的网页。
- 右边是它们所指向的网页，得到的“票数”表示一定的认可度



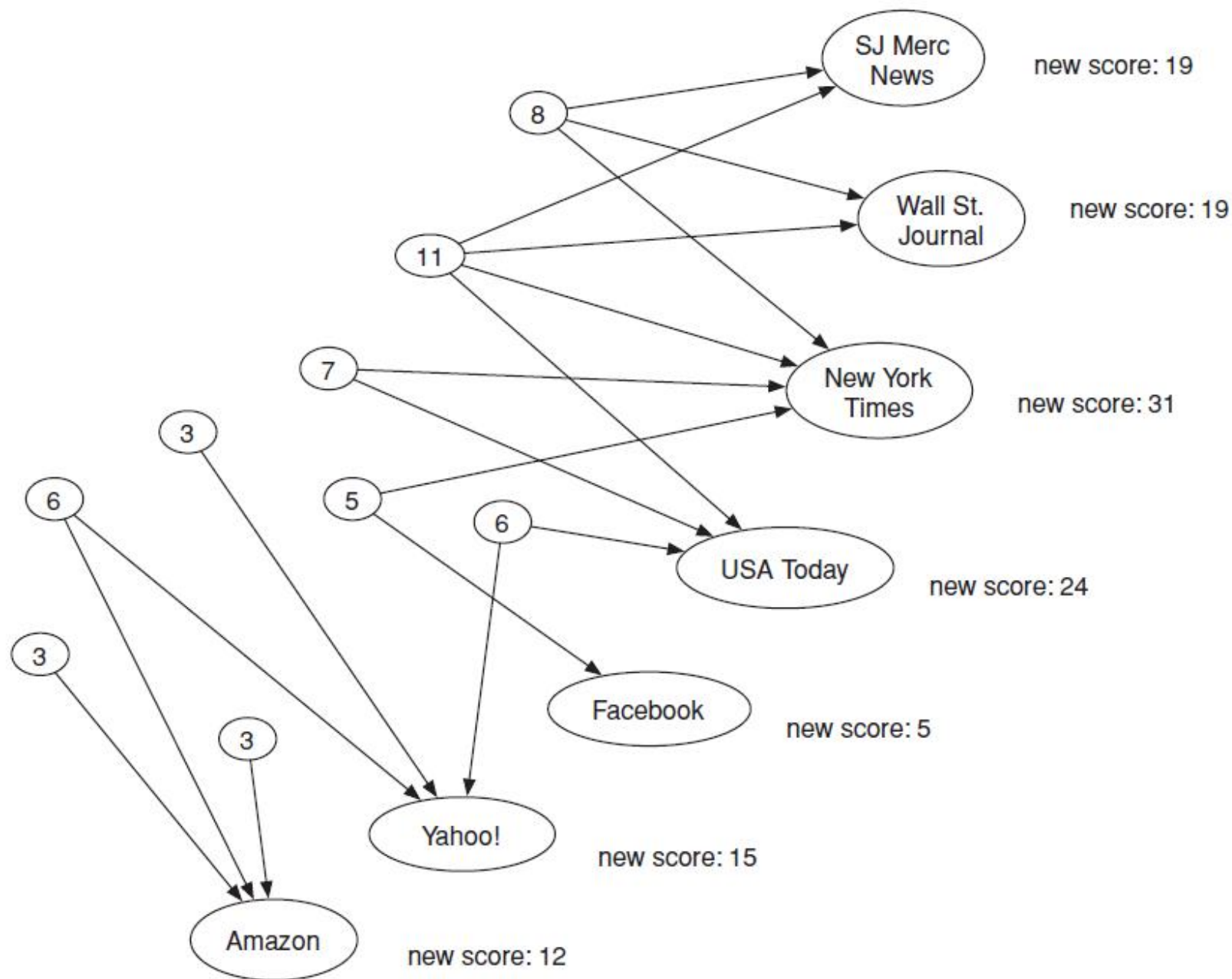
# 反复改进原理（续）

## （principle of repeated improvement）

- 也可以反过来评估“推荐者”的份量
- 然后可以在考虑推荐者份量的情况下重新评估网站相对于“newspaper”的重要性



# 反复改进原理



# “中枢” (hub) 与 “权威” (authority)

- 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
  - 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
- 在实际中算法实施的针对性：相关网页集合
  - 不是全部网页集合。为避免赘述，这一点后面不再总强调，认为所讨论的就是相关网页集合。

# auth(p) 和 hub(p) 的计算方法

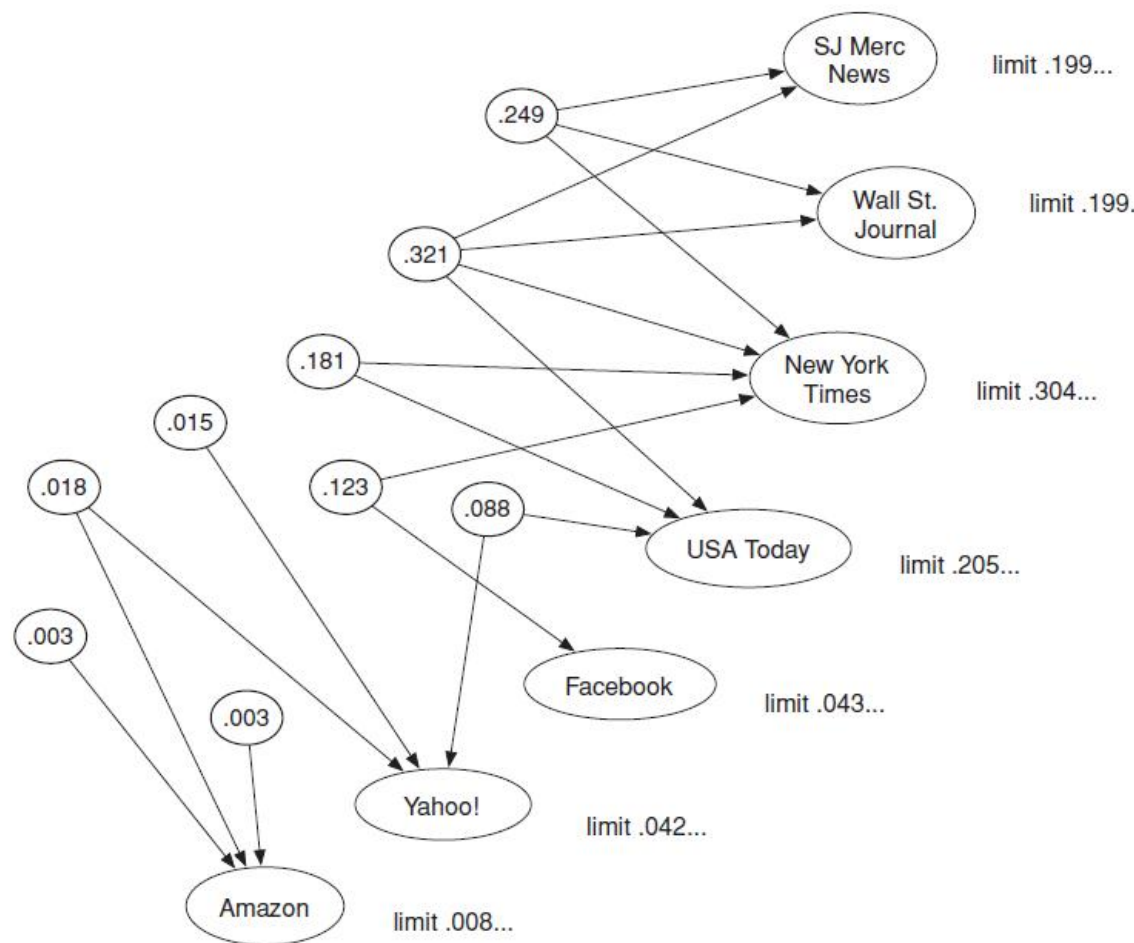
- 输入：一个有向图
- 初始化：对于每一个节点 $p$ ， $\text{auth}(p)=1$ ， $\text{hub}(p)=1$
- 利用中枢值更新权威值
  - 对于每一个节点 $p$ ，让 $\text{auth}(p)$ 等于指向 $p$ 的所有节点 $q$ 的 $\text{hub}(q)$ 之和
- 利用权威值更新中枢值
  - 对于每一个节点 $p$ ，让 $\text{hub}(p)$ 等于 $p$ 指向的所有节点 $q$ 的 $\text{auth}(q)$ 之和
- 重复上述两步若干（ $k$ ）次

在搜索引擎领域， $\text{auth}$ 值或 $\text{hub}$ 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。

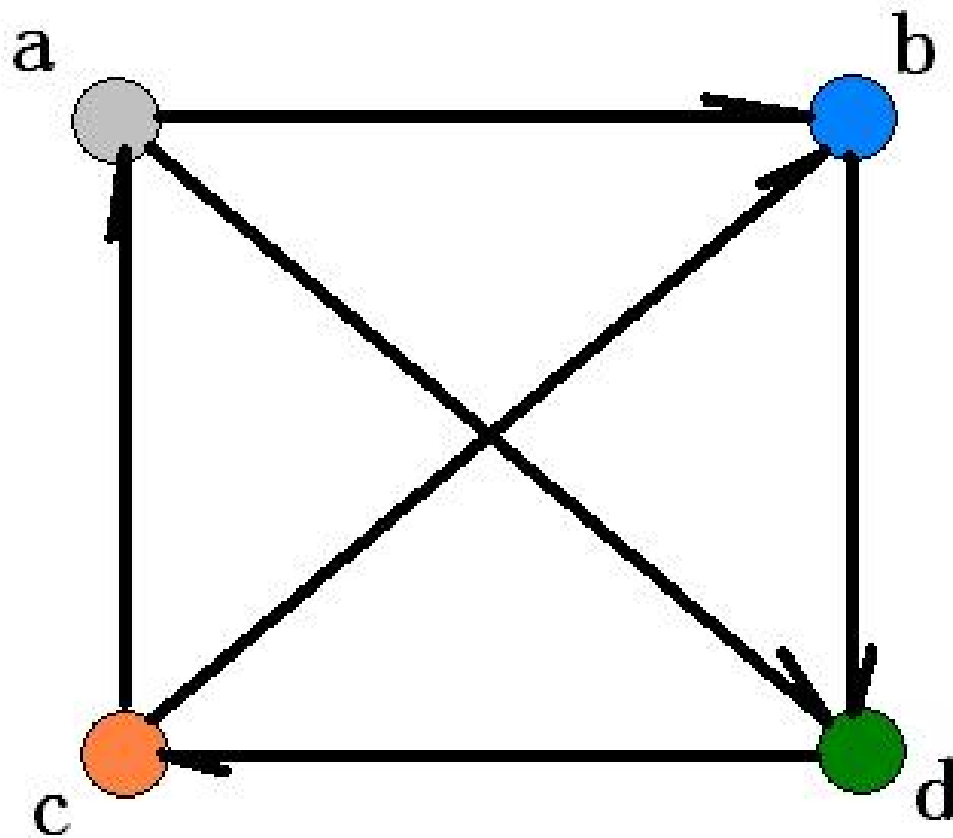


# 归一化与极限

- 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”



# PageRank: 节点的一种重要性测度



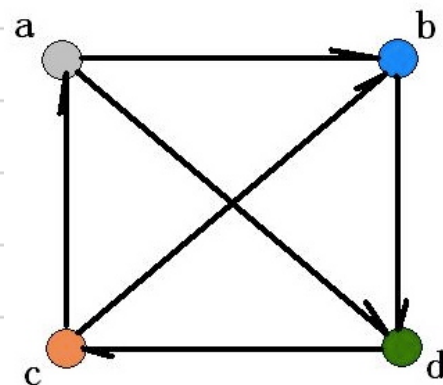
$$\begin{aligned}a &= c/2 \\ b &= a/2 + c/2 \\ c &= d \\ d &= a/2 + b\end{aligned}$$

搜索引擎形成查询结果网页排序的重要参数

# 上图的算例

经过约70次迭代，最后收敛到：  
A=0.615, B=0.923, C=D=1.231

	A	B	C	D	E
1	$a=c/2$	$b=a/2+c/2$	$c=d$	$d=a/2+b$	传递关系
2	a	b	c	d	
3	1	1	1	1	初值
4	0.5	1	1	1.5	按传递关系计算
5	0.5	0.75	1.5	1.25	
6	0.75	1	1.25	1	
7	0.625	1	1	1.375	
8	0.5	0.8125	1.375	1.3125	
9	0.6875	0.9375	1.3125	1.0625	
10	0.65625	1	1.0625	1.28125	
11	0.53125	0.859375	1.28125	1.328125	
12	0.640625	0.90625	1.328125	1.125	
13	0.6640625	0.984375	1.125	1.2265625	

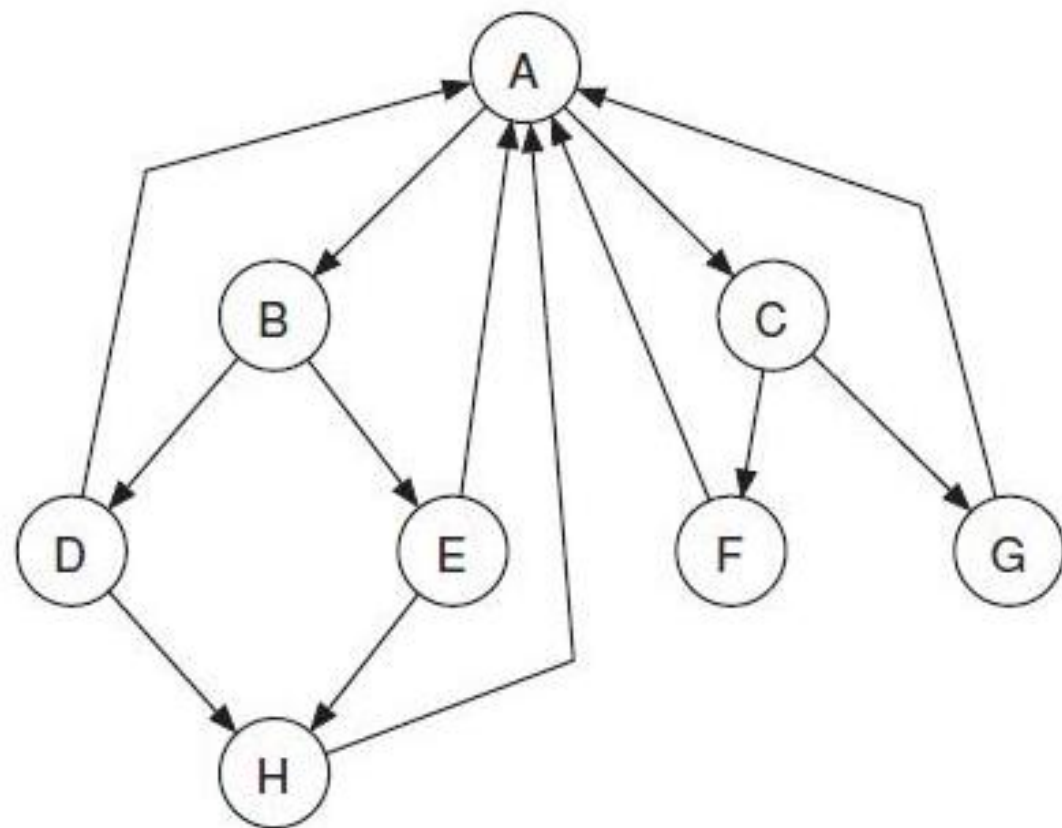


# PageRank基本算法描述

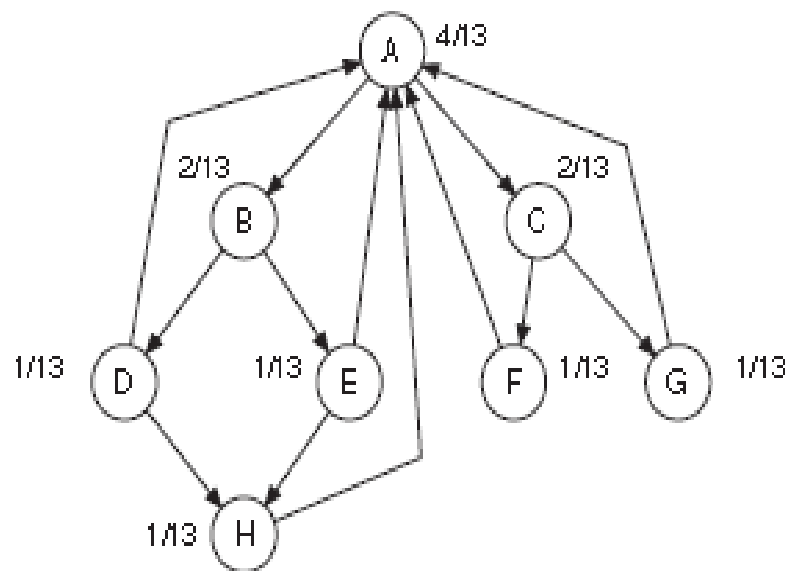
- 输入：一个有 $n$ 个节点的网络（有向图），设所有节点的PageRank初始值为 $1/n$ 。
- 选择操作的步骤数 $k$
- 对PageRank做 $k$ 次更新操作，每次使用以下规则：
  - 每个节点将自己当前的PageRank值通过出向链接均分传递给所指向的节点
    - 若没有出向链接，则认为传递给自己
  - 每个节点以从入向链接获得的（包括可能自传的）所有值之和更新它的PageRank

# 一个计算网页排名的实例

- 每个节点的初值都是 $1/8$
- 最后收敛结果见下图

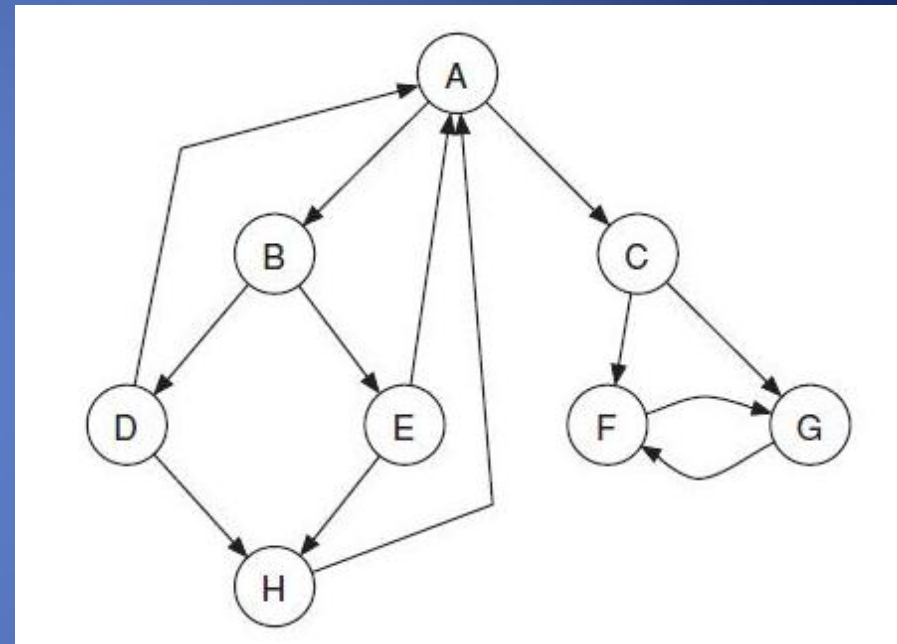


Step	A	B	C	D	E	F	G	H
1	$\frac{1}{2}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$
2	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$



# PageRank基本算法在某些网络结构上表现不好

- PageRank算法不象HITS算法那样需要归一化问题，但有新问题
- F和G两个节点显得很“自私”：吸收别人的价值，但不向外传
  - 导致它们最后各自1/2，其他人都0



这也显示了共谋（**colluding**）制造垃圾网页的一个原理



# PageRank值很快集中到F和G

	A	B	C	D	E	F	G	H
62	$a=d/2+e/2+1$	$b=a/2$	$c=a/2$	$d=b/2$	$e=b/2$	$f=c/2+g$	$g=c/2+f$	$h=d/2+e/2$
63	a	b	c	d	e	f	g	h
64	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
65	0.25	0.0625	0.0625	0.0625	0.0625	0.1875	0.1875	0.125
66	0.1875	0.125	0.125	0.03125	0.03125	0.21875	0.21875	0.0625
67	0.09375	0.09375	0.09375	0.0625	0.0625	0.28125	0.28125	0.03125
68	0.09375	0.046875	0.046875	0.046875	0.046875	0.328125	0.328125	0.0625
69	0.109375	0.046875	0.046875	0.0234375	0.0234375	0.3515625	0.3515625	0.046875
70	0.0703125	0.0546875	0.0546875	0.0234375	0.0234375	0.375	0.375	0.0234375
71	0.046875	0.03515625	0.03515625	0.02734375	0.02734375	0.40234375	0.40234375	0.0234375
72	0.05078125	0.0234375	0.0234375	0.01757813	0.017578125	0.41992188	0.41992188	0.02734375
73	0.04492188	0.02539063	0.02539063	0.01171875	0.01171875	0.43164063	0.43164063	0.01757813
74	0.02929688	0.02246094	0.02246094	0.01269531	0.012695313	0.44433594	0.44433594	0.01171875
75	0.02441406	0.01464844	0.01464844	0.01123047	0.011230469	0.45556641	0.45556641	0.01269531
76	0.02392578	0.01220703	0.01220703	0.00732422	0.007324219	0.46289063	0.46289063	0.01123047
77	0.01855469	0.01196289	0.01196289	0.00610352	0.006103516	0.46899414	0.46899414	0.00732422
78	0.01342773	0.00927734	0.00927734	0.00598145	0.005981445	0.47497559	0.47497559	0.00610352
79	0.01208496	0.00671387	0.00671387	0.00463867	0.004638672	0.47961426	0.47961426	0.00598145
80	0.01062012	0.00604248	0.00604248	0.00335693	0.003356934	0.48297119	0.48297119	0.00463867
81	0.00799561	0.00531006	0.00531006	0.00302124	0.00302124	0.48599243	0.48599243	0.00335693
82	0.00637817	0.0039978	0.0039978	0.00265503	0.002655029	0.48864746	0.48864746	0.00302124

# PageRank的同比缩减与统一补偿规则

- 同比缩减
  - 在每次运行基本PageRank更新规则后，将每一节点的PageRank值都乘以一个比例因子 $s$ ， $0 < s < 1$ ，经验值在0.8-0.9之间。
- 统一补偿
  - 在每一节点的PageRank值上统一加上 $(1-s)/n$ 。

这样，既维持了所有PageRank之和等于1的性质，也防止了PageRank值不恰当地集中到个别节点。





# 随机游走：PageRank的另一种等价理解

- 想象一个人从一篇**随机选择**的网页开始，随机选择其中的链接浏览到下一篇文章，并不断如此进行，称为“随机游走”。
- 考虑一篇网页 $x$ ，问：经过 $k$ 步随机游走到达 $x$ 的概率是多少？
- 可以证明：到达 $x$ 的概率等于运行PageRank基本算法 $k$ 步得到的值。
- 随机游走概念稍加修改也可以和同比缩减统一补偿的PageRank等价。

# 链接分析小结

- 有向图作为万维网的结构模型
  - 概貌形状：领结
  - 得到领结概貌形状的方法：宽度优先搜索
- 万维网信息节点的（结构性）重要程度
  - 权威性，中枢性：一个节点的双重作用
    - 反复改进原理（交叉支持原理）
    - 相对查询需求或者某个主题
  - PageRank（网页排名）：节点的全局相对重要性
    - 局部结构信息在全局扩散达到的均衡