第四章 PageRank

主要内容

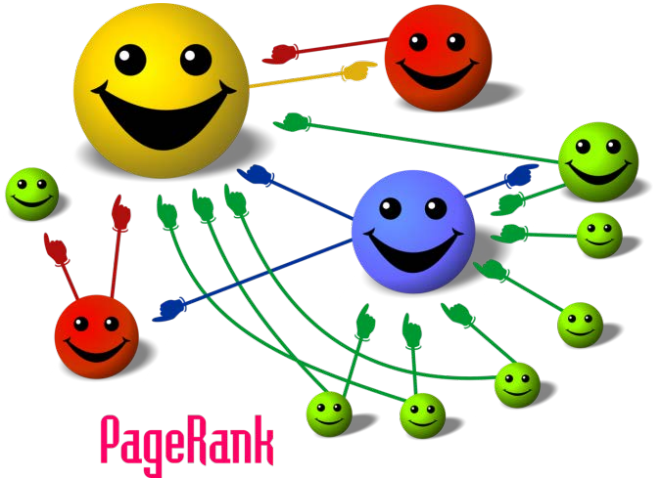
1 问题的引出

2 PageRank计算

3 链接作弊

4 导航页与权威页

5 小结



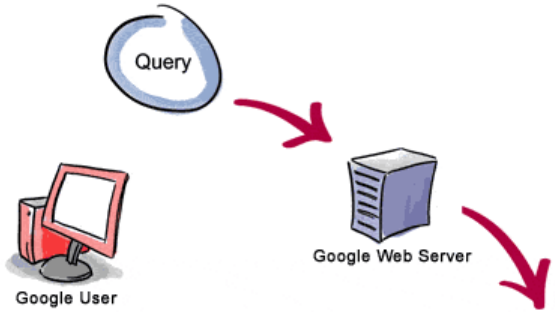
1 问题的引出

1.1 Web结构与搜索

1.2 页面排序

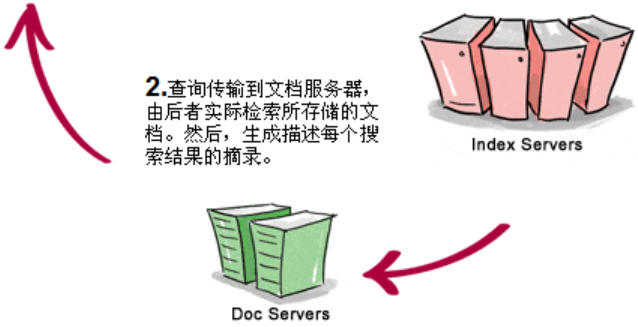
1.3 PageRank的定义

1.1 Web结构与搜索



3. 瞬间返回用户需要的搜索结果。

1. 网络服务器将查询发送到索引服务器。索引服务器所包含的内容与书本末尾的索引目录相似，即说明哪些网页包含与查询匹配的文字。



2. 查询传输到文档服务器，由后者实际检索所存储的文档。然后，生成描述每个搜索结果的摘录。

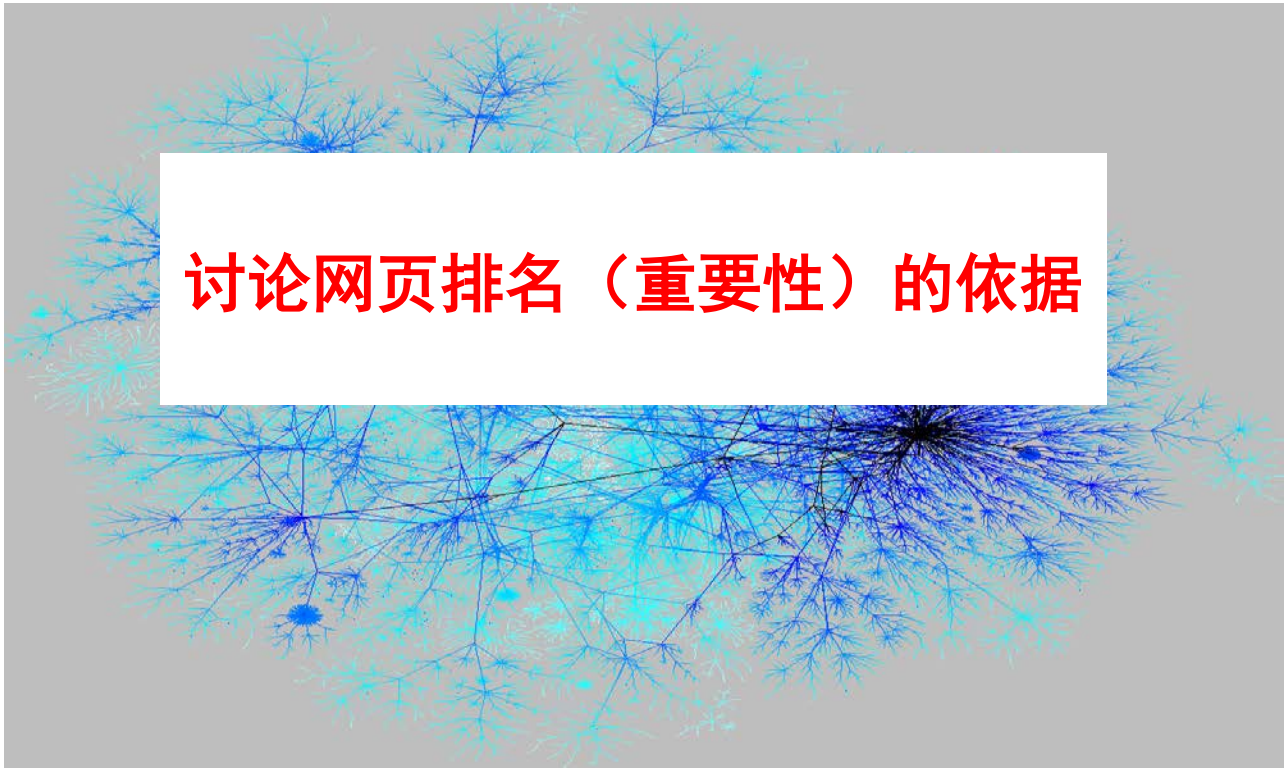
1.2 页面排序


Bai du 百度
java 程序设计
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约32,000,000个
搜索工具
为您推荐: java 程序设计 pdf java语言程序设计 pdf java语言程序设计
百度传课-java编程实例，百度旗下网络课程专业平台
java编程实例百度传课名师主讲，快速掌握技巧，多种免费课程，在线随时无限学百度传课-中国网络课程专业平台，集合百度优势资源，涉及英语/职场/生活等任何技能!
ios应用开发教程
ios应用开发教程-移动开发应用教程尽在百度传课,全球名校名师课程专享,多种..
高中数学题及答案
高中数学题及答案例题分析-历年中高考错题分类解析-知识点总结.
www.chuanke.com 2015-10 V3 - 评价
java 程序设计官方主页 点击进入
java 程序设计，免费入学，保障就业，保底薪保就业，不高薪不就业!java 程序设计， 10年已培养20万多名学员，高薪入职IT名企，年薪10万!
www.beifeng.com 2015-10 V2 - 评价
Java语言程序设计-基础篇(原书第8版) - 下载频道 - CSDN.NET
2012年11月15日 - 《Java语言程序设计:基础篇(原书第8版)》是Java语言的经典教材,中文版分为《Java语言程序设计基础篇》和《Java语言程序设计进阶篇》,主要介绍程序设计...
download.csdn.net/deta... - 百度快照 - 75%好评


Google
java 程序设计
网页 图片 视频 新闻 地图 更多 搜索工具
找到约 1,350,000 条结果 （用时 0.29 秒）
Java程序设计- Peking University | Coursera
https://www.coursera.org/course/pkujava
Java程序设计from Peking University. 《Java程序设计》课程是使用Java语言进行应用程序设计的课程，针对各专业的大学本科生开设。课程的主要目标有三： 一、 ...
Java程序设计基础教程_互动百科
www.baike.com/wiki/Java程序设计基础教程
《Java程序设计基础教程》从Java语言编程的入门概念开始，对Java面向对象编程基本概念和技术等内容进行了较为全面和详细地的讲解。《Java程序设计基础教程》 ...
java程序设计（吴萍蒲鹏朱丽娟编清华大学出版社教材）_百度...
baike.baidu.com/view/303655.htm
《java程序设计》是2006年清华大学出版社北京交通大学出版社出版的图书，作者是吴萍、蒲鹏、朱丽娟。主要讲述了本书通过对Java编程语言的全面介绍，引导读者 ...

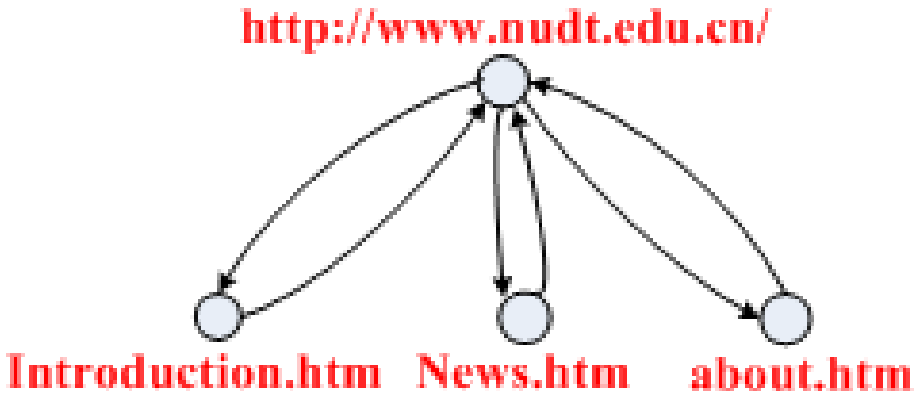
1.2 页面排序

- **互联网Web链接**



1.2 页面排序

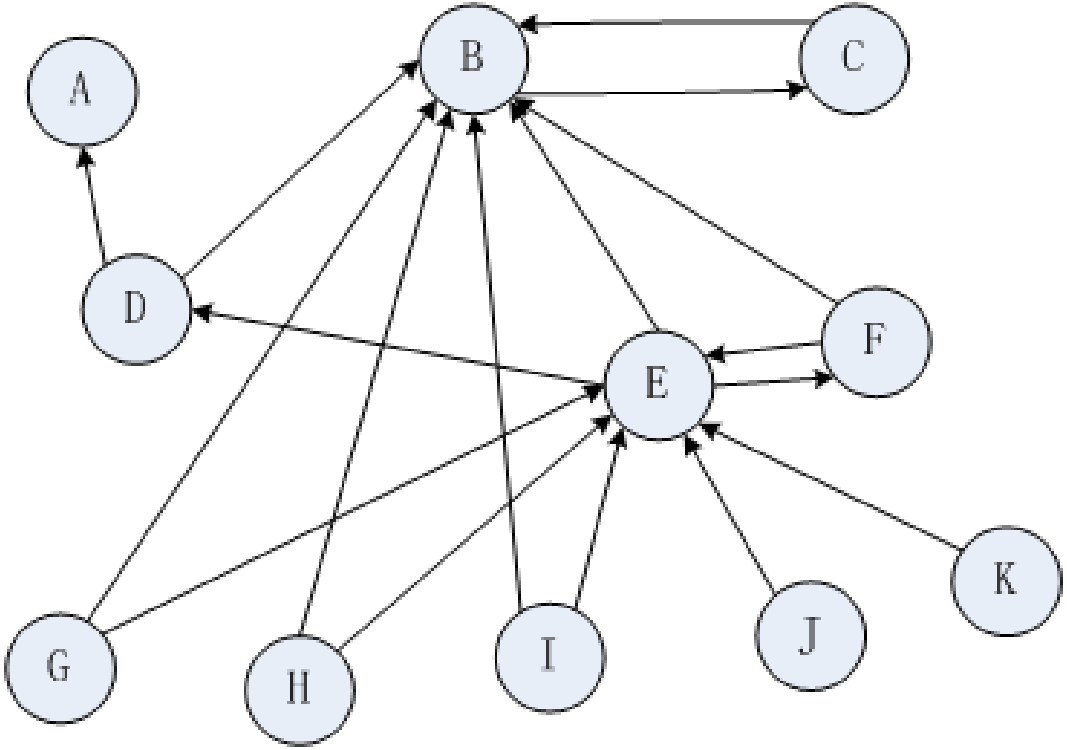
- **网页链接关系**

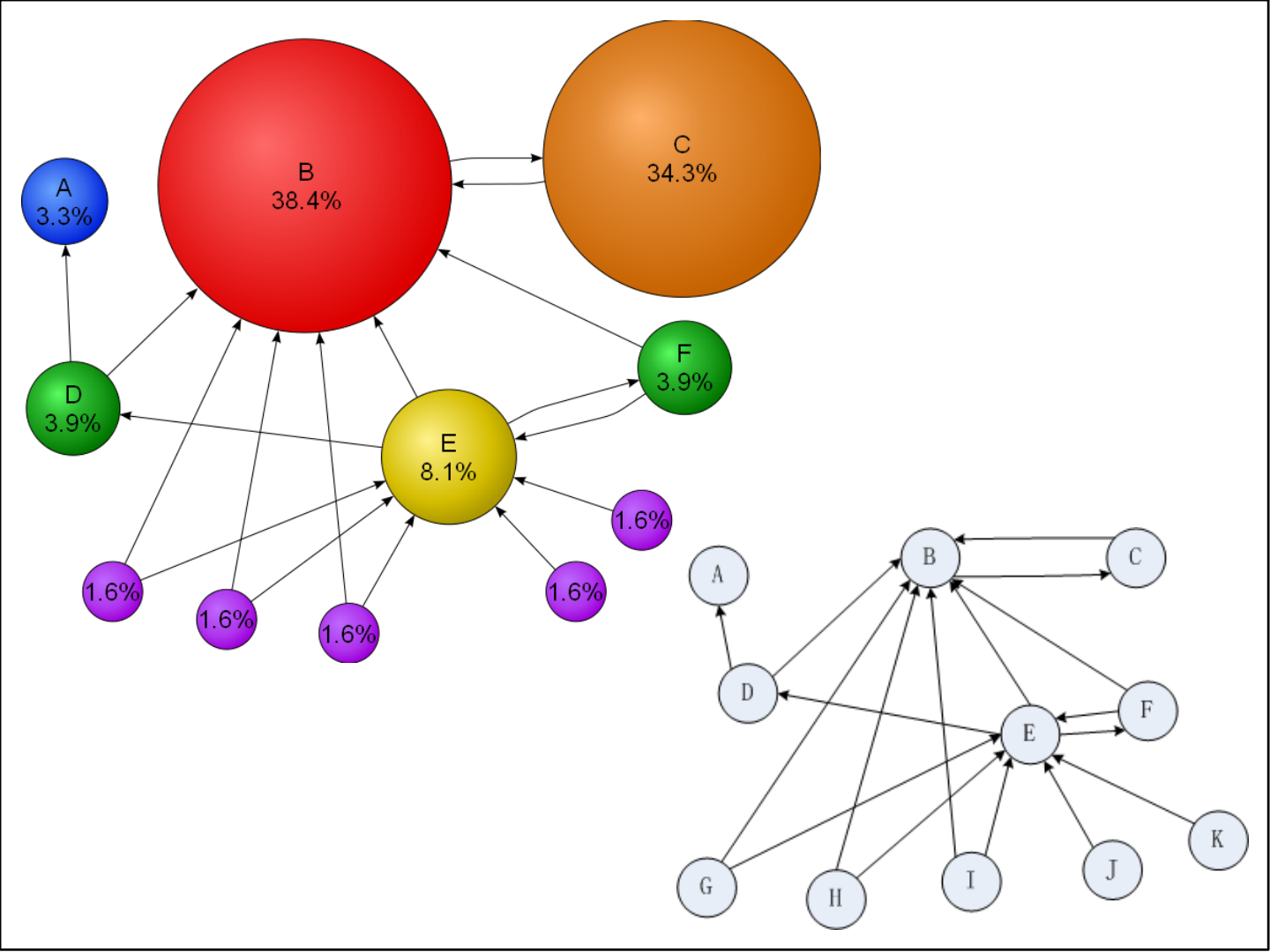


- 网页A链向网页B，可认为B比较重要
- 网页被指向越多越重要
- 被重要网页指向的网页更重要

1.2 页面排序

- 排序案例




A
3.3%
B
38.4%
C
34.3%
D
3.9%
E
8.1%
F
3.9%
1.6%
1.6%
1.6%
1.6%
1.6%
A
B
C
D
E
F
G
H
I
J
K

1.3 PageRank的定义

- **概述**
  - 用于对网页进行排名的一种算法
  - 最早由谷歌总裁Lawrence Page提出

The PageRank citation ranking: bringing order to the Web.
L Page, S Brin, R Motwani, T Winograd - 1999 - ilpubs.stanford.edu
The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, ...
被引用次数：8577

1.3 PageRank的定义

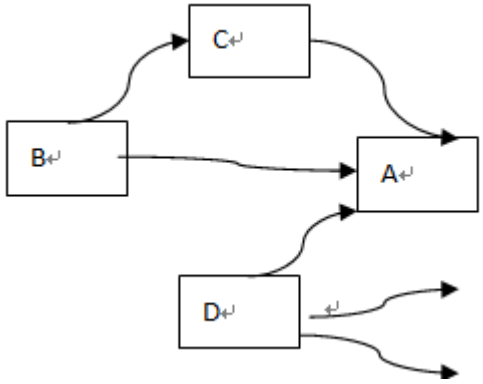
- **重要性测度**
  - 图的每个结点赋予一个PR值
  - PR值用于度量结点的重要程度
  - PR值越大结点越重要

1.3 PageRank的定义

- **量化计算**
  - 结点的PR值等于链入邻居PR值的加权和

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

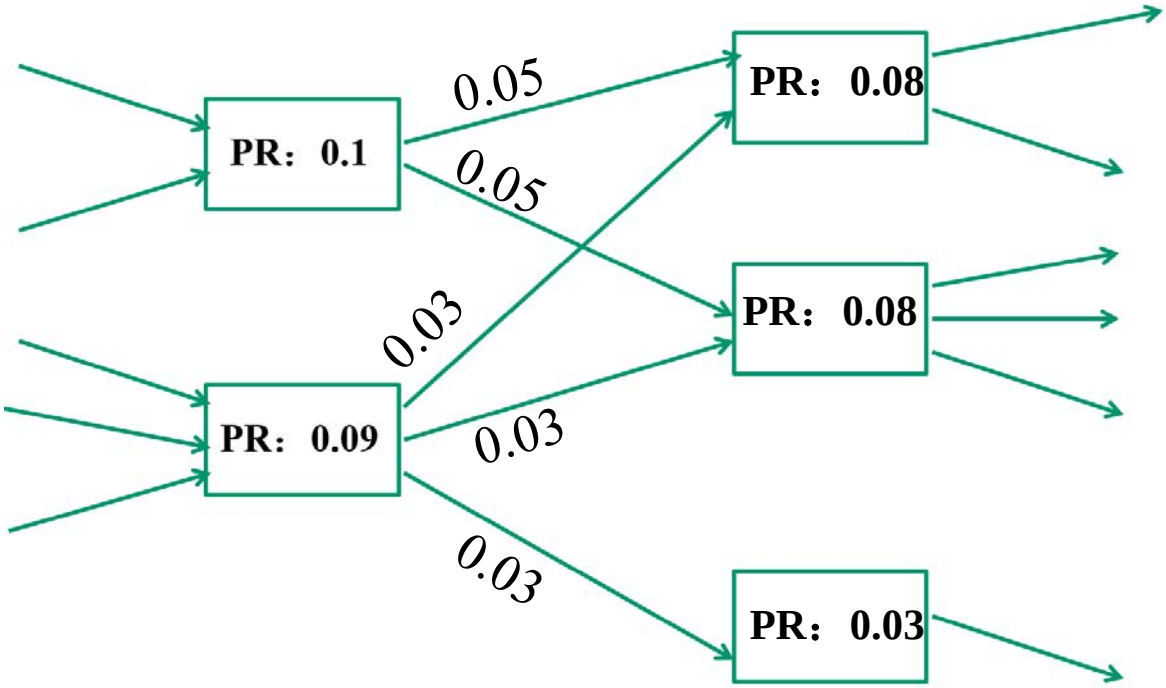
$$PR_i = \sum_{j \in B_i} \frac{PR_j}{L_j}$$



PR_i：页面i的PageRank值
PR_j：页面j的PageRank值
L_j：页面j链出的连接数量
$\mathbf{B}_i$：链接到网页i的页面集合

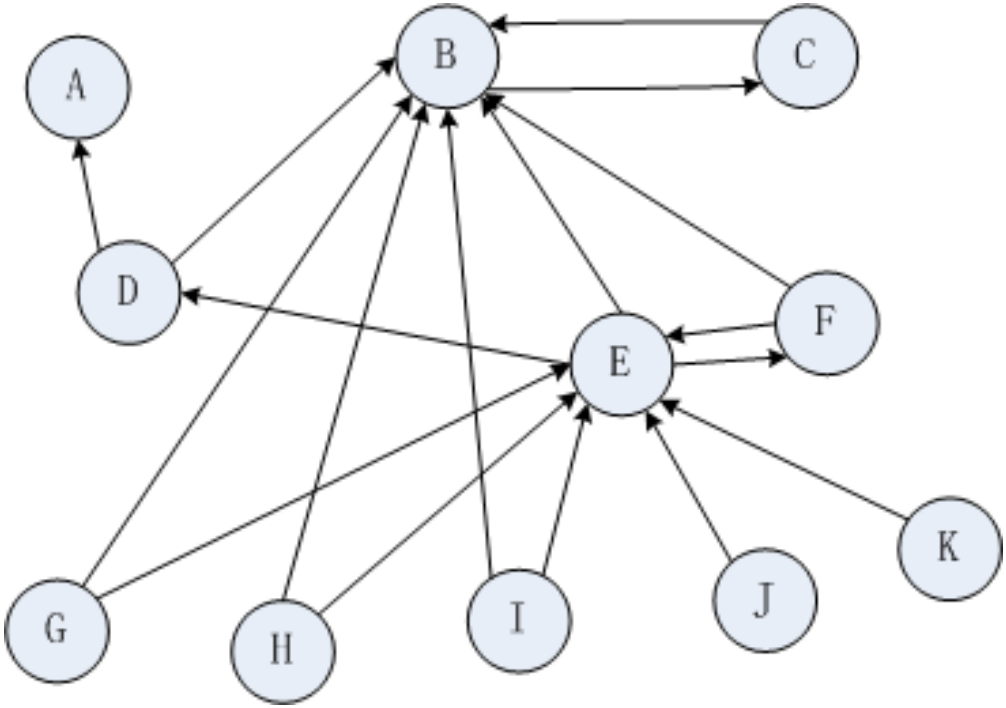
1.3 PageRank的定义

- 量化计算(示例)



1.3 PageRank的定义

- 如何计算？



主要内容

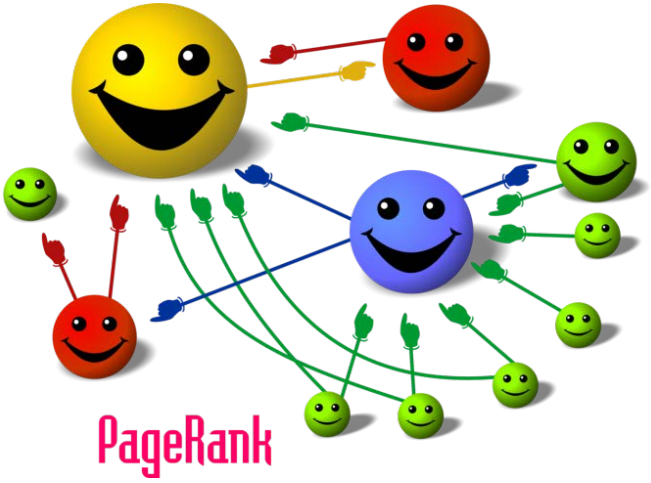
1 问题的引出

2 PageRank计算

3 链接作弊

4 导航页与权威页

5 小结



2 PageRank计算

2.1 转移矩阵

2.2 PageRank求解

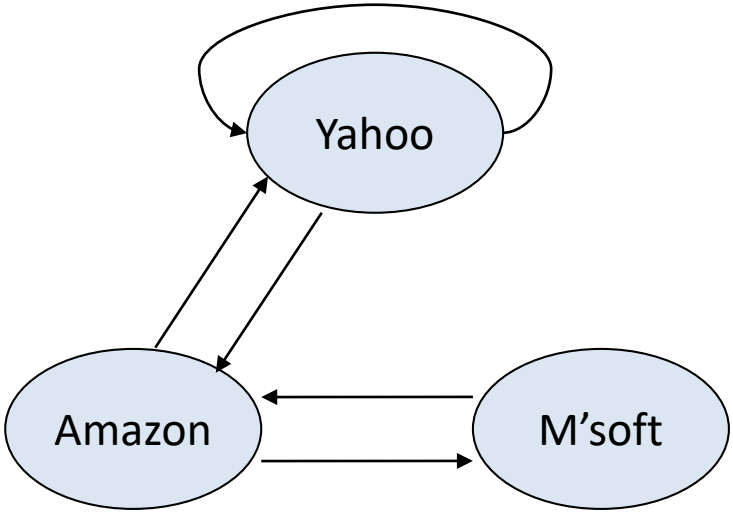
2.3 终止节点

2.4 抽税法

2.5 面向主题的PageRank

2.1 转移矩阵

- 所有网页构成一个有向图
- 每个网页是图中的一个节点
- 转移矩阵$M[i,j]$表示网页j链向网页i的概率

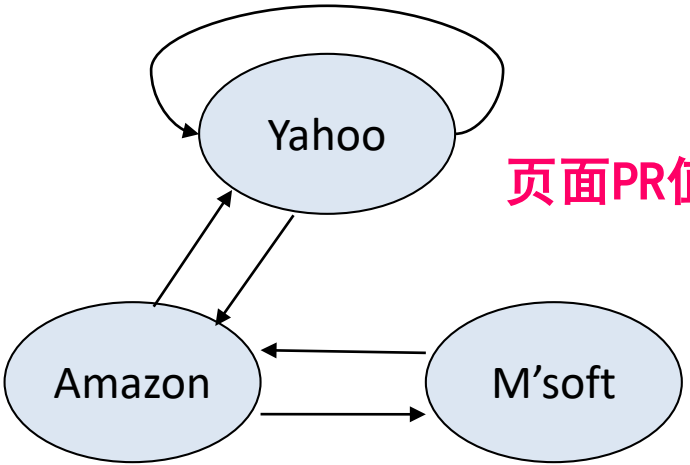


	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

2.1 转移矩阵

- **PageRank物理含义**
  - 随机页面访问，各页面被访问的概率分布
  - 设平稳条件下概率分布$x=(x_1,\ldots\ldots,x_n)$

$$Mx=x \quad \sum_i x_i = 1$$



页面PR值对应转移矩阵的特征向量

2.2 PageRank求解

- **迭代法**

公式：$Mx \rightarrow x$

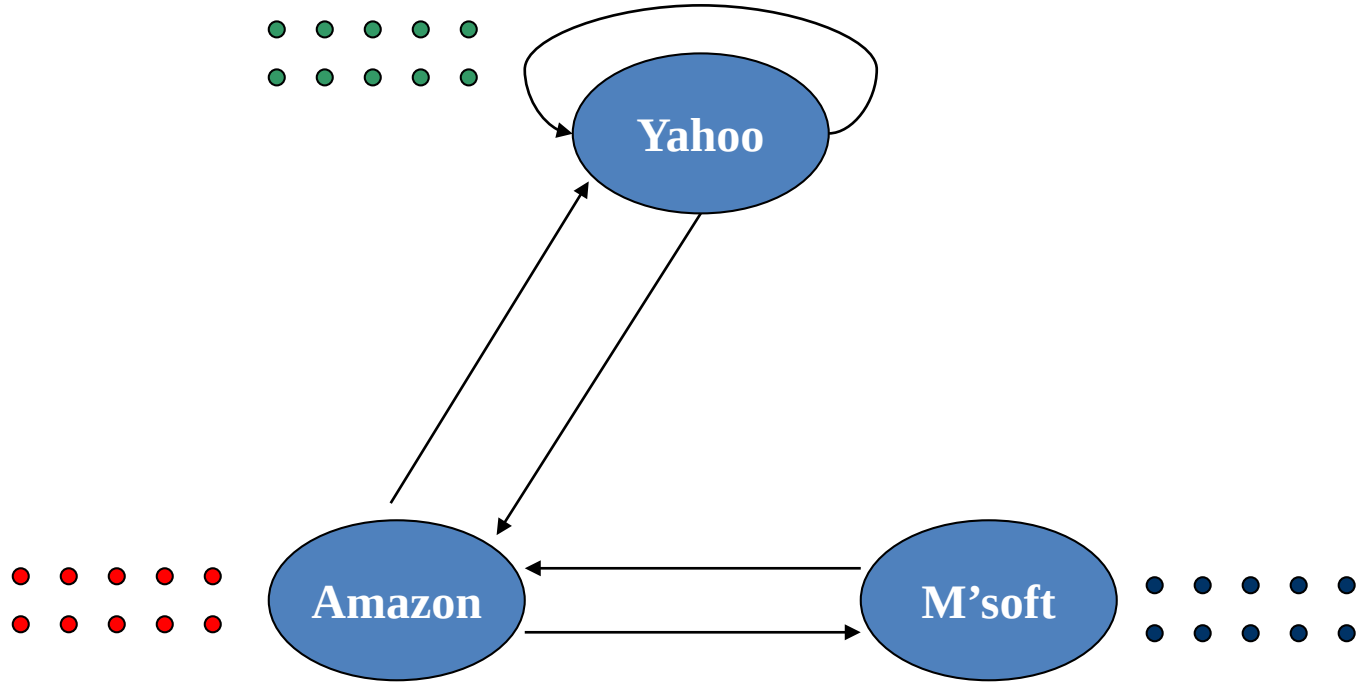
$y \leftarrow y/2 + a/2$
$a \leftarrow y/2 + m$
$m \leftarrow a/2$

	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

y		1/3	1/3	5/12	3/8		2/5
a	=	1/3	1/2	1/3	11/24	. . .	2/5
m		1/3	1/6	1/4	1/6		1/5

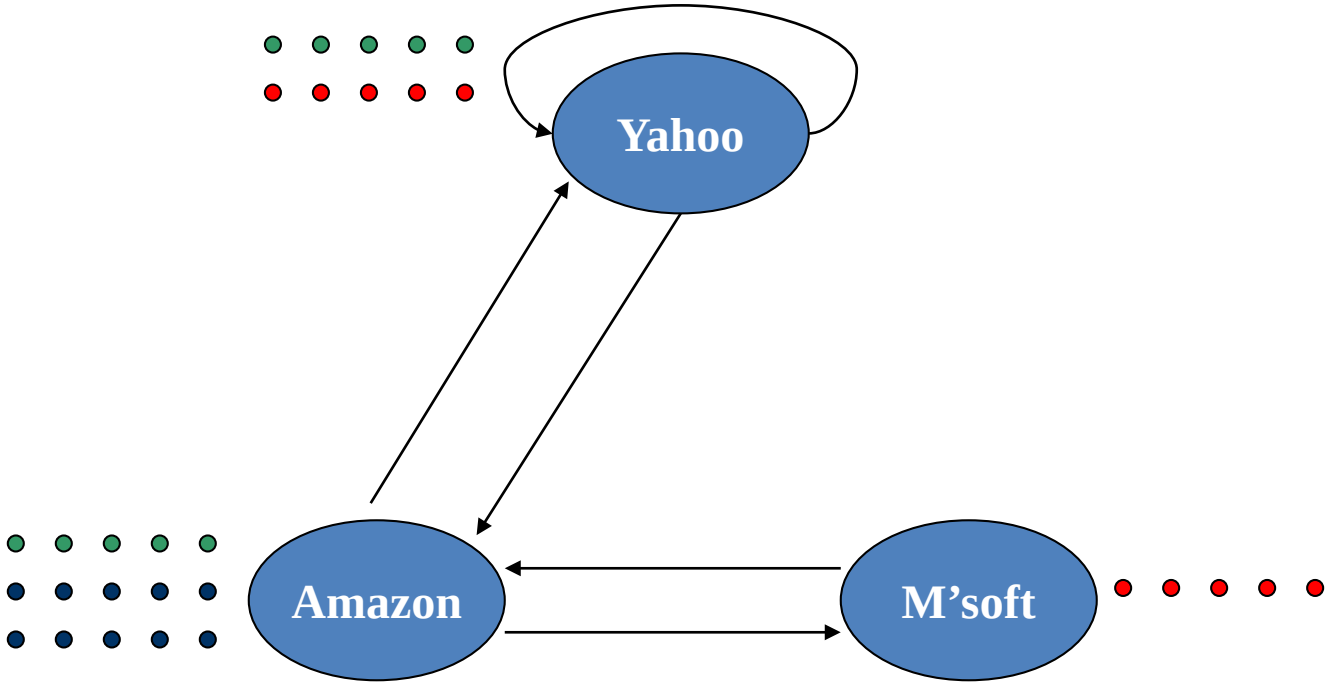
2.2 PageRank求解

- 迭代示意



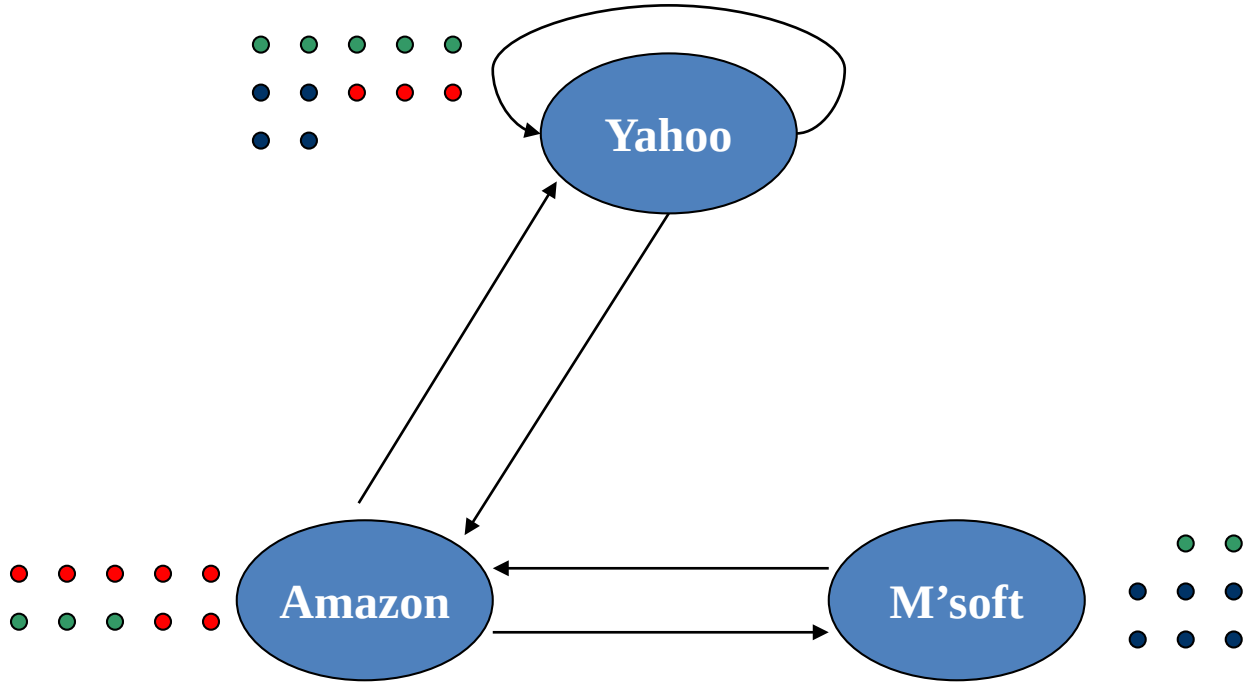
2.2 PageRank求解

- 迭代示意



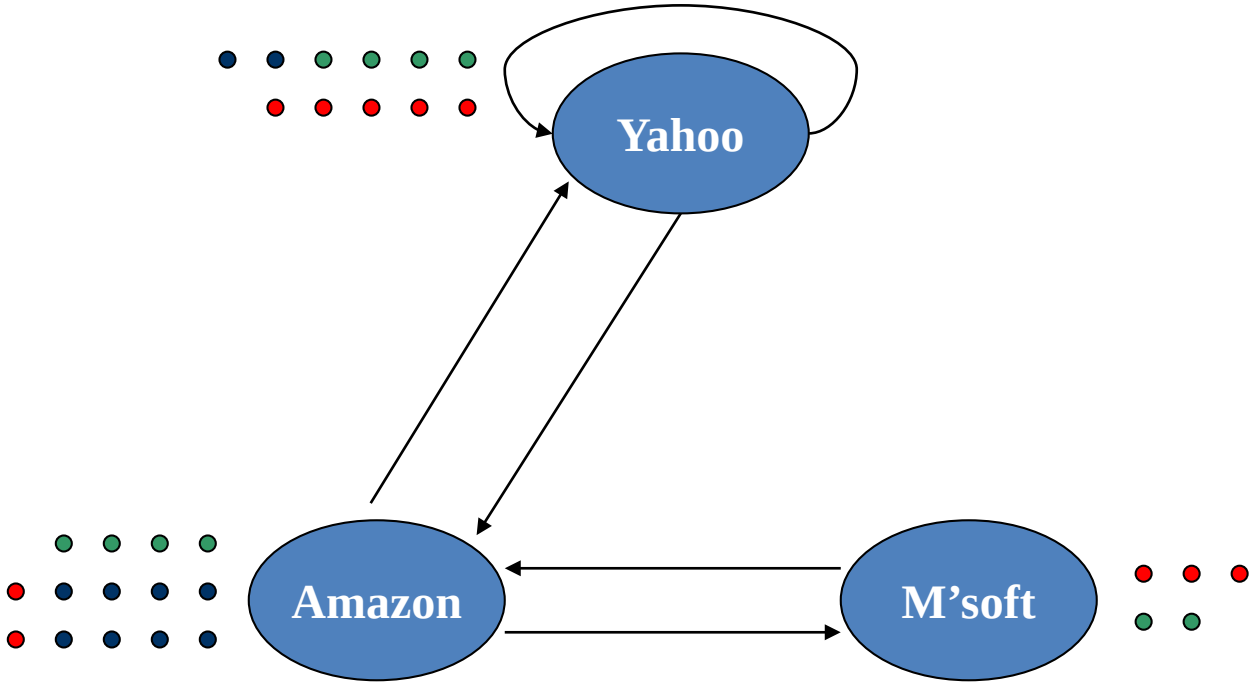
2.2 PageRank求解

- 迭代示意



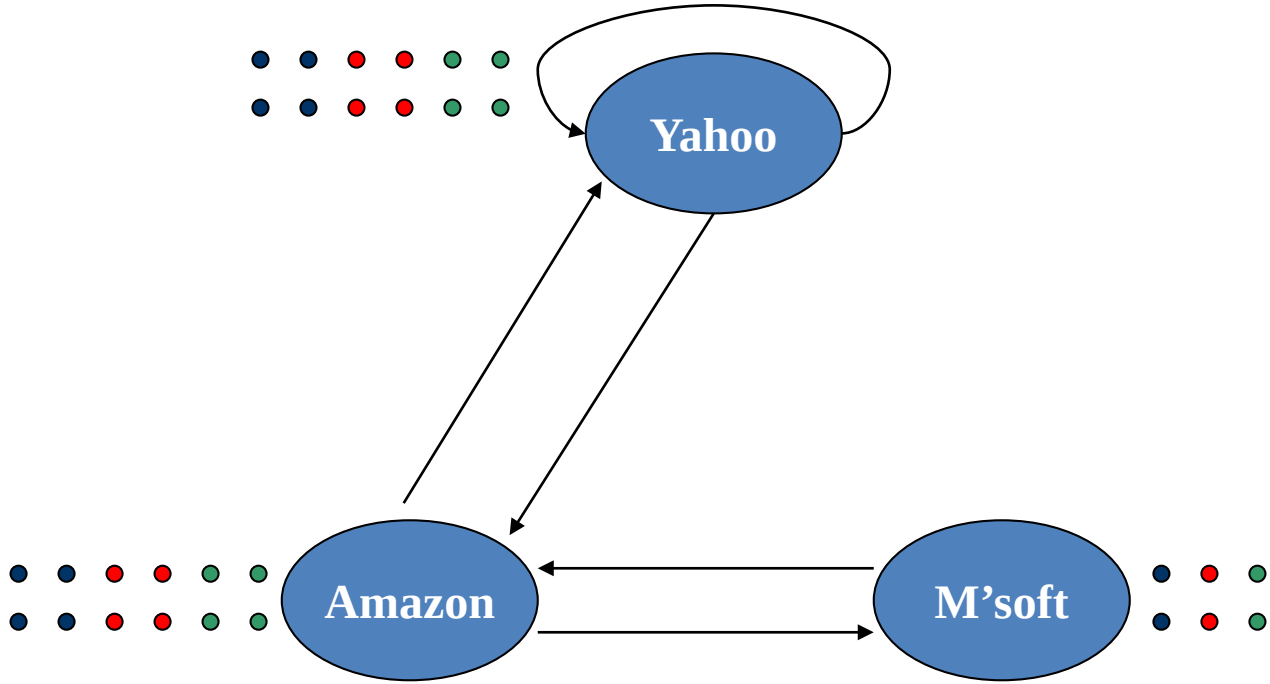
2.2 PageRank求解

- 迭代示意



2.2 PageRank求解

- 迭代示意



2 PageRank计算

2.1 转移矩阵

2.2 PageRank求解

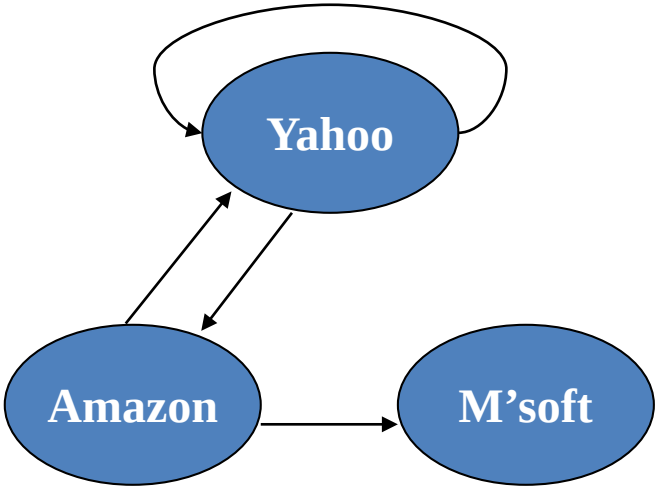
2.3 终止节点

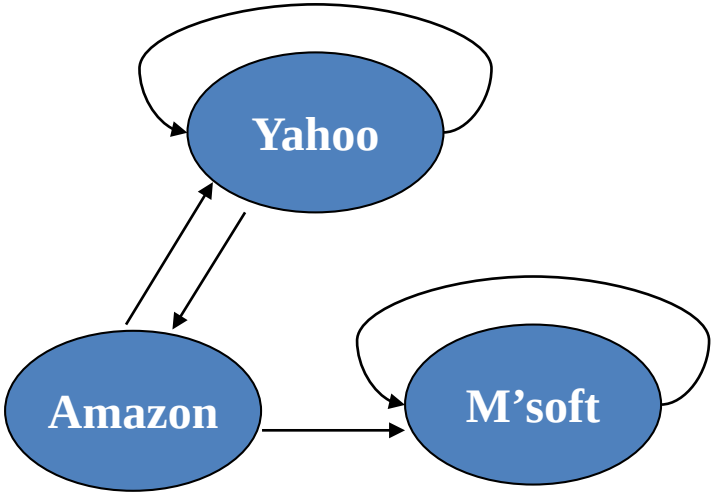
2.4 抽税法

2.5 面向主题的PageRank

2.3 终止节点

- 无出度的节点
- 采集器陷阱

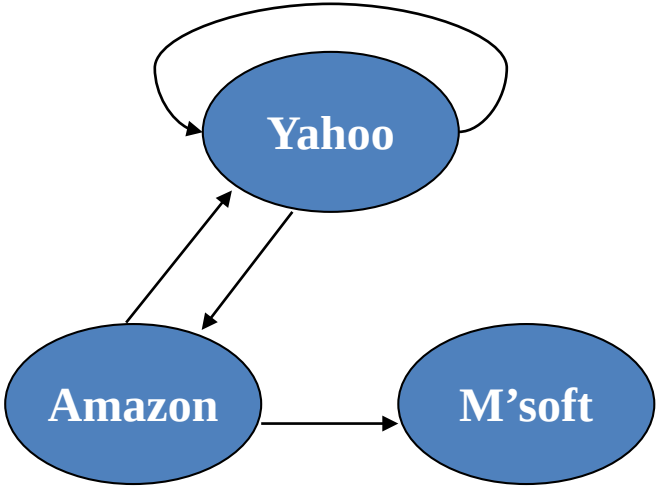




2.3 终止节点

- **无出度的节点**

$$y \leftarrow y/2 + a/2$$
$$a \leftarrow y/2$$
$$m \leftarrow a/2$$

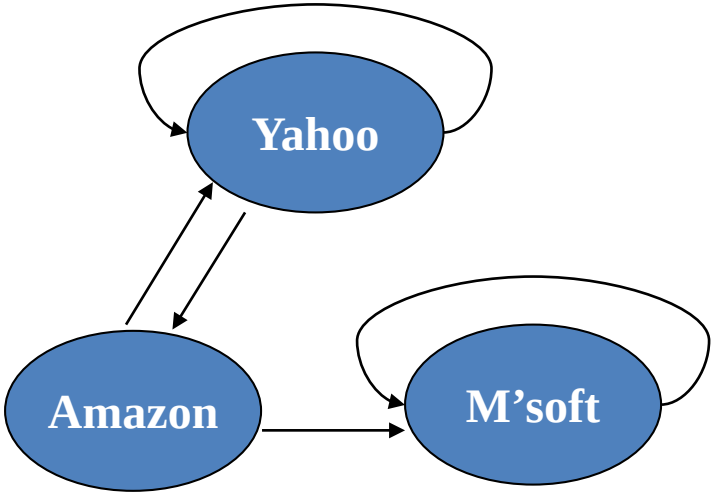


$$\begin{matrix} y \\ a \\ m \end{matrix} = \begin{matrix} 1/3 \\ 1/3 \\ 1/3 \end{matrix} \quad \begin{matrix} 1/3 \\ 1/6 \\ 1/6 \end{matrix} \quad \begin{matrix} 1/4 \\ 1/6 \\ 1/12 \end{matrix} \quad \begin{matrix} 5/24 \\ 1/8 \\ 1/12 \end{matrix} \quad \cdots \quad \begin{matrix} 0 \\ 0 \\ 0 \end{matrix}$$

2.3 终止节点

- **采集器陷阱**

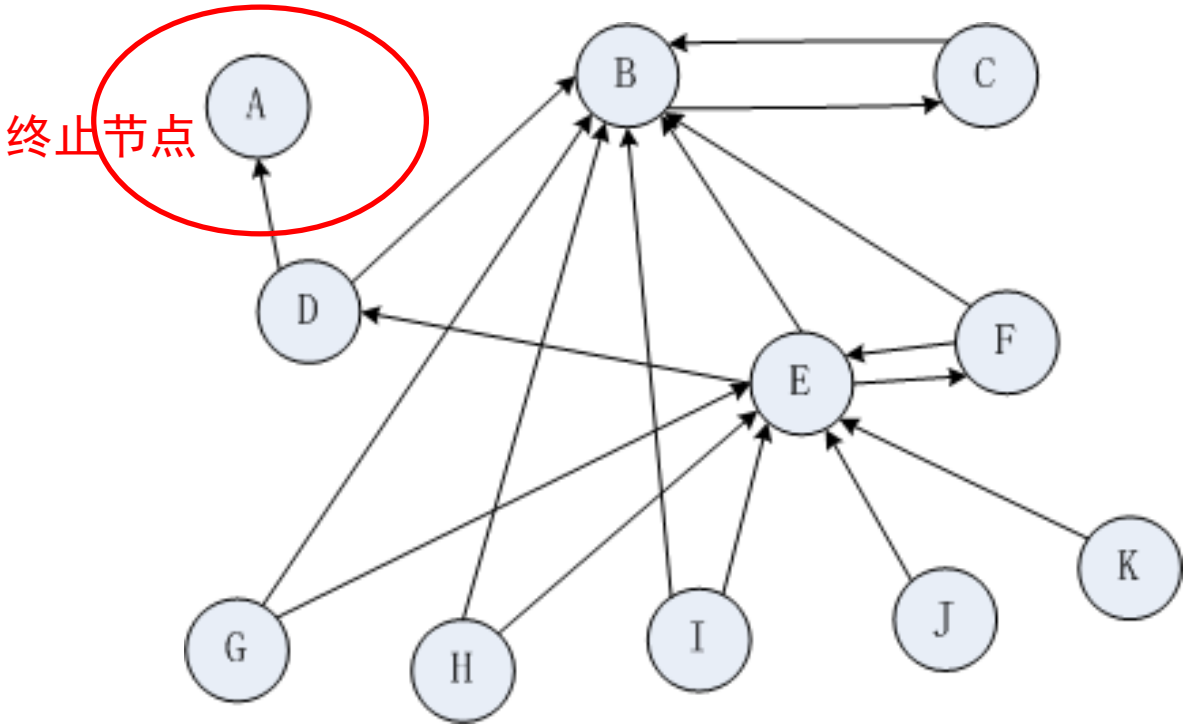
$$y \leftarrow y/2 + a/2$$
$$a \leftarrow y/2$$
$$m \leftarrow a/2 + m$$



y		1/3	1/3	1/4	5/24		0
a	=	1/3	1/6	1/6	1/8	. . .	0
m		1/3	1/2	7/12	2/3		1

2.4 抽税法

- 如何计算？

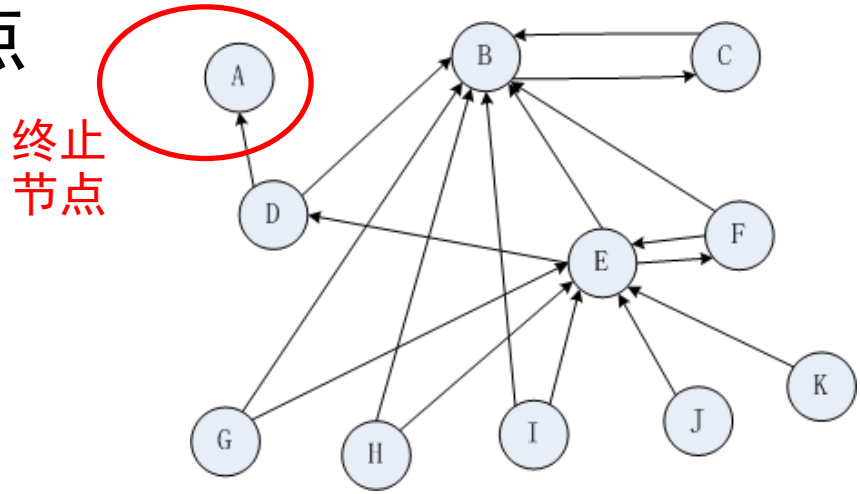


内容回顾

2.1 转移矩阵

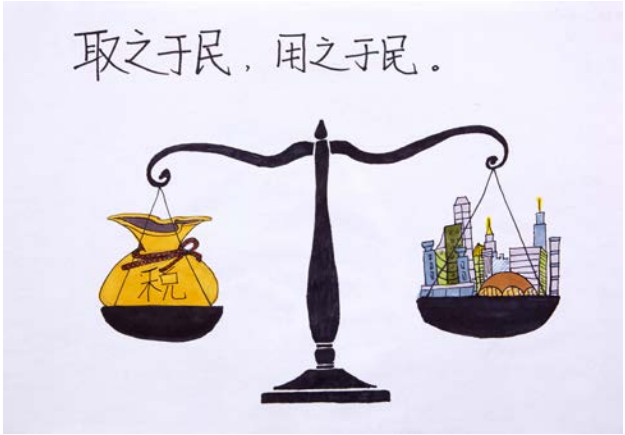
2.2 PageRank求解

2.3 终止节点



2.4 抽税法

- **抽税的基本思想**

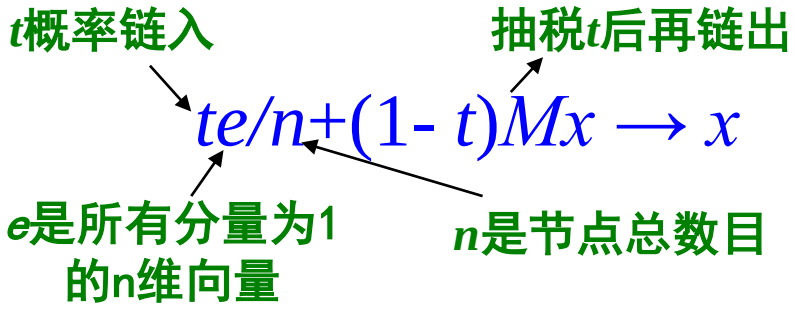


取之于民：每个节点都抽取一定比例的链出

用之于民：每个节点平均分配总税收作链入

2.4 抽税法

- **每次迭代固定税率 t**



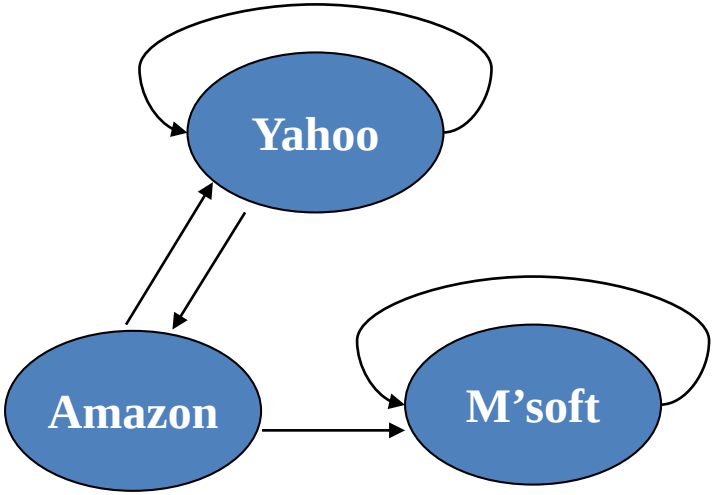
- 抽税：以t的概率随机链出致任意网页
- 剩余$(1-t)$的概率按照转移矩阵M链出
- 每个页面有 te/n的概率链入

2.4 抽税法

- **抽税示例（$t = 0.2$）**

$0.8Mx+0.2/3 \rightarrow x$

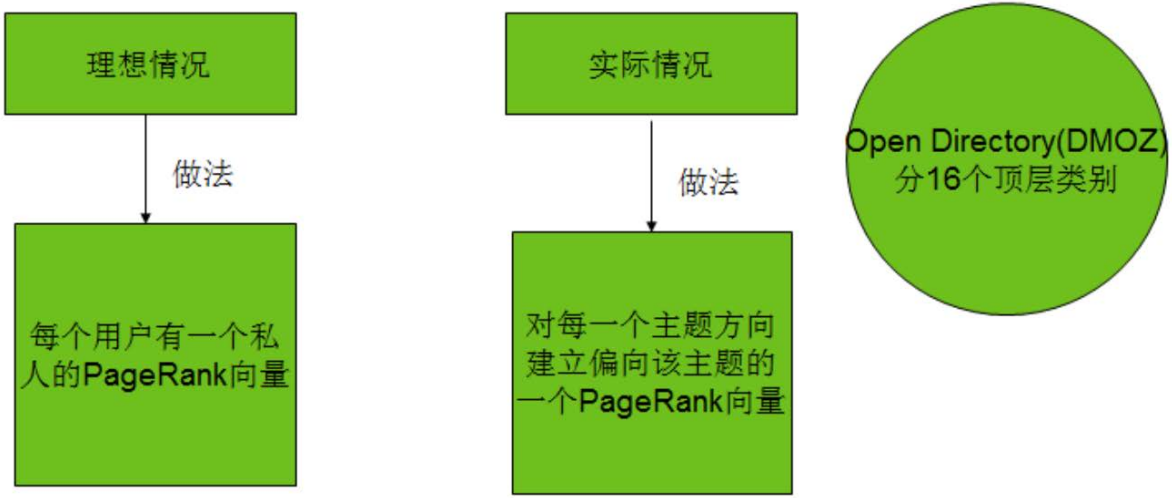
$y \leftarrow 0.8(y/2 + a/2) + 0.067$
$a \leftarrow 0.8(y/2) + 0.067$
$m \leftarrow 0.8(a/2 + m) + 0.067$



y		1/3	0.33	0.24	0.2		0.21
a	=	1/3	0.20	0.20	0.145	. . .	0.15
m		1/3	0.46	0.52	0.56		0.64

2.5 面向主题的PageRank

- **应用需求**
  - 不同的人有不同的兴趣
  - 搜索引擎能推断出不同用户的兴趣，体验会更棒



2.5 面向主题的PageRank

- **解决思路**
  - 获得各个网页的主题
  - 构建面向主题的PageRank
    - ✓随机跳转只能到达给定主题的页面
  - 迭代算式

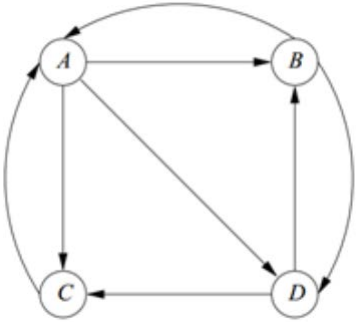
给定主题的页面集合S

$$x = (1- t)Mx + te_s/|S|$$

分量对应的页面属于S，
则分量为1，否则为0

2.5 面向主题的PageRank

- 示例
  - 假设S={B, D}, t=0.2



M = ?

$$(1-t)M = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix}$$

e_s = ?

x = ?

$$\boldsymbol{x} = \begin{bmatrix} \mathbf{0} & \mathbf{2/5} & \mathbf{4/5} & \mathbf{0} \\ \mathbf{4/15} & 0 & \mathbf{0} & \mathbf{2/5} \\ \mathbf{4/15} & 0 & \mathbf{0} & \mathbf{2/5} \\ \mathbf{4/15} & \mathbf{2/5} & 0 & \mathbf{0} \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} \mathbf{0} \\ \mathbf{1/10} \\ \mathbf{0} \\ \mathbf{1/10} \end{bmatrix}$$

主要内容

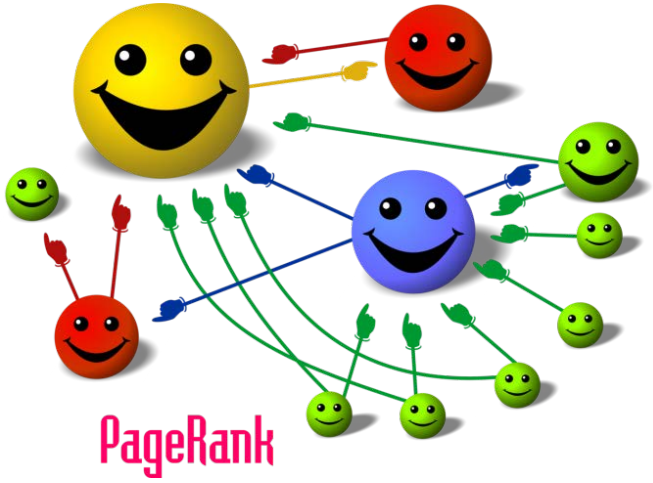
1 问题的引出

2 PageRank计算

3 链接作弊

4 导航页与权威页

5 小结



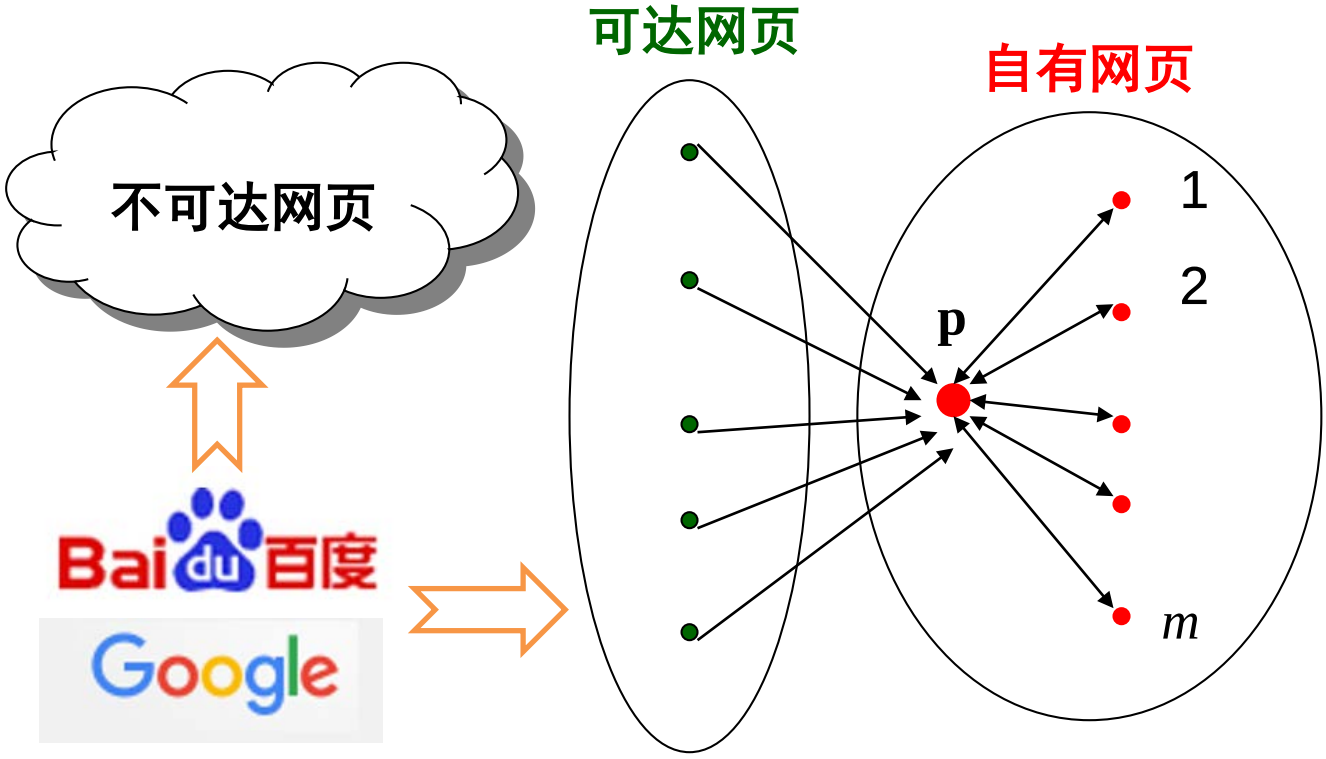
3 链接作弊

3.1 自建网页的发布

3.2 垃圾农场

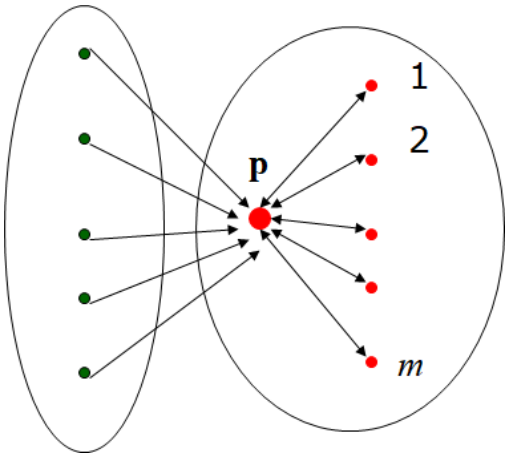
3.3 TrustRank

3.1 自建网页的发布



3.2 垃圾农场

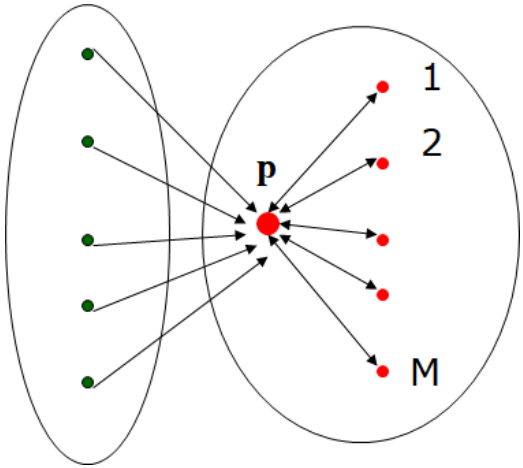
- PageRank值分析



- 从可达网页获取的PR值可记作x（已知）
- 网页P的PR值为y (未知)
- 抽税比率为t
- 每个农场网页PR值$(1-t)y/m + t/n$

3.2 垃圾农场

- PageRank值分析



每个农场网页PR值

$$y = x + (1-t)m[(1-t)y/m + t/n] + t/n$$

$$y \approx x + (1-t)^2 y + t(1-t)m/n$$

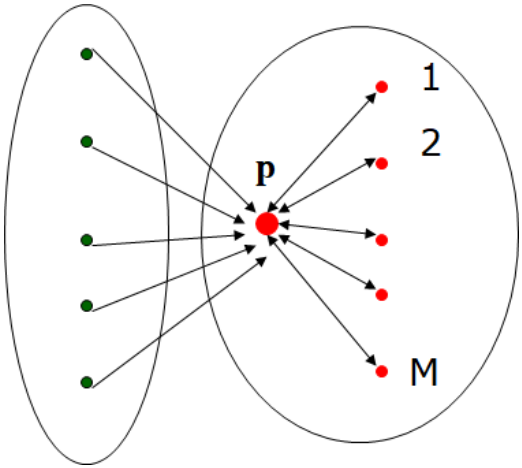
$$y = x/t(2-t) + t(1-t)m/t(2-t)n$$

值很小，可忽略

3.2 垃圾农场

- **PageRank值分析**

$y = x/t(2-t) + t(1-t)m/t(2-t)n$



- $t(1-t)m/t(2-t)n \approx m/2n$
- $x/t(2-t) = x/[-(t-1)^2+1]$

$t \to 0$? 相当于不抽税，P成为终止节点

$t=0.2$? $x/t(2-t) \approx 2.8x$

3. 3 TrustRank

- **与作弊做斗争**
  - 检测垃圾农场，删除之
  - 跟踪SEO(Search Engine Optimization)技术
  - 构造可信PR值计算策略

3. 3 TrustRank

- **基本思路**
  - 构建可信赖的可靠网页集合S
    - ✓人工标注Top-K排名的网页
    - ✓选择成员受限的域名集合(edu、org、…)
  - 仅可信网页分享抽税值

$$x = (1- t)Mx + te_s/|S|$$

主要内容

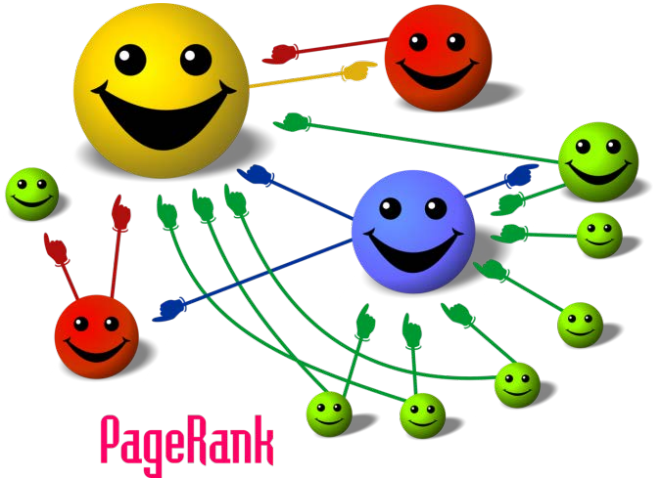
1 问题的引出

2 PageRank计算

3 链接作弊

4 导航页与权威页

5 小结



4 导航页与权威页

4.1 导航页

4.2 权威页

4.3 HITS算法

4.1 导航页

- **提供重要信息的入口，不提供信息本身**

4.2 权威页

- 提供重要的信息

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Article Talk

HITS algorithm

From Wikipedia, the free encyclopedia
(Redirected from Hubs and authorities)

Hyperlink-Induced Topic Search (**HITS**; also known as **hubs and authorities**) is a link analysis algorithm and Authorities stemmed from a particular insight into the creation of web pages when the Internet was orig directories that were not actually authoritative in the information that they held, but were used as compilatio authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and hubs.[1]

The scheme therefore assigns two scores for each page: its authority, which estimates the value of the cont other pages.

Contents [hide]

1 History
1.1 In journals
1.2 On the Web

4.3 HITS算法

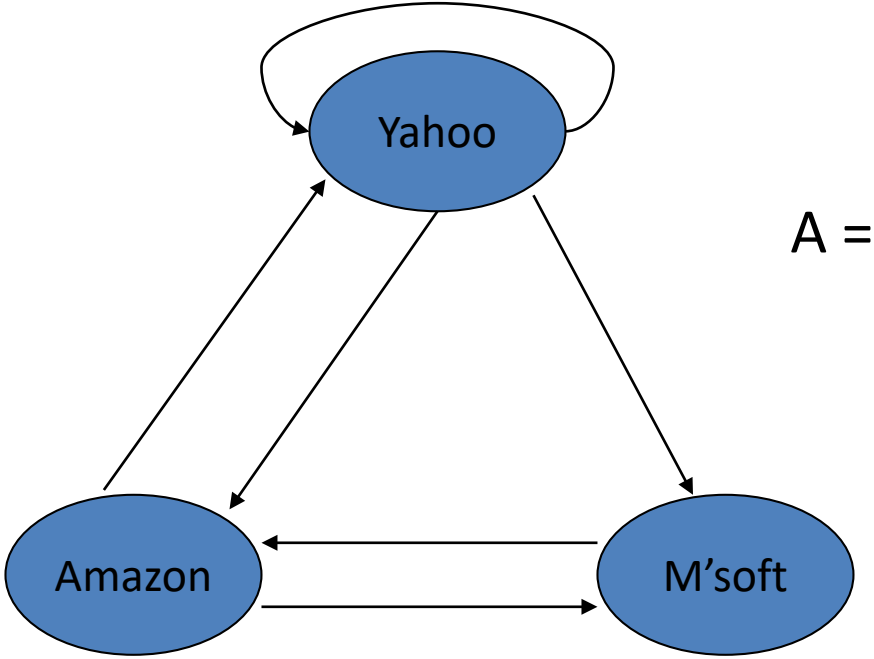
- **度量指标**
  - 导航度：充当导航页的良好程度(h)
    - ✓累加所有链出网页的权威度来估计导航页的导航度
  - 权威度：充当权威页的良好程度(a)
    - ✓累加所有链入网页的导航度来估计权威页的权威度

4.3 HITS算法

- **转移矩阵A**
  - $A[i, j] = 1$ 如果页面i链接到页面j，否则 $A[i, j] = 0$
  - A^T 与PageRank转移矩阵M类似， A^T 元素仅为0/1，M元素可为小数
  - 列向量h、a分别表示导航度和权威度
  - h、a的计算与PageRank类似

4.3 HITS算法

- **转移矩阵A示例**



$$A = \begin{array}{c|ccc} & y & a & m \\ \hline y & 1 & 1 & 1 \\ a & 1 & 0 & 1 \\ m & 0 & 1 & 0 \end{array}$$

4.3 HITS算法

- **量化计算**

  - 导航度正比于所有链出网页的权威度之和

$$\mathbf{h} = \lambda A\mathbf{a}$$

  - 权威度正比于所有链入网页的导航度之和

$$\mathbf{a} = \mu A^T\mathbf{h}$$

$$\mathbf{h} = \lambda A\mathbf{a} = \lambda A\mu A^T\mathbf{h} \qquad \mathbf{a} = \mu A^T\mathbf{h} = \mu A^T\lambda A\mathbf{a}$$

$$\mathbf{h} = \lambda\mu AA^T\mathbf{h} \qquad \mathbf{a} = \mu\lambda A^TA\mathbf{a}$$

4.3 HITS算法

- 示例　　$\mathbf{a} = \lambda\mu A^T A\mathbf{a}$; $\mathbf{h} = \lambda\mu AA^T\mathbf{h}$

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad A^T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad AA^T = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad A^TA = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

a(yahoo)	=	1	5	24	114	· · ·	$1+\sqrt{3}$
a(amazon)	=	1	4	18	84	· · ·	2
a(microsoft)	=	1	5	24	114	· · ·	$1+\sqrt{3}$
h(yahoo)	=	1	6	28	132	· · ·	1.000
h(amazon)	=	1	4	20	96	· · ·	0.735
h(microsoft)	=	1	2	8	36	· · ·	0.268

主要内容

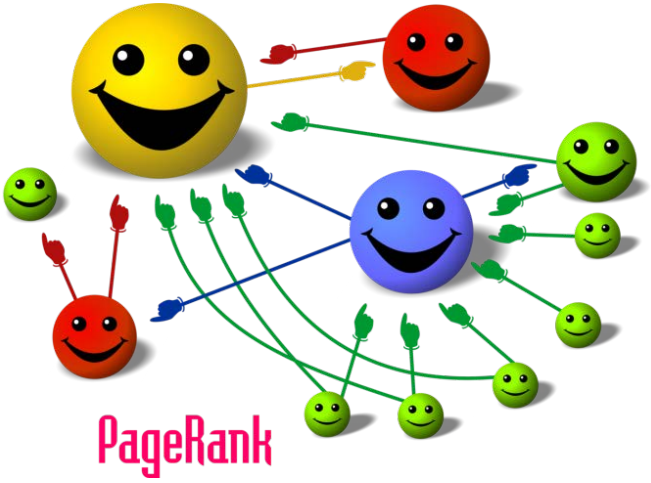
1 问题的引出

2 PageRank计算

3 链接作弊

4 导航页与权威页

5 小结



5 小结

1 问题的引出

2 PageRank计算

3 链接作弊

4 导航页与权威页