

信息检索导论

An Introduction to Information Retrieval

第19讲 链接分析

Link Analysis

授课人：古晓艳

中国科学院信息工程研究所/国科大网络空间安全学院

*改编自“An introduction to Information retrieval”网上公开的课件，地址 <http://nlp.stanford.edu/IR-book/>

提纲

- ① 上一讲回顾
- ② 锚文本
- ③ 引用分析
- ④ PageRank
- ⑤ HITS: Hub节点&Authority节点
- ⑥ Web作弊与反作弊

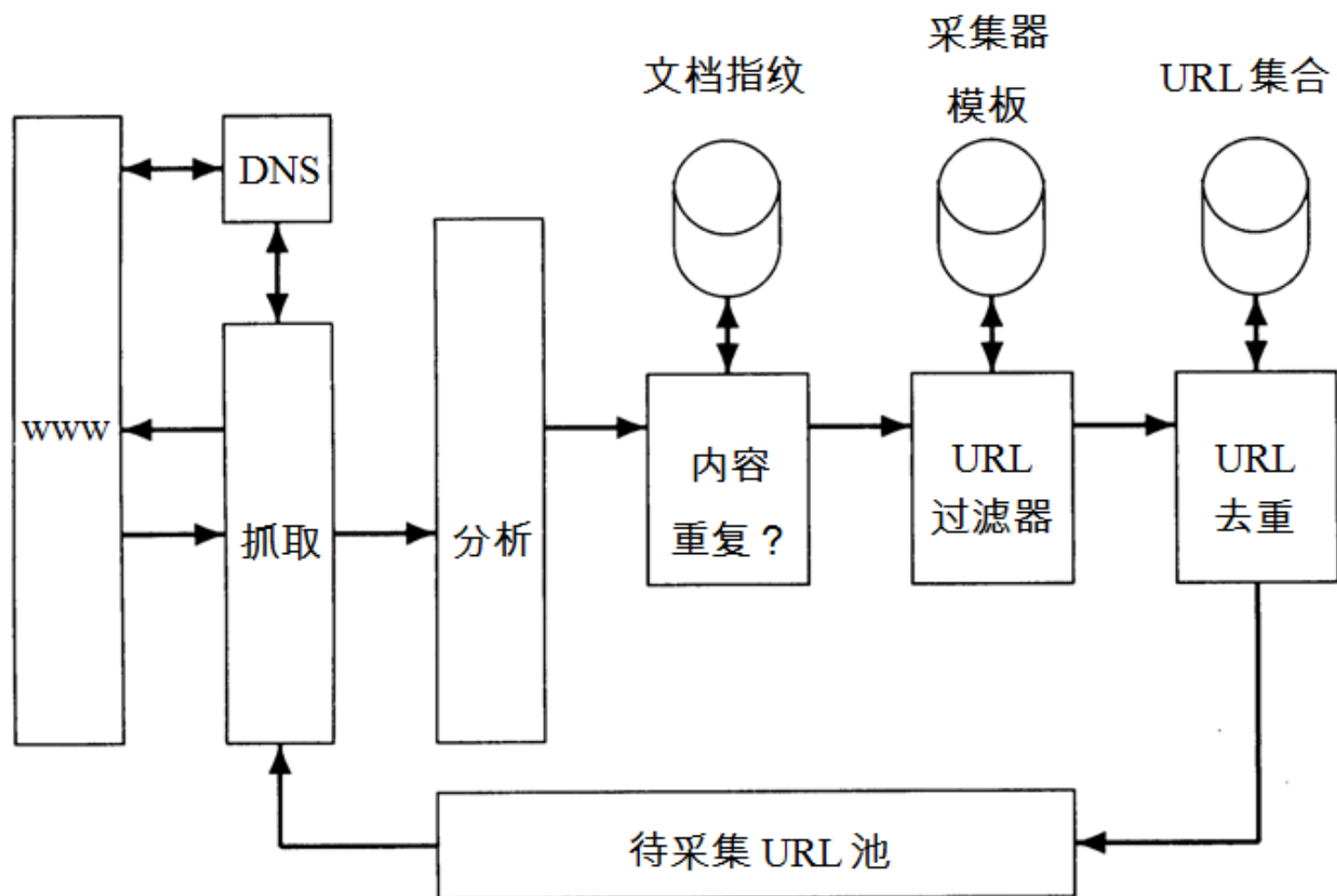
提纲

- ① 上一讲回顾
- ② 锚文本
- ③ 引用分析
- ④ PageRank
- ⑤ HITS: Hub节点&Authority节点
- ⑥ Web作弊与反作弊

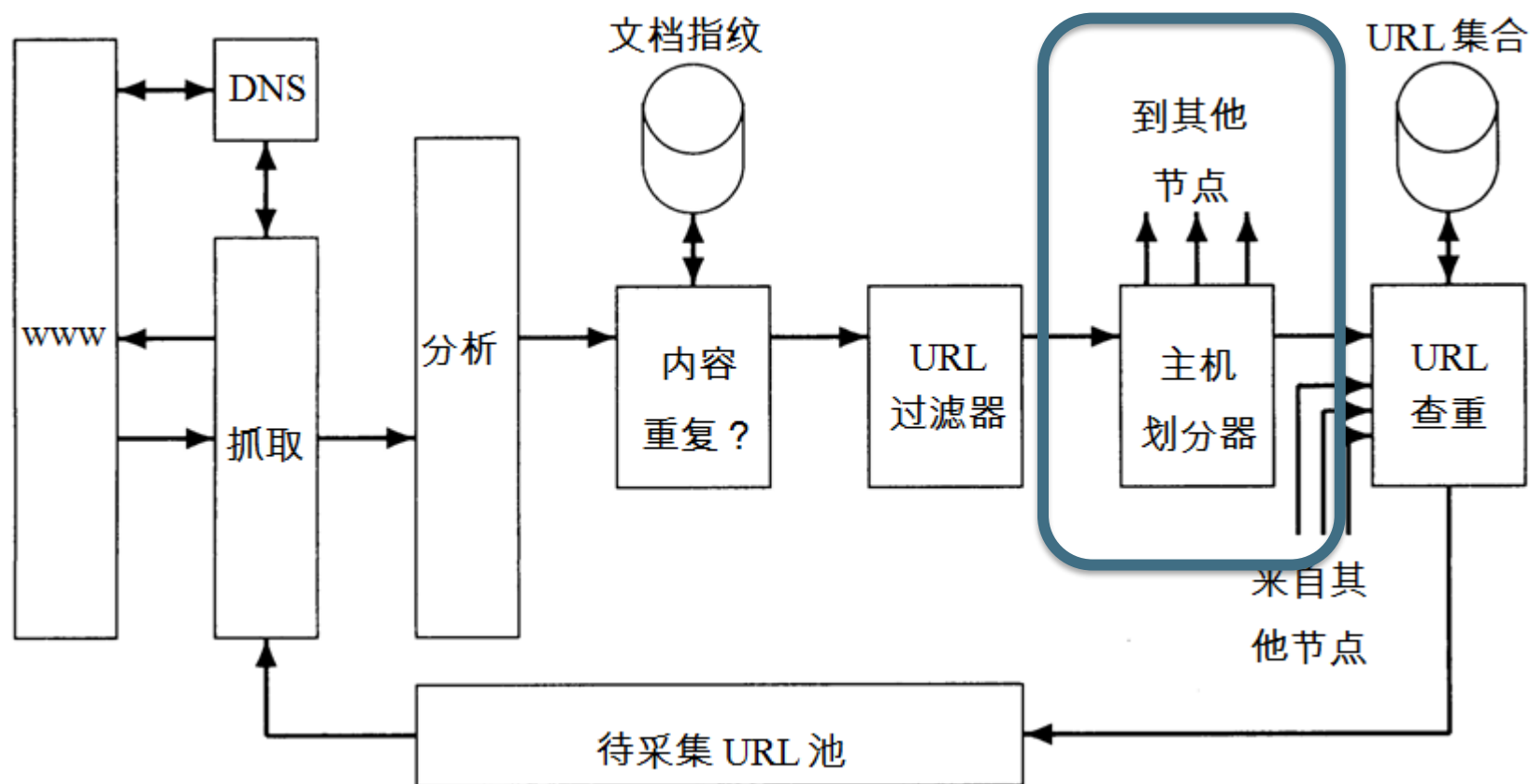
基本的采集过程

- 初始化采集URL种子队列;
- 重复如下过程:
 - 从队列中取出URL
 - 下载并分析网页
 - 从网页中抽取更多的URL
 - 将这些URL放到队列中
- 这里有个“Web的连通性很好”的基本假设

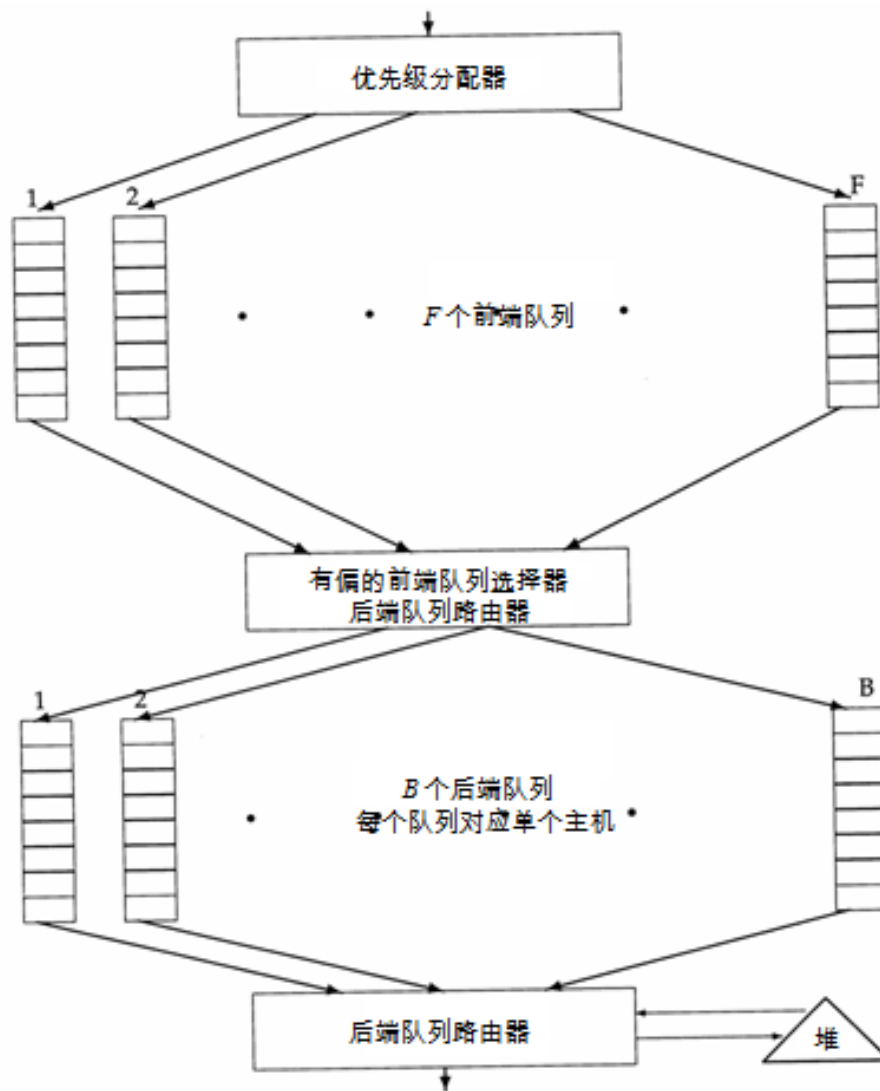
基本的采集架构



分布式采集器



Mercator 中的待采集URL缓冲池



分布式索引

- 随着文档规模不断增长，IR系统难以管理
- 检索和索引的成本随着文档规模增大而增加
- 文档集合越大，检索响应时间越长
- 系统管理的文档数不断增加，性能不断下降，直至系统不可用
- 需要采用分布式架构和算法

文档集分区方法

- 文档分区法（横向切分法）
 - N个文档在系统的P个处理器/节点间进行分布
 - P个文档子集，每个子集包含文档数 N/P
 - 查询过程中，每个处理器/节点仅处理其文档子集
 - 每个文档子集的结果集汇聚后形成最终结果
- 词项分区法（纵向切分法）
 - 将词项分布到P个处理器/节点上
 - 针对每个文档的检索，将分布到多个处理器/节点上进行

总体对比

- 文档分区法的主要缺点：
 - 需要向文档子集发起很多不必要的查询操作，这些文档集中可能基本没有相关文档
- 词项分区法的主要缺点：
 - 需要构建和维护全局索引，限制了该方法的可扩展性
- 此外，词项分区法的响应时间的波动范围较大，需要引入复杂的均衡机制以保证响应时间

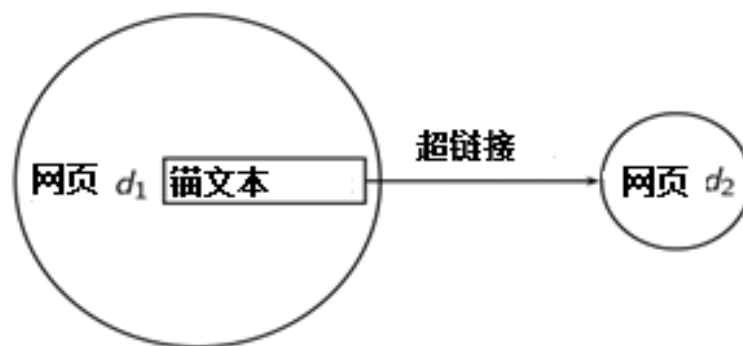
本讲内容

- 锚文本: Web上的链接相关信息为什么对IR有用?
- 引用分析(Citation analysis): PageRank及其他基于链接排序方法的数学基础
- PageRank : 一个著名的基于链接分析的排序算法(Google)
- HITS : 另一个著名的基于链接分析的排序算法(IBM)

提纲

- ① 上一讲回顾
- ② 锚文本
- ③ 引用分析
- ④ PageRank
- ⑤ HITS: Hub节点&Authority节点
- ⑥ Web作弊与反作弊

Web可以看成一个有向图



- 假设1: 超链接代表了某种质量认可信号
 - 超链 $d_1 \rightarrow d_2$ 表示 d_1 的作者认可 d_2 的质量和相关性
- 假设 2: 锚文本描述了文档 d_2 的内容
 - 这里的锚文本定义比较**宽泛**，包括链接周围的文本
 - 例子: “You can find cheap cars here .”
 - 锚文本: “You can find cheap cars here”

$[d_2 \text{ 中 文 本}] \text{ vs. } [d_2 \text{ 中 文 本}] + [\text{锚文本} \rightarrow d_2]$

- 后者往往效果好于前者

- 例子: 查询 *IBM*
 - IBM 的版权页匹配上
 - 很多作弊网页匹配上
 - IBM的wikipedia页面
 - 可能与IBM 的主页并不匹配!
 - ... 也许 IBM 的主页上大部分都是图

- 而按照 $[\text{锚文本} \rightarrow d_2]$ 来搜索效果会比较好
 - 这种表示下, 出现IBM最多的是其主页 www.ibm.com

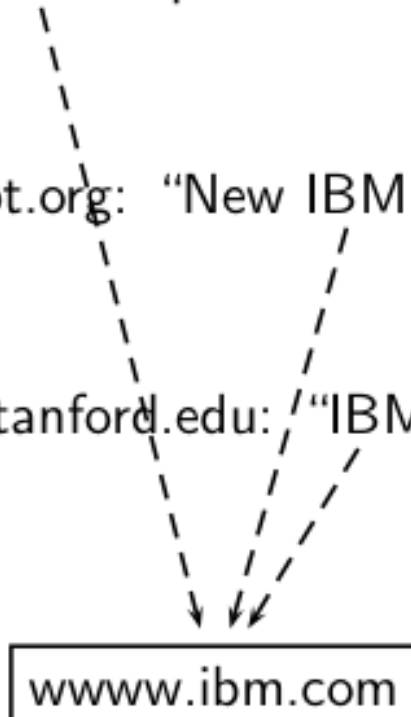
指向www.ibm.com的很多锚文本中包含IBM

www.nytimes.com: "IBM acquires Webify"

www.slashdot.org: "New IBM optical chip"

www.stanford.edu: "IBM faculty award recipients"

www.ibm.com



对锚文本构建索引

- 因此，锚文本往往比网页本身更能揭示网页的内容
- 在计算过程中，锚文本应该被赋予比文档中文本更高的权重

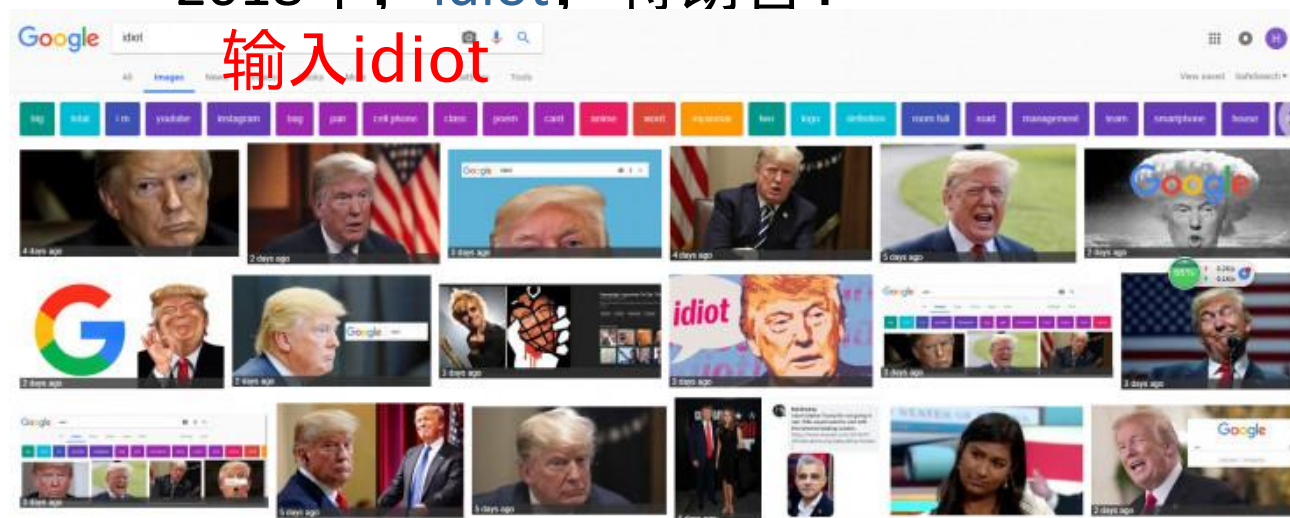
课堂练习：PageRank背后的假设

- 假设1: Web上的链接是网页质量的标志—链出网页的作者认为链向的网页具有很高的质量
- 假设2: 锚文本能够描述链向网页的内容

- 通常情况下假设1是否成立？
- 通常情况下假设2是否成立？

Google炸弹(Google bomb)

- Google炸弹是指由于人为恶意构造锚文本而导致的搜索结果很差的搜索
- 典型案例：
 - 2003年10月，miserable failure，George W. Bush?
 - 2018年，idiot，特朗普？



- The 10 Most Incredible Google Bomb
(<https://www.searchenginepeople.com/blog/incredible-google-bombs.html>)

提纲

- ① 上一讲回顾
- ② 锚文本
- ③ 引用分析
- ④ PageRank
- ⑤ HITS: Hub节点&Authority节点
- ⑥ Web作弊与反作弊

PageRank的起源: 引用分析(1)

- 引用分析: 科技文献中的引用分析
- 一个引用的例子: “[Miller \(2001\)](#) has shown that physical activity alters the metabolism of estrogens.”
- 可以把“Miller (2001)”看成是两篇学术文献之间的超链接
- 在科技文献领域使用这些“超链接”的一个应用:
 - 根据他人引用的重合率来度量两篇文献的相似度, 这称为共引相似度
 - 在Web上也存在共引相似度: Google中提供的“find pages like this”或者“Similar”功能

Generative models. Recently, unsupervised learning has been making a lot of progress on image generation. Typically, a parametrized mapping is learned between a predefined random noise and the images, with either an autoencoder [4, 22, 29, 40, 62] a generative adversarial network (GAN) [20] or more directly with a reconstruction loss [6]. Of particular interest, the discriminator of a GAN can produce visual features, but their performance are relatively disappointing [15]. Donahue *et al.* [15] and Dumoulin *et al.* [17] have shown that adding an encoder to a GAN produces visual features that are much more competitive.

PageRank起源: 引用分析(2)

- 另一个应用: 引用频率可以用来度量一篇文档的影响度
 - 最简单的度量指标: 每篇文档都看成一个投票单位, 引用可以看成是投票, 然后计算一篇文档被投票的票数。当然这种方法不太精确。

- 在V Bert: Pre-training of deep bidirectional transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv ..., 2018 - arxiv.org
We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and ...
☆ 99 被引用次数: 7390 相关文章 所有 25 个版本

2020.06

- 更好 Bert: Pre-training of deep bidirectional transformers for language understanding
J Devlin, MW Chang, K Lee, K Toutanova - arXiv preprint arXiv ..., 2018 - arxiv.org
We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations ...
☆ 99 被引用次数: 13780 相关文章 所有 25 个版本

率进行加权

2020.12

这里可以采用良好的

形式化定义

PageRank的起源: 引用分析(3)

- 更好的度量方法: 加权的引用频率
- 这就是PageRank的基本思路
- PageRank 最早起源于1960年代Pinsker和Narin提出的引用分析
- 引用分析不是小事情, 在美国, 任何教职人员的薪水取决于其发表文章的影响力!



ELSEVIER 2020年5月第六次发布“中国高被引学者”榜单

提纲

- ① 上一讲回顾
- ② 锚文本
- ③ 引用分析
- ④ PageRank
- ⑤ HITS: Hub节点&Authority节点
- ⑥ Web作弊与反作弊

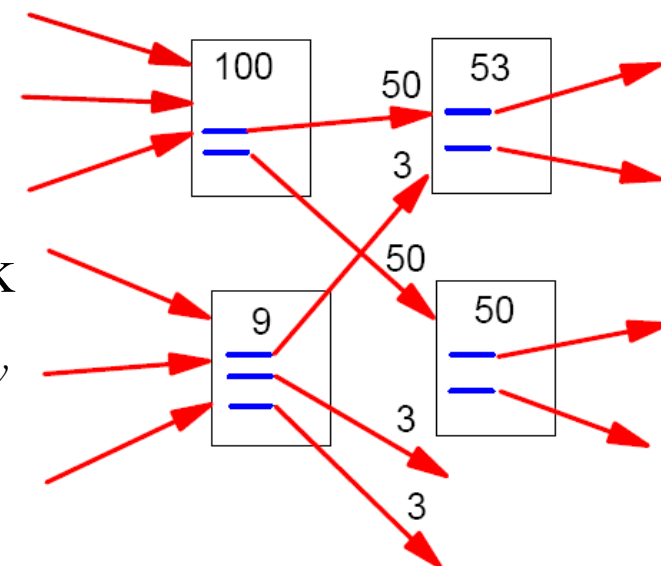
PageRank的含义

- 假设一个Web使用者，在Web上随机游走（浏览）
 - 开始于一个随机页面
 - 在接下来每一步，以相同的概率访问当前页面所指向的其它链接
- 在稳态情况下，每个页面都有一个长时访问频率（long-term visit rate）
- 这个长时访问频率即为网页的PageRank值
- $\text{PageRank} = \text{长时访问频率} = \text{稳态概率}$

原始的PageRank公式

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

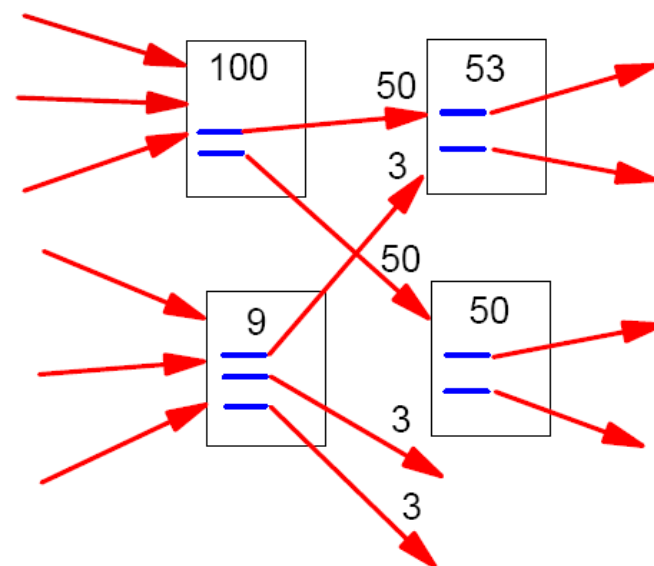
$R(u)$ 和 $R(v)$ 是分别是网页 u 、 v 的PageRank值， B_u 指的是指向网页 u 的网页集合、 N_v 是网页 v 的出链数目。



一个网页的PageRank等于所有的指向它的网页的PageRank的分量之和(c 为归一化参数)。网页的每条出链上每个分量上承载了相同的PageRank分量。

PageRank的特点

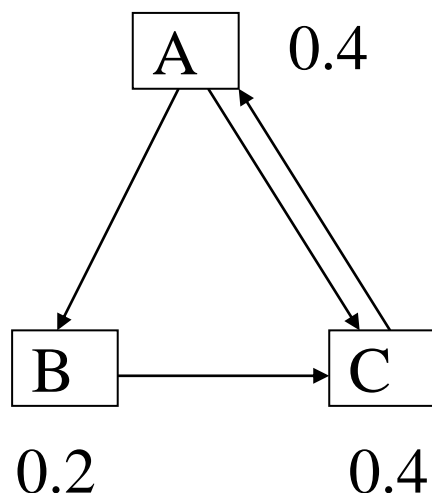
- (1) 一个网页如果它的入链越多, 那么它也越重要(PageRank越高);
- (2) 一个网页如果被越重要的网页所指向, 那么它也越重要(PageRank越高)。



类比: 微博粉丝

简单计算的例子($c=1$)

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$



$$R(A)=R(C)$$

$$R(B)=0.5R(A)$$

$$R(C)=R(B)+0.5R(A)$$

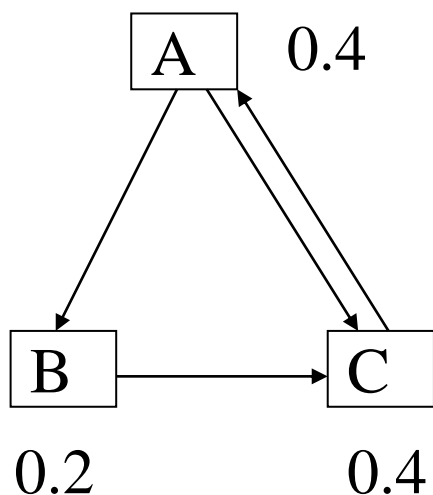
$$R(A)+R(B)+R(C)=1$$

解上述方程得:

$$R(A)=R(C)=0.4$$

$$R(B)=0.2$$

简单计算的例子($c=1$): 迭代法求解



$$R(A)=R(C)$$

$$R(B)=0.5R(A)$$

$$R(C)=R(B)+0.5R(A)$$

$$R(A)+R(B)+R(C)=1$$

迭代次数	R(A)	R(B)	R(C)
0	1/3	1/3	1/3
1	1/3	1/6	1/2
2	1/2	1/6	1/3
3	1/3	1/4	5/12
...
收敛	2/5	1/5	2/5

转化成矩阵形式

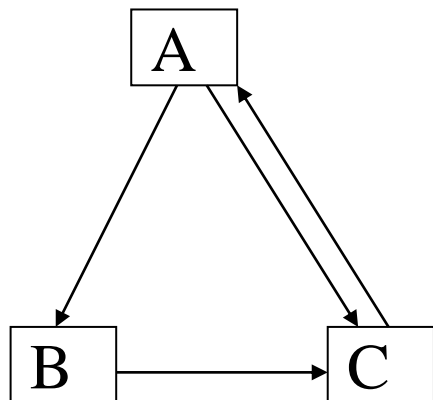
- 令 \mathbf{R} 表示所有 N 个网页的PageRank组成的列向量，令网页间的连接矩阵 $\mathbf{L}=\{l_{ij}\}$ ， P_i 有链接指向 P_j 时， $l_{ij}=1$ ，否则 $l_{ij}=0$ 。对 \mathbf{L} 的每行进行归一化，即用 P_i 的出度 N_i 去除得到矩阵 $\mathbf{A}=\{a_{ij}\}$ ， $a_{ij}=l_{ij}/N_i$ ，则有 (\mathbf{A}^T 表示 \mathbf{A} 的转置矩阵)：

$$\mathbf{R}=c\mathbf{A}^T\mathbf{R} \iff c^{-1}\mathbf{R}=\mathbf{A}^T\mathbf{R}$$

根据线性代数中有关特征向量和特征值的理论， \mathbf{R} 是矩阵 \mathbf{A}^T 的 c^{-1} 特征值对应的特征向量

$$\begin{aligned} R(A) &= R(C) \\ R(B) &= 0.5R(A) \\ R(C) &= R(B) + 0.5R(A) \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} R(A) \\ R(B) \\ R(C) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix} \begin{bmatrix} R(A) \\ R(B) \\ R(C) \end{bmatrix}$$

转化成矩阵形式



$$R(A)=R(C)$$

$$R(B)=0.5R(A)$$

$$R(C)=R(B)+0.5R(A)$$



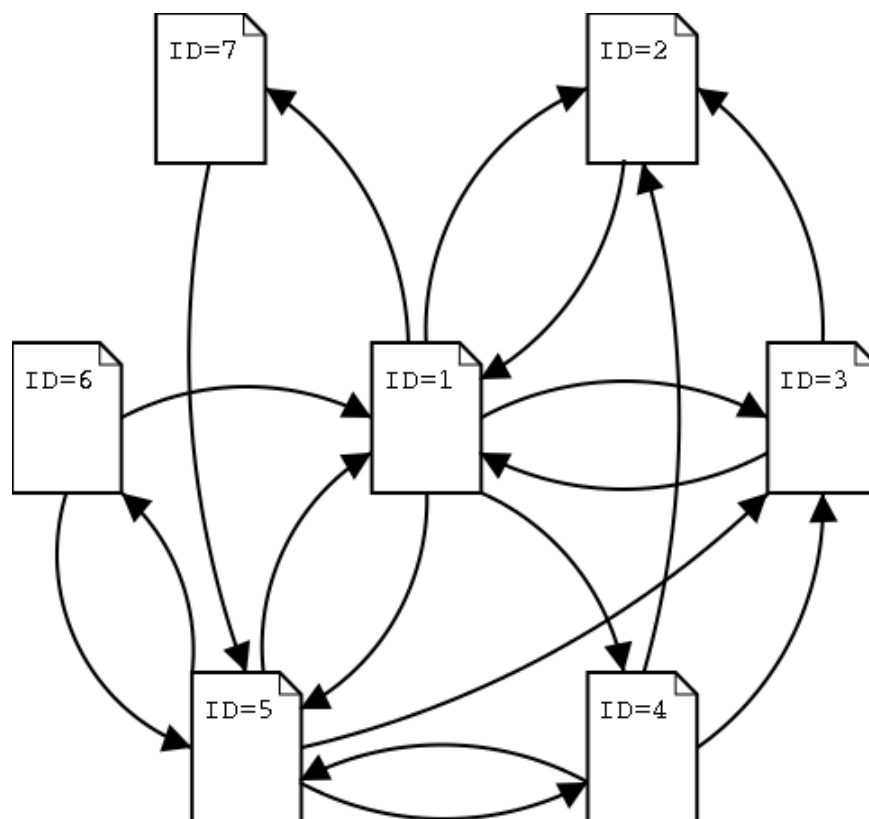
$$L=\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$$A=\begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

$$A^T=\begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix}$$

$$c^{-1}\mathbf{R}=\mathbf{A}^T\mathbf{R}$$

一个稍微复杂的例子



Page ID	OutLinks
1	2,3,4,5,7
2	1
3	1,2
4	2,3,5
5	1,3,4,6
6	1,5
7	5

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

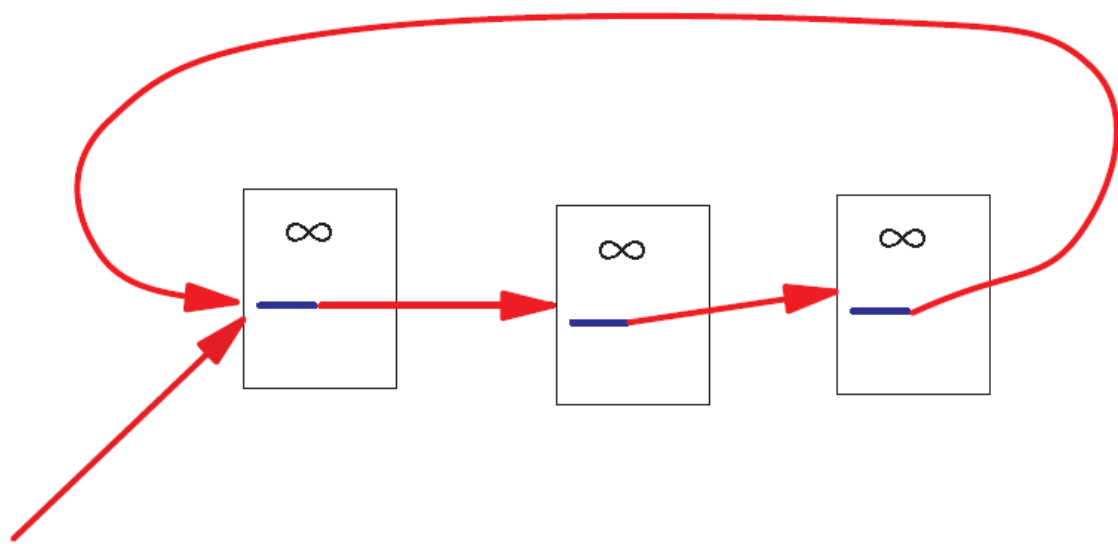
计算过程

$$\text{则归一化后} A = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} = cA^T R, \text{ 令 } c=1, \text{ 解得}$$

$$R = \begin{pmatrix} 0.69946 \\ 0.38286 \\ 0.32396 \\ 0.24297 \\ 0.41231 \\ 0.10308 \\ 0.13989 \end{pmatrix}$$

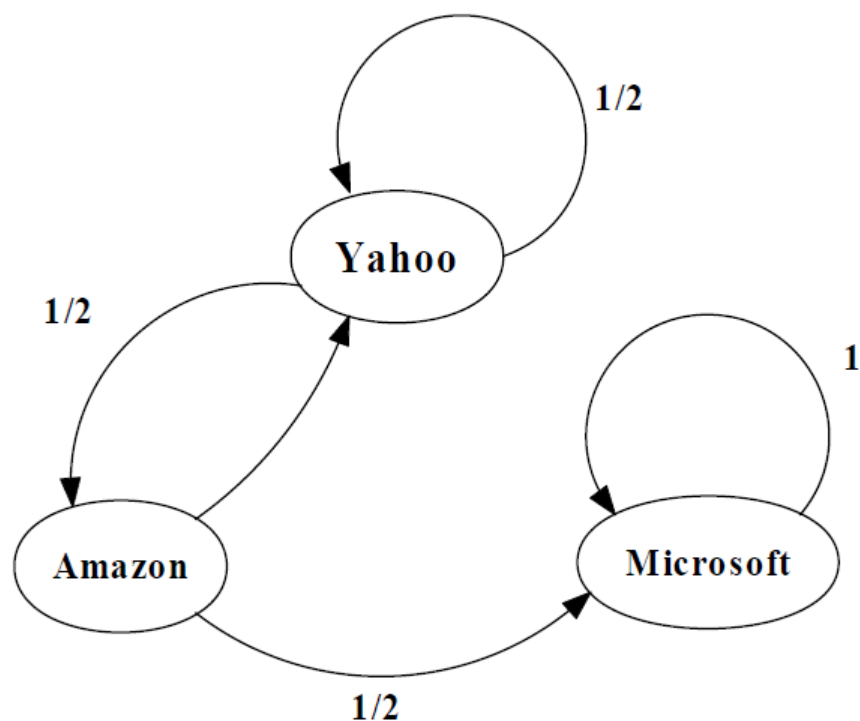
$$\text{Normalized} = \begin{pmatrix} 0.303514 \\ 0.166134 \\ 0.140575 \\ 0.105431 \\ 0.178914 \\ 0.044728 \\ 0.060703 \end{pmatrix}$$

原始PageRank的一个不足



图中存在一个循环通路，每次迭代，该循环通路中的每个节点的PageRank不断增加，但是它们并不指出去，即不将PageRank分配给其他节点！

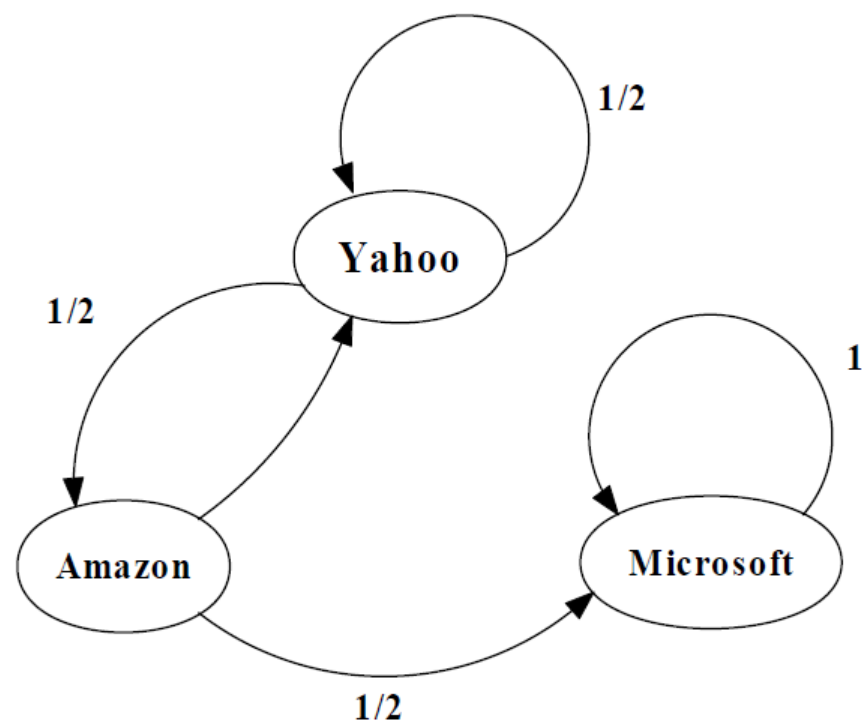
课堂练习



$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/6 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

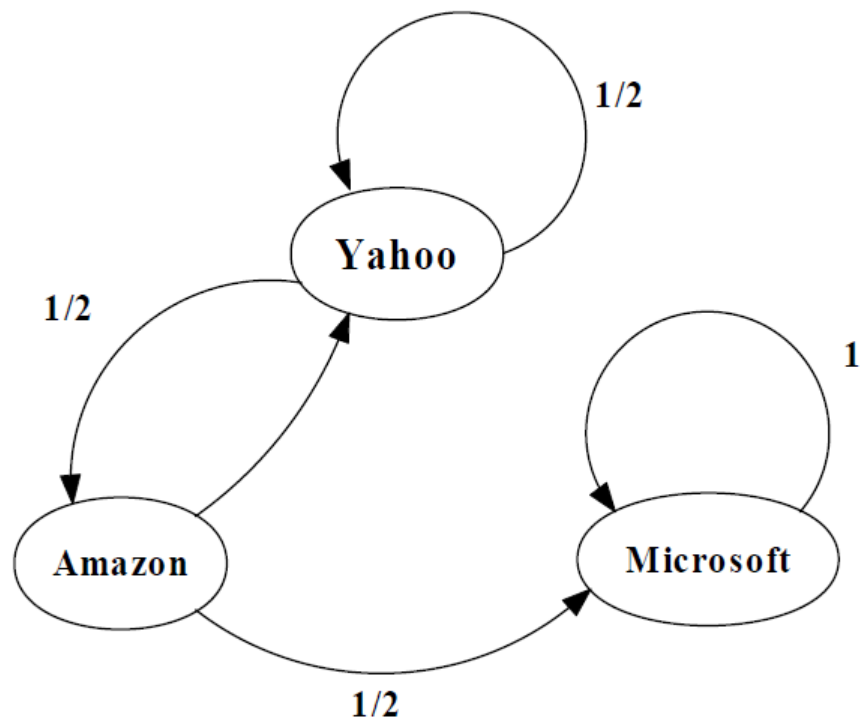
课堂练习



$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/6 \\ 7/12 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/6 \\ 1/2 \end{bmatrix}$$

课堂练习



$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 5/24 \\ 1/8 \\ 2/3 \end{bmatrix} \begin{bmatrix} 1/6 \\ 5/48 \\ 35/48 \end{bmatrix} \dots \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \leftarrow$$

改进的PageRank公式

随机冲浪或随机游走(Random Walk)模型：到达 u 的概率由两部分组成，一部分是直接随机选中的概率 $(1-d)$ 或 $(1-d)/N$ ，另一部分是从指向它的网页顺着链接浏览的概率 d ，则有

$$R(u) = (1-d) + d \sum_{v \in B_u} \frac{R(v)}{N_v} \quad \text{或} \quad R(u) = \frac{(1-d)}{N} + d \sum_{v \in B_u} \frac{R(v)}{N_v}$$

上述两个公式中，后一个公式所有网页PageRank的和为1，前一个公式的PageRank和为 $N(1-d)+d$ 。

可以证明，PageRank是收敛的。计算时，PageRank很难通过解析方式求解，通常通过迭代方式求解。 d 通常取0.85

PageRank面对的Spamming问题

- SEO (Search Engine Optimization): 通过正当或者作弊等手段提高网站的检索排名(包括PageRank)排名。
- 因此, 实际中的PageRank实现必须应对这种作弊, 实际实现复杂得多。实际中往往有多个因子(比如内容相似度)的融合。

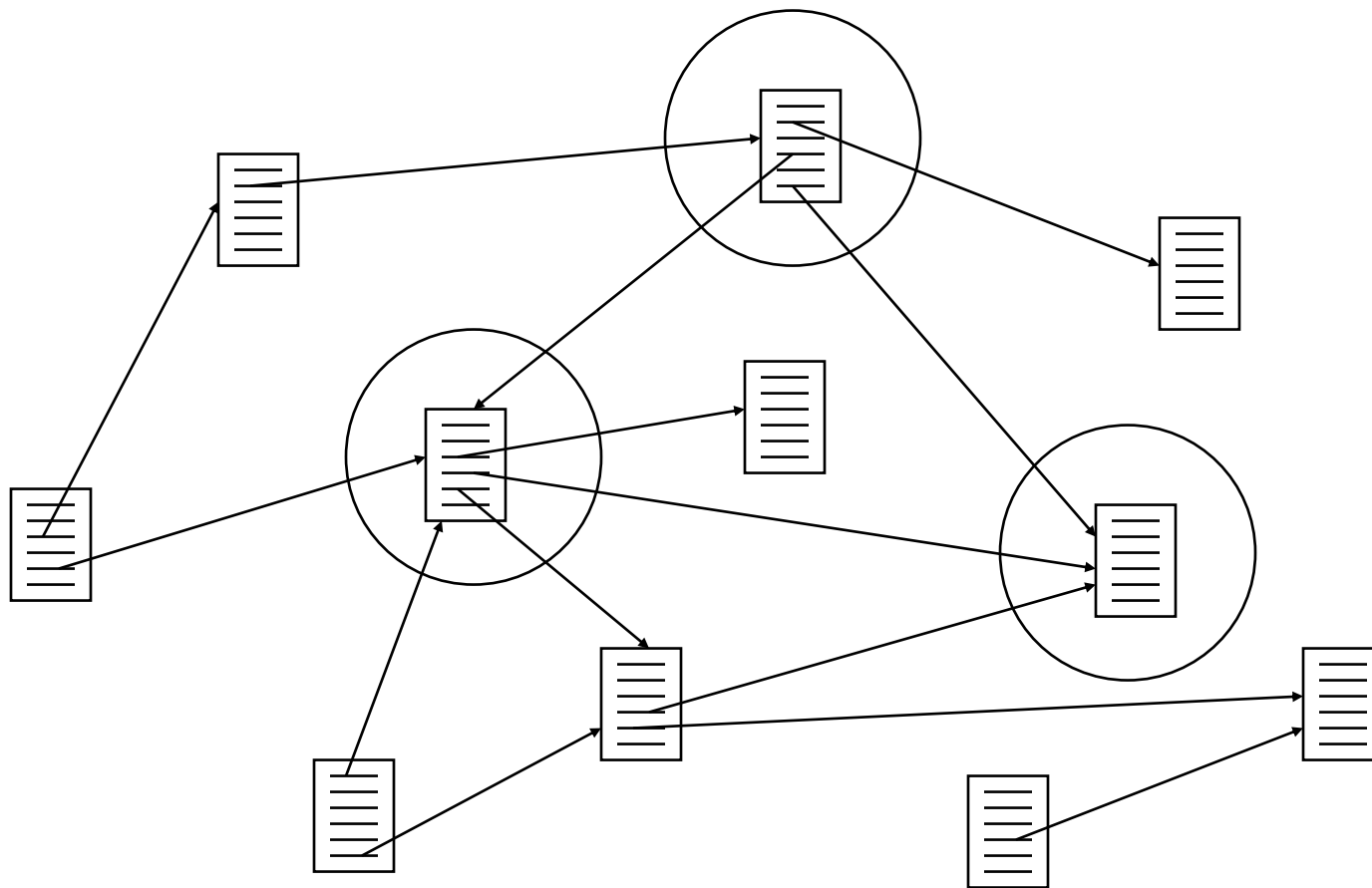
提纲

- ① 上一讲回顾
- ② 锚文本
- ③ 引用分析
- ④ PageRank
- ⑤ HITS: Hub节点&Authority节点
- ⑥ Web作弊与反作弊

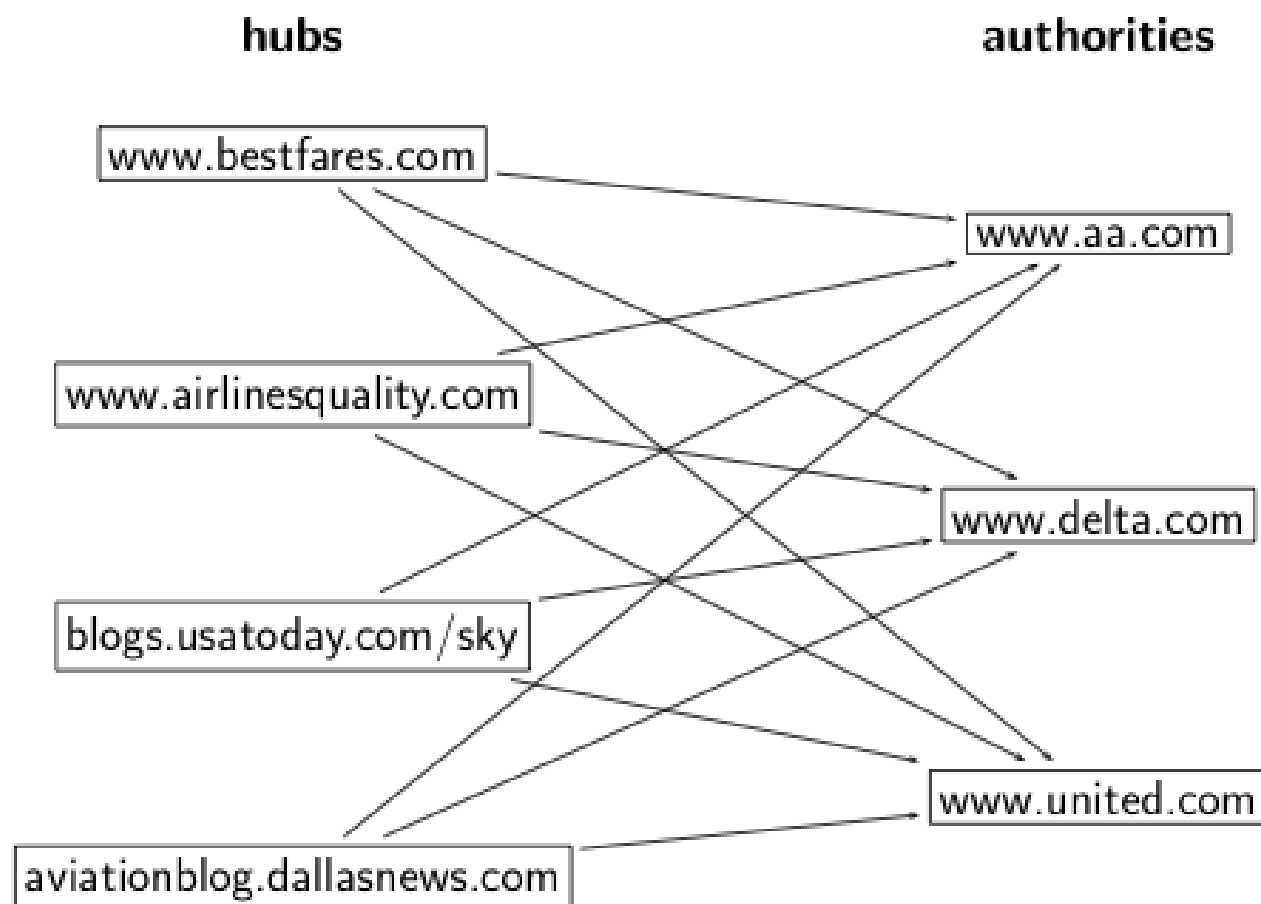
IBM的HITS算法

- HITS(Hyperlink-Induced Topic Search)
- 每个网页计算两个值
 - Authority: 作为权威型网页的权重
 - Hub: 作为目录型或导航型网页的权重

Hub & Authority



例子



查询[Chicago Bulls]的权威网页

0.85	www.nba.com/bulls
0.25	www.essex1.com/people/jmiller/bulls.htm “da Bulls”
0.20	www.nando.net/SportServer/basketball/nba/chi.html “The Chicago Bulls”
0.15	Users.aol.com/rynecub/bulls.htm “The Chicago Bulls Home Page ”
0.13	www.geocities.com/Colosseum/6095 “Chicago Bulls”

(Ben Shaul et al, WWW8)

[Chicago Bulls]的权威网页



KIA LATEST NEWS



WENDELL CARTER JR: THROUGH THE LENS



RYAN ARCDIACONO: THROUGH THE LENS

July 1, 2020

BULLS.COM



BULLS 19-20 SEASON HIGHLIGHTS: RYAN ARCDIACONO

July 1, 2020



BULLS JOIN CHICAGO IN PEACEFUL JUNETEENTH MARCH

June 19, 2020

查询[Chicago Bulls]的导航型网页

1.62 www.geocities.com/Colosseum/1778
 “Unbelieveabulls!!!!”

1.24 www.webring.org/cgi-bin/webring?ring=chbulls
 “Chicago Bulls”


0.74 www.geocities.com/Hollywood/Lot/3330/Bulls.html
 “Chicago Bulls”

0.52 www.nobull.net/web_position/kw-search-15-M2.html
 “Excite Search Results: bulls ”

0.52 www.halcyon.com/wordltd/bball/bulls.html
 “Chicago Bulls Links”

(Ben Shaul et al, WWW8)

[Chicago Bulls]导航型网页的例子



great tickets from nice people

Returning Customer

City Guide | V

Minnesota Timberwolves Tickets
 New Jersey Nets Tickets
 New Orleans Hornets Tickets
 New York Knicks Tickets
 Oklahoma City Thunder Tickets
 Orlando Magic Tickets
 Philadelphia 76ers Tickets
 Phoenix Suns Tickets
 Portland Trail Blazers Tickets
 Sacramento Kings Tickets
 San Antonio Spurs Tickets
 Toronto Raptors Tickets
 Utah Jazz Tickets
 Washington Wizards Tickets
NBA All-Star Weekend
NBA Finals Tickets
NBA Playoffs Tickets
All NBA Tickets

Official Website Links:

Chicago Bulls (official site)
<http://www.nba.com/bulls/>

Fan Club - Fan Site Links:

Chicago Bulls
 Chicago Bulls Fan Site with Bulls Blog, News, Bulls Forum, Wallpapers and all your basic Chicago Bulls essentials!!
<http://www.bullscentral.com>

Chicago Bulls Blog
 The place to be for news and views on the Chicago Bulls and NBA Basketball!
<http://chi-bulls.blogspot.com>

News and Information Links:

Chicago Sun-Times (local newspaper)
<http://www.suntimes.com/sports/basketball/bulls/index.html>

Chicago Tribune (local newspaper)
<http://www.chicagotribune.com/sports/basketball/bulls/>

Wikipedia - Chicago Bulls
 All about the Chicago Bulls from Wikipedia, the free online encyclopedia.
http://en.wikipedia.org/wiki/Chicago_Bulls

Merchandise Links:

Chicago Bulls watches
http://www.sportswatches.com/NBA_watches/Chicago-Bulls-watches.html

Event Selections

Sporting Events

MLB Baseball Tickets

NFL Football Tickets

NBA Basketball Tickets

NHL Hockey Tickets

NASCAR Racing Tickets

PGA Golf Tickets

Tennis Tickets

NCAA Football Tickets

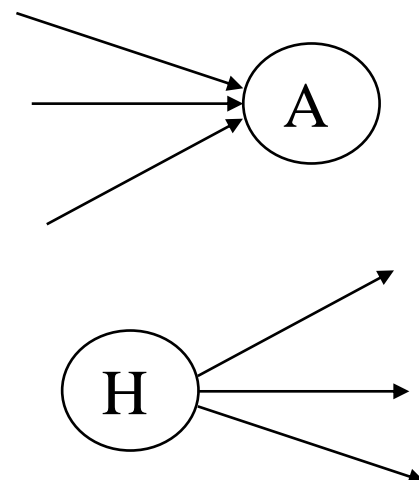
计算方法

$$A(p) = \sum H(q_i)$$

(其中 q_i 是所有链接到 p 的页面)

$$H(p) = \sum A(r_i)$$

(其中 r_i 是所有页面 p 链接到的页面)



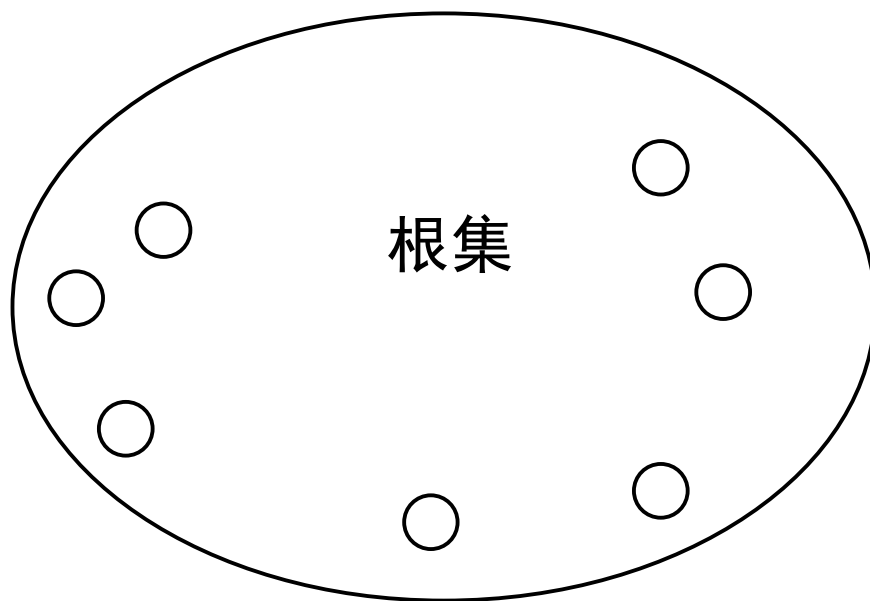
- (1) 一个网页被越重要的导航型网页指向越多，那么它的Authority越大；
- (2) 一个网页指向的高重要度权威型网页越多，那么它的Hub越大。

HITS算法也是收敛的，也可以通过迭代的方式计算。

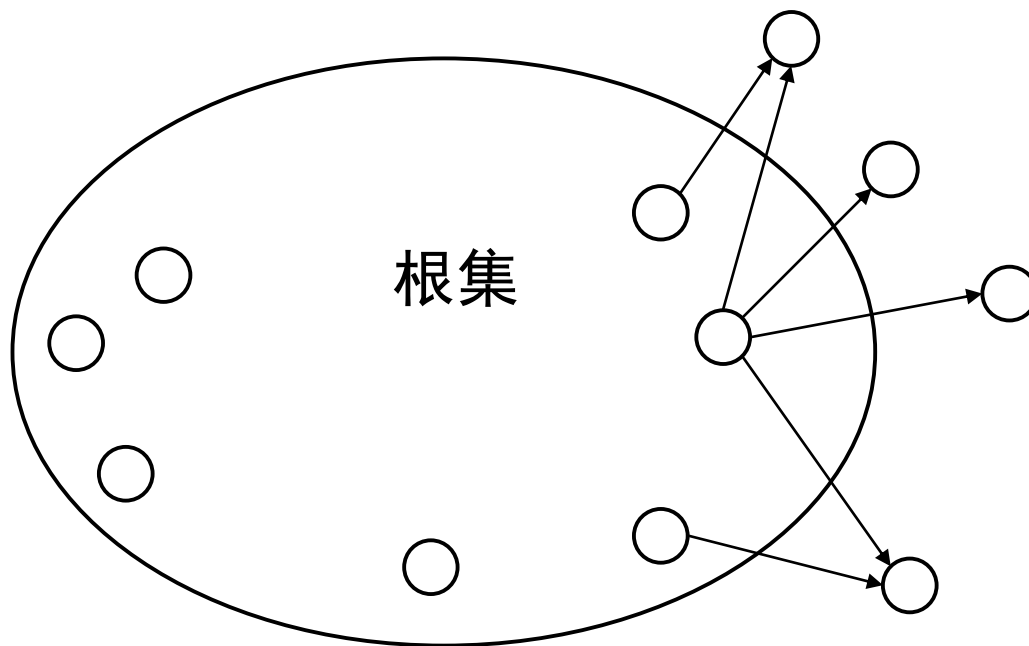
HITS算法的实际计算过程

- 首先进行Web搜索；
- 搜索的结果称为根集(**root set**)；
- 将所有链向种子集合和种子集合链出的网页加入到种子集合；
- 新的更大的集合称为基本集(**base set**)；
- 最后，在基本集上计算每个网页的hub值和authority值 (该基本集可以看成一个小Web图)。

根集和基本集 (1)

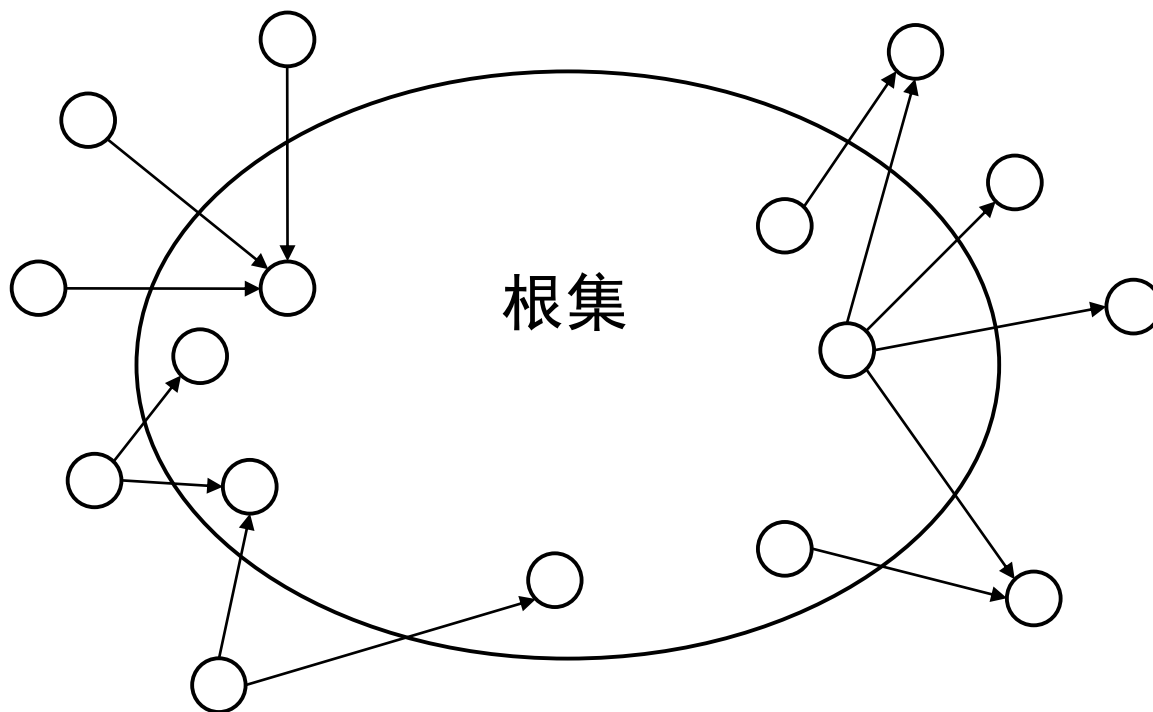


根集和基本集 (2)



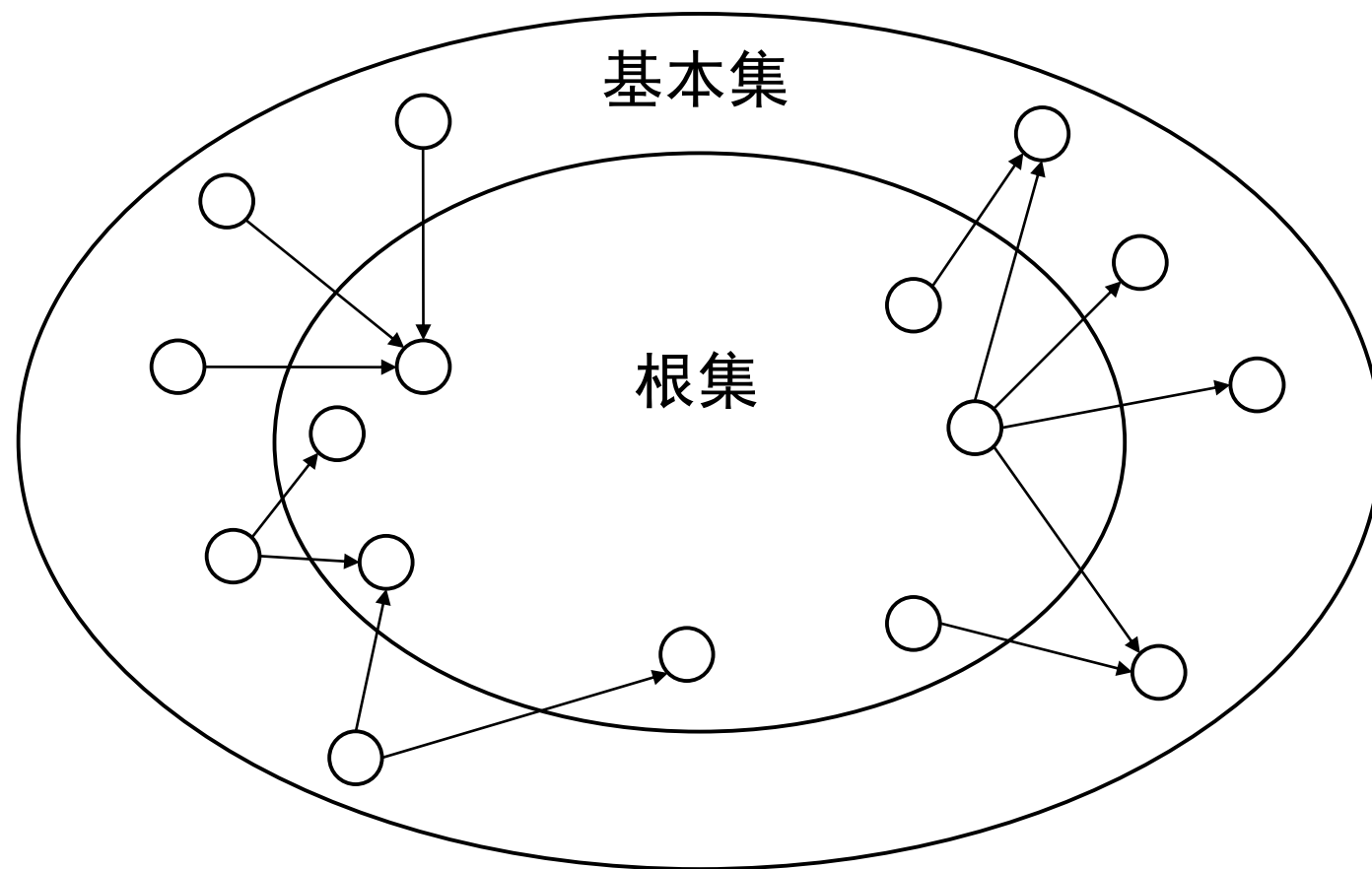
根集中节点链向的网页节点

根集和基本集 (3)



指向根集节点的那些节点

根集和基本集 (4)



基本集

根集和基本集 (5)

- 根集往往包含200-1000个节点
- 基本集可以达到5000个节点

输出排序结果

- 对基本集的每个页面初始化Hub和Authority值
- 迭代计算

$$A(p) = \sum H(q_i)$$

(其中 q_i 是所有链接到 p 的页面)

$$H(p) = \sum A(r_i)$$

(其中 r_i 是所有页面 p 链接到的页面)

- 按照Authority排序，输出结果页面

PageRank vs. HITS

- 网页的PageRank与查询主题无关，可以事先算好，因此适合于大型搜索引擎的应用。
- HITS算法的计算与查询主题相关，检索之后再行计算，因此，不适合于大型搜索引擎。

提纲

- ① 上一讲回顾
- ② 锚文本
- ③ 引用分析
- ④ PageRank
- ⑤ HITS: Hub节点&Authority节点
- ⑥ Web作弊与反作弊

Web作弊与反作弊

- Web作弊(Web Spam)是指采取一些迷惑、欺骗搜索引擎的手段,使某些Web页面在检索结果中的排名高于实际应得的排名的行为。
- 有人估计WEB中有10%~15%的作弊内容。
- 搜索引擎优化(Search Engine Optimizing) 行业的诞生
 - 正当手段: 对网页进行优化(标题、布局)
 - 作弊手段: 欺骗搜索引擎的手段
- 反作弊(anti-spam)是搜索引擎公司的一项重要任务
- 学术界2005年开始就有**AIRWeb: Adversarial Information Retrieval**的Workshop (<http://airweb.cse.lehigh.edu/>), 其中最重要的一个任务就是Web反作弊

Web作弊的危害

- 降低用户体验的满意程度，降低用户对搜索引擎的信任
- 搜索引擎公司会因用户的满意度降低而使其商业价值受到损害
- 作弊或者垃圾页面也消耗了大量时间和空间

Web作弊的方法

- 一、各种提高排名的技术
- 二、各种隐蔽技术，用于使第一类技术的使用不被发现

Web作弊的方法

- 一、各种提高排名的技术
- 二、各种隐蔽技术，用于使第一类技术的使用不被发现

利用关键词提高排名

- 内容匹配仍是大部分搜索引擎排名算法的重要组成部分。 $TF*IDF$ 仍是基本思想。
 - 作弊方法一：在网页(标题或者元信息域)中加入大量关键词，使得查询和目标网页匹配上的关键词个数增多，从而提高排名。
 - 作弊方法二：在网页中(标题或者元信息域)加入大量与某些查询相关的重复“关键词”，使得网页排名上升。

利用链接提高排名(1)

- 根据搜索引擎所采用的链接分析算法，构造具有某些链接结构的作弊网站，迷惑搜索引擎，提高排名。
 - 出链接作弊(破坏HITS算法)：在网页上加入大量的出链接指向著名站点，提高本网页的Hub值。如采用目录克隆(directory cloning)方法直接拷贝如DMOZ Open Directory Project上的全部或者部分目录。

利用链接提高排名(2)

- 入链接作弊：
 - 蜜罐诱饵(honey pot): 一组提供有用资源的网页，包含了许多指向目标作弊网页的链接，它们像蜜罐一样引诱其他页面指向它们，从而间接提高的目标作弊网页的排名。
 - 渗入Web目录: 作弊者提交网页到一些著名的WEB目录，编辑者可能没有严格审查，而上述提交网页中含有指向目标作弊网页的链接，由于WEB分类目录通常具有很高的PageRank和Hub，所以目标网页的排名也能提高。

利用链接提高排名(3)

- 入链接作弊：
 - 张贴法：在Blog、BBS、留言板或Wiki上张贴链接指向目标作弊网页。
 - 链接交换：作弊者联合作案，作弊网站互相链接。
 - 购买过期域名：过期域名指向作弊链接。
 - 构造链接农场(link farm)：操作大量网站，构造能够提高PageRank的任意网站。现在投资已经很少。
 - 泛域名作弊(二级域名作弊)：最低一级域名是随机生成的，这些域名代表的页面要么互相链接，要么指向同一作弊网页，要么重定向到一个作弊页面。如：中文互联网上的881166.com等

Web作弊的方法

- 一、各种提高排名的技术
- 二、各种隐蔽技术，用于使第一类技术的使用不被发现

内容隐藏

- 浏览器显示页面时，用户看不到作弊的关键词或者链接。
 - 通过颜色配置使得关键词和背景颜色一样
 - 作弊链接不加上文字后不可见
 - 将作弊链接加在非常小的透明或者和背景一样颜色的图片上
 - 使用脚本技术来隐藏网页中的一些可见成分，如将HTML风格中的Visible属性设为false

覆盖(Cloaking)

- 通过识别网站的访问者是否是搜索引擎的爬虫，提供不同的URL。作弊网页被提供给搜索引擎用于建立索引。而用户访问时显示为另一个正常页面。

重定向

- 网页在被浏览器载入时自动重定向到另一个URL。这样的网页仍然可以被搜索引擎抓取，但是用户却看不到它。这样作弊网页被抓取，而用户看到的却是重定向后的目标文件。简单的方法就是在网页头部meta中的refresh时间设为0。更高级的方法采用一些脚本技术。

一些反作弊技术

- TrustRank: 为网页建立信任值。
- 改进的PageRank方法: 识别链接农场作弊方法。
- 语言模型方法: 根据不同类型网页内容的语言模型的差别进行判别
- 网页版本差异判断方法: 采用浏览器方法和爬虫方法同时抓取。
- 目前这些方法的精度仍然不是很高, 因受各种限制, 很多方法在搜索引擎中并没大量使用。

WEB作弊和反作弊的长期斗争

- 道高一尺魔高一丈
- 魔高一尺道高一丈

本讲内容

- 锚文本: Web上的链接相关信息为什么对IR有用?
- 引用分析(Citation analysis): PageRank及其他基于链接排序方法的数学基础
- PageRank : 一个著名的基于链接分析的排序算法(Google)
- HITS : 另一个著名的基于链接分析的排序算法(IBM)
- Web作弊与反作弊

参考资料

- 《信息检索导论》 第21章
- American Mathematical Society article on PageRank (popular science style)
- Jon Kleinberg's home page (main person behind HITS)
- A Google bomb and its defusing
- Google's official description of PageRank: *PageRank reflects our view of the importance of web pages by considering more than 500 million variables and 2 billion terms. Pages that believe are important pages receive a higher PageRank and are more likely to appear at the top of the search results.*