WDS

DataScience@Web

GRAPH | NETWORK ANALYSIS

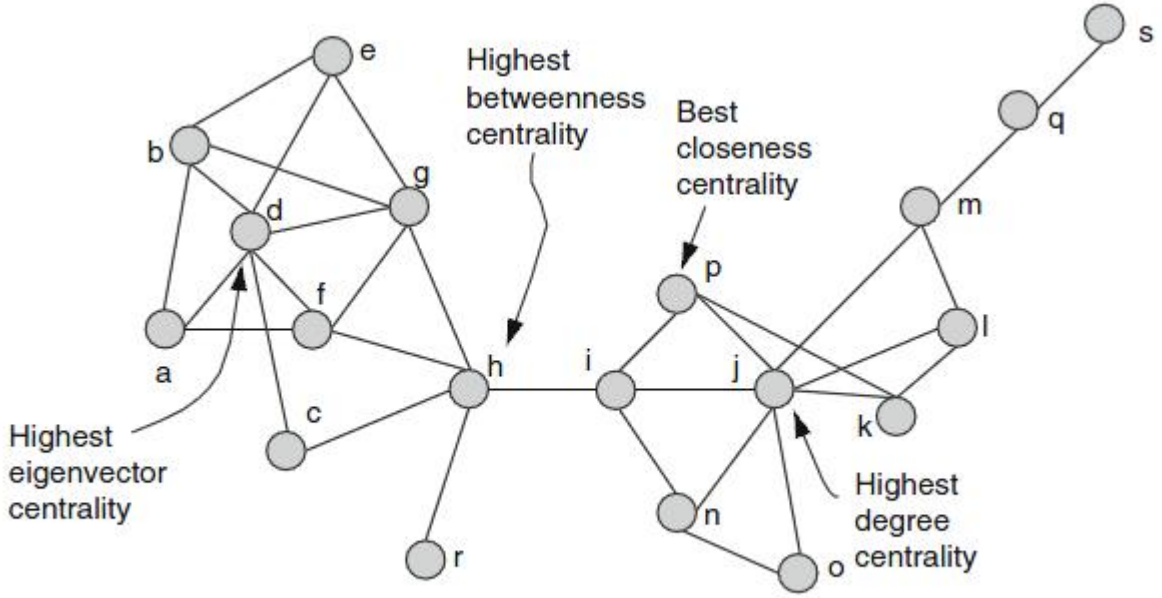
Peng Wang

School of Computer Science and Engineering, Southeast University

MEASURES & METRICS

中心度

Centrality ? finding the key players



Ortizarroyo D. Discovering Sets of Key Players in Social Networks[J]. Computational Social Network Analysis Computer Communications & Networks, 2010:27-47.

特征向量中心度

Eigenvector Centrality

- an extension of degree centrality
- a player with many important friends is important

x_i的表示节点i的中心度，是邻居节点的中心度之和

$$x_i' = \sum_j A_{ij} x_j$$

i节点重要有两个条件：
- i的邻居要多
- i的邻居要重要

第t次迭代后的中心度向量

$$x(t) = A^t x(0)$$

x是元素为x_i的向量

$$Ax = k_1 x$$

k_1是邻接矩阵A的最大特征值(主特征值)

x是邻接矩阵A的主特征向量

Eigenvector Centrality

- For directed graph, such as WWW

$$Ax = k_1 x$$

k_1是邻接矩阵A的最大特征值(主特征值)

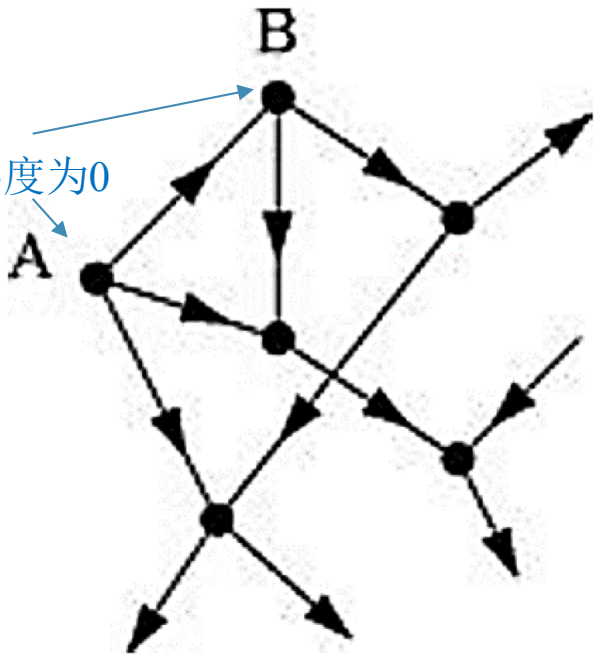
x是邻接矩阵A的右主特征向量

为什么是右主特征向量，而不是左主特征向量

Eigenvector Centrality - problems

- the EC of nodes with no in-links are 0 (like node A);
- the EC of nodes with in-links from A are 0 (like node B)
- Only nodes in a SCC or its out-component are not 0;
- for DAG, each node is 0…

外向分支



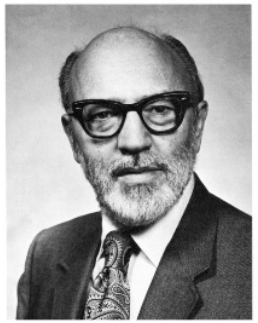
Katz Centrality

为每个节点加上一些“免费”的中心度，β常设为1

$$x_i' = \alpha \sum_j A_{ij} x_j + \beta$$

自由参数，调节特征向量与常数项的平衡

- $\alpha \to 0$: $x_i \to \beta$
- $\alpha > \frac{1}{k_1}$: x_i does not converge 不收敛
- $\alpha \in \left(0, \frac{1}{k_1}\right)$



Leo Katz

Katz Centrality

What is the problem of Katz Centrality?

十大经典数据挖掘算法之一

PageRank [1] Centrality


Larry Page


Sergey Brin

The PageRank Citation Ranking: Bringing Order to the Web

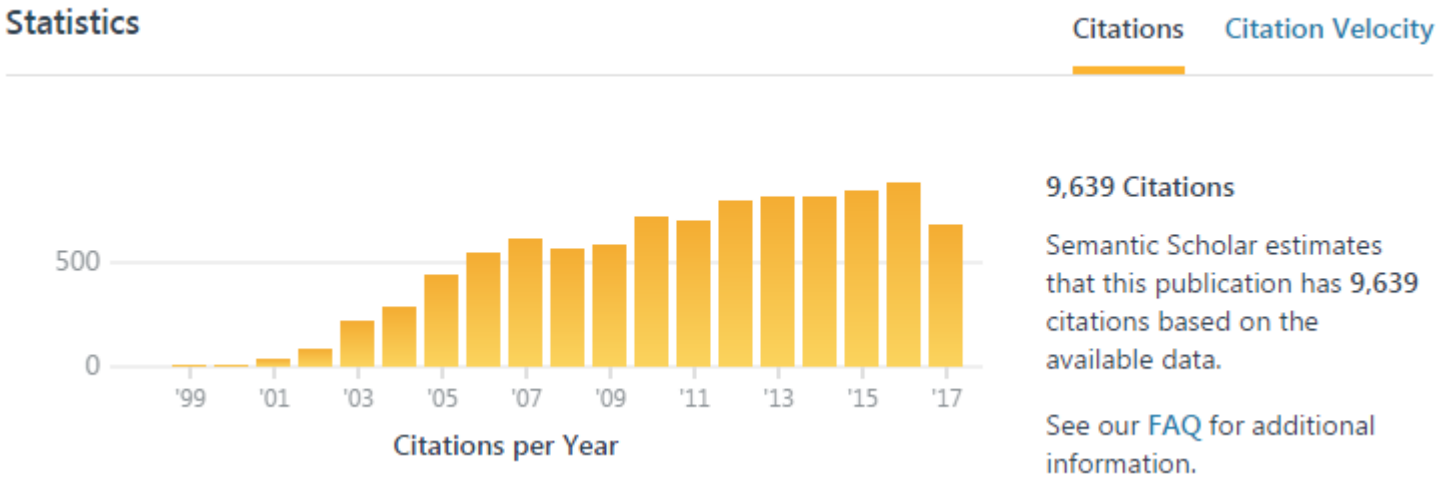
January 29, 1998

Abstract

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, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

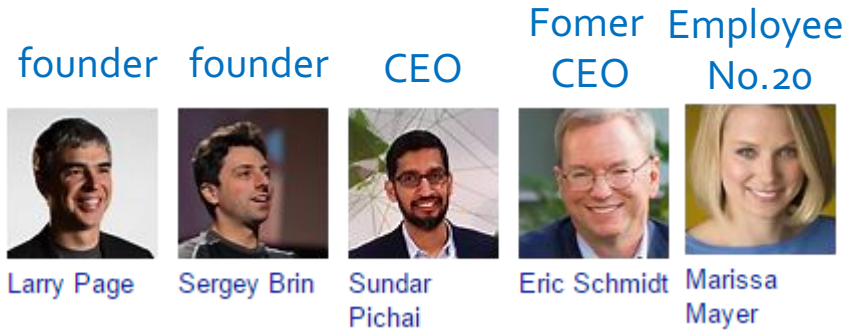
PageRank


Statistics
Citations
Citation Velocity
500
0
'99
'01
'03
'05
'07
'09
'11
'13
'15
'17
Citations per Year
9,639 Citations
Semantic Scholar estimates that this publication has 9,639 citations based on the available data.
See our FAQ for additional information.

The Rise Of Google™

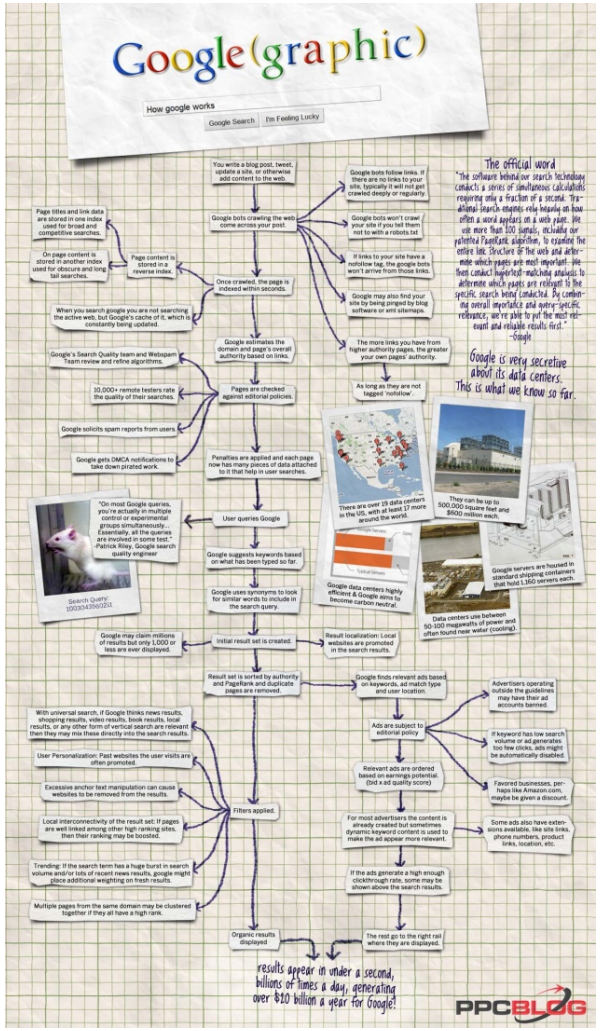
Google History

- Larry Page and Sergey Brin from Stanford
- From “What Box”to 10^{100}
- Refused by Altavista (say no for $1M), InfoSeek (“滚!”)
- 100,000 dollars from Sun
- Beat Yahoo, Excite.com, InfoSeek, Lycos



Google Facts

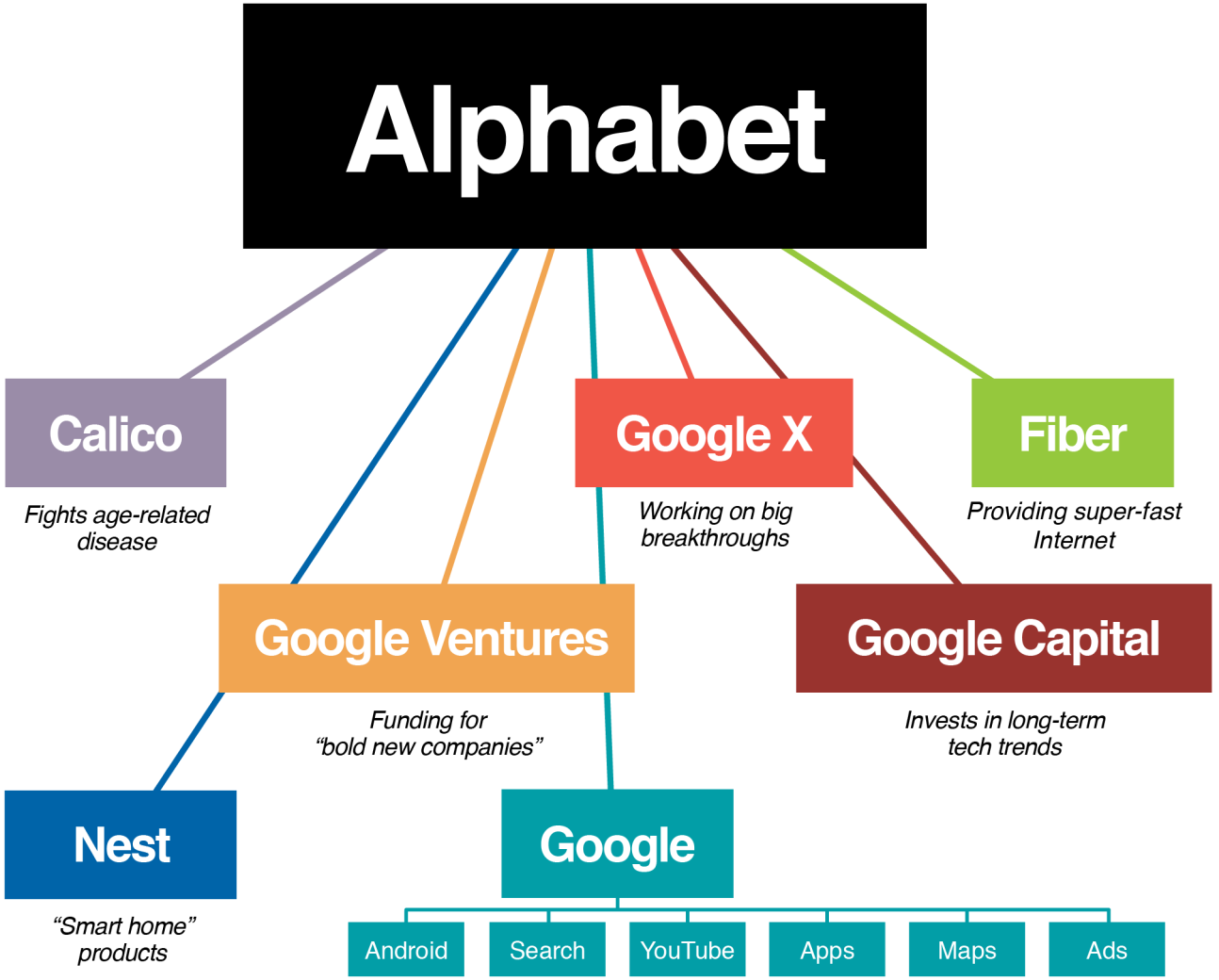
- Google receives over 63,000 searches per second on any given day.
- Google takes over 200 factors into account before delivering you the best results to any query in a fraction of a second.
- Google has a market value of $739 billion.
- Google user statistics: 2 billion users on Android and 500M on Google Photos.
- Google search statistics: Google's search engine performed 1.2 trillion searches in 2016.
- Google has 90.46% of the search engine market share worldwide.
- Google office facts: the company has 88,110 full-time employees.
- 30.9% of the global Google employees are female.
- An average person conducts 3–4 searches every single day.
- There are more than 70,000 Google queries per second.



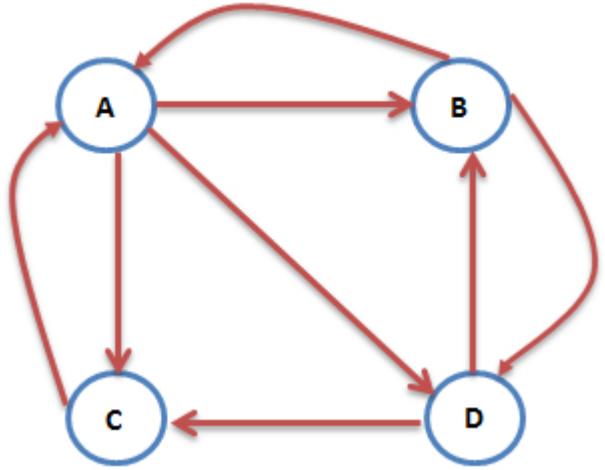
Understanding How Google Works

English: http://ppcblog.com/how-google-works/

Chinese:
http://www.kuqin.com/searchengine/20130320/334054.html


Alphabet
Calico
Fights age-related disease
Google X
Working on big breakthroughs
Fiber
Providing super-fast Internet
Google Ventures
Funding for "bold new companies"
Google Capital
Invests in long-term tech trends
Nest
"Smart home" products
Google
Android
Search
YouTube
Apps
Maps
Ads

PageRank Basic Idea 1:



PageRank算法从原理到实现
https://www.cnblogs.com/rubinorth/p/5799848.html

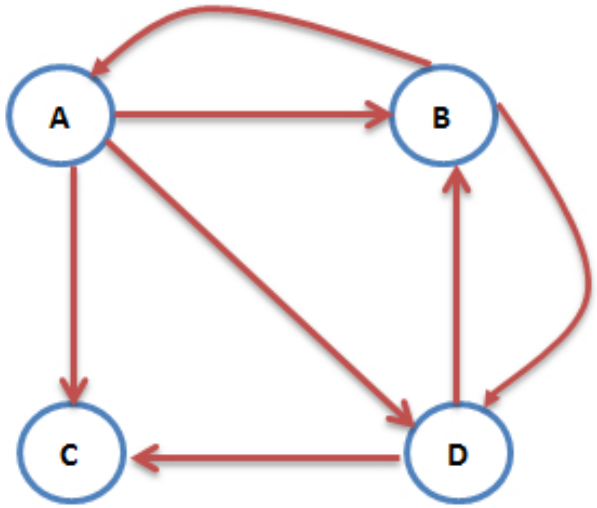
每一位指向A的邻居为A投票

$$PR(A) = \frac{PR(B)}{k_{out}(B)} + \frac{PR(C)}{k_{out}(C)}$$

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

票面价值按出度等分

PageRank Basic Idea 2:



没有出度

$$PR(A) = \frac{PR(B)}{k_{out}(B)} + \frac{PR(C)}{N}$$

$$= \frac{PR(B)}{2} + \frac{PR(C)}{4}$$

对“自私”的节点进行自动投票

? 为何需要为自私节点进行自动投票？
为了保证马尔科夫过程的收敛性：“马尔科夫收敛的条件之一是*能从任意一个状态能够变到任意其他一个状态。*”

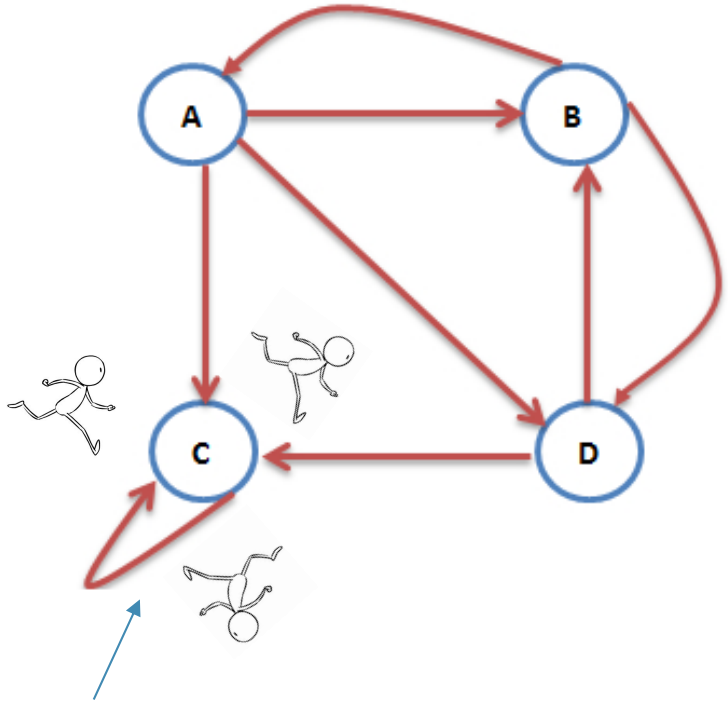
马尔科夫收敛定理

Markov Convergence Theorem

如果满足下述四个条件，一个马尔科夫过程将收敛到一个均衡状态，且此均衡唯一：

1. 可能的状态数量是有限的。
2. 转移概率固定不变。
3. 从任意一个状态能够变到任意其他一个状态。有可能不是从状态A直接变到状态C，而是先变到状态B再变到C，但只要有路径从状态A变成状态C就行。
4. 过程不是简单循环。比如不能是从全A变到全B，然后又自动从全B变到全A。

PageRank Basic Idea 3：



存在自指的节点，一个浏览者(random surfer)会陷入无限循环

Surfer以α的概率顺着超链接浏览网页

$$PR(A) = \alpha \frac{PR(B)}{k_{out}(B)} + \frac{1-\alpha}{N}$$

$$= \alpha \frac{PR(B)}{2} + \frac{1-\alpha}{4}$$

Surfer以$1-\alpha$的概率随机跳转至任意网页

PageRank Basic Idea 3：

阻尼系数，damping factor

$$PR(i) = \alpha \sum_{j \to i} \frac{PR(j)}{k_{out}(j)} + \frac{(1-\alpha)}{N}$$

一般α设为0.85

Calculating PageRank with Matrix

$$PR(i) = \alpha \sum_{j \to i} \frac{PR(j)}{k_{out}(j)} + \frac{(1-\alpha)}{N}$$

网页PR值向量 → $\mathbf{PR} = \alpha \mathbf{A} \mathbf{D}^{-1} \mathbf{PR} + \frac{(1-\alpha)}{N} \mathbf{1}$ ← 全1向量

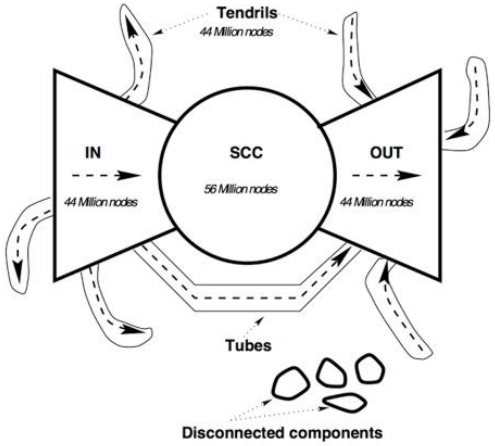
作为因子可以省略 →

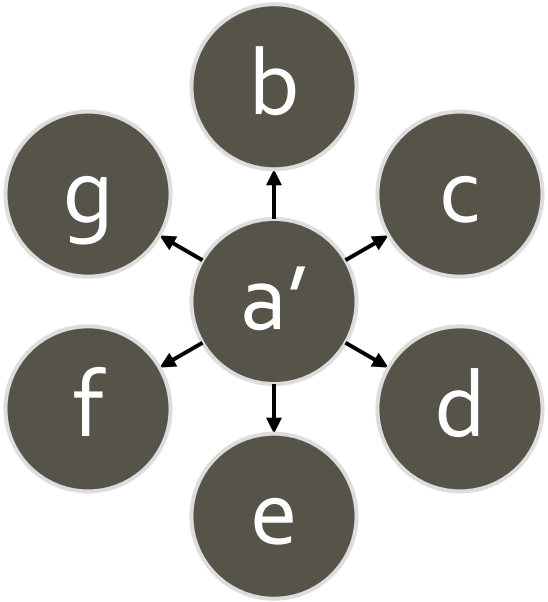
$$= \frac{(1-\alpha)}{N} (\mathbf{I} - \alpha \mathbf{A} \mathbf{D}^{-1})^{-1} \mathbf{1}$$

$$= \mathbf{D} (\mathbf{D} - \alpha \mathbf{A})^{-1} \mathbf{1}$$

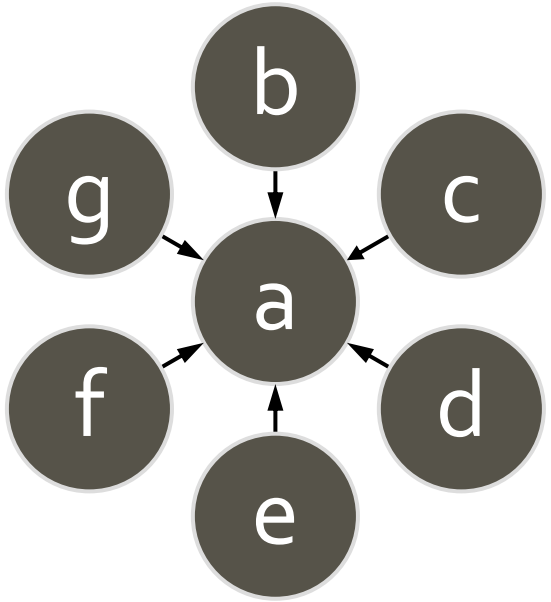
对角矩阵，元素为$max(k_{out}(i), 1)$

HITS [2]





Hub Nodes
枢纽节点



Authoritative Nodes
权威节点

Jon kleinberg

HITS: Basic Idea

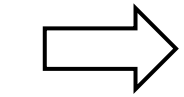
权威度 $\longrightarrow$ $authority(i) = \sum_{j \to i} hubness(j)$

枢纽度 $\longrightarrow$ $hubness(i) = \sum_{i \to j} authority(j)$

Calculating HITS with Matrix

$$a(i) = \alpha \sum_{j} A_{ij} h(j)$$

$$h(i) = \beta \sum_{i \to j} A_{ij} a(j)$$



权威度矩阵

枢纽度矩阵

常数

邻接矩阵

$$\boldsymbol{\mathcal{A}} = \alpha \mathbf{A} \boldsymbol{\mathcal{H}}$$

$$\boldsymbol{\mathcal{H}} = \beta \mathbf{A}^{\mathrm{T}} \boldsymbol{\mathcal{A}}$$

$$\boldsymbol{\lambda} = (\boldsymbol{\alpha\beta})^{-1}$$

$$\mathbf{A}\mathbf{A}^{\mathrm{T}} \boldsymbol{\mathcal{A}} = \boldsymbol{\lambda} \boldsymbol{\mathcal{A}}$$

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} \boldsymbol{\mathcal{H}} = \boldsymbol{\lambda} \boldsymbol{\mathcal{H}}$$

$$\boldsymbol{\mathcal{H}} = \mathbf{A}^{\mathrm{T}} \boldsymbol{\mathcal{A}}$$

Calculating HITS with Matrix

$\mathcal{A}$由$\mathbf{A}\mathbf{A}^{\mathrm{T}}$特征向量决定

$$\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathcal{A} = \lambda\mathcal{A}$$

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathcal{H} = \lambda\mathcal{H}$$

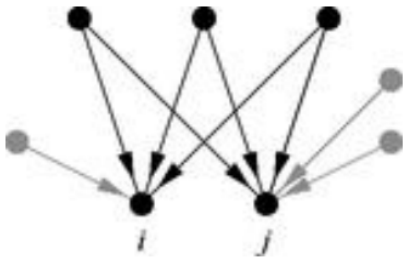
$\mathcal{H}$由$\mathbf{A}^{\mathrm{T}}\mathbf{A}$特征向量决定

$\mathbf{A}\mathbf{A}^{\mathrm{T}}$和$\mathbf{A}^{\mathrm{T}}\mathbf{A}$分别是什么矩阵？

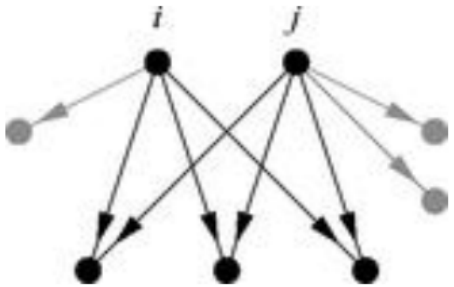
Calculating HITS with Matrix

- $\mathbf{AA}^{\mathbf{T}}$: Cocitation Matrix 共引矩阵
- $\mathbf{A}^{\mathbf{T}}\mathbf{A}$: Bibliographic Coupling Matrix 文献耦合矩阵
- Authority: Eigenvector Centrality of Cocitation Network
- Hubness: Eigenvector Centrality of Bibliographic Coupling Network

共引矩阵



文献耦合矩阵



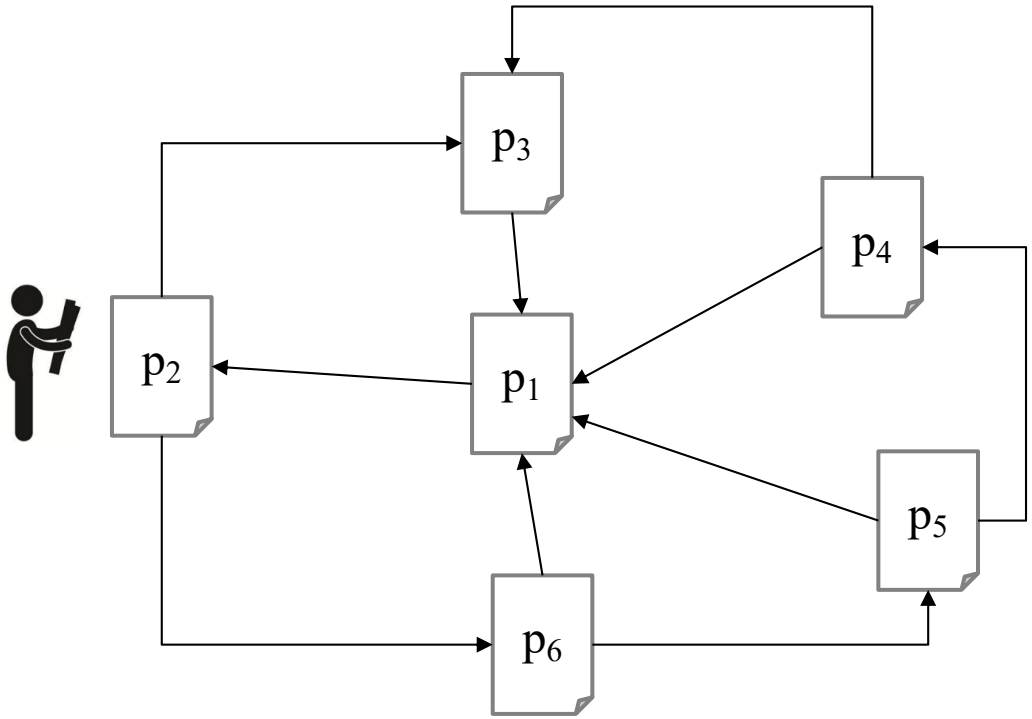
PageRank vs. HITS

PageRank从原理到实现: http://blog.csdn.net/rubinorth/article/details/52215036
HITS 从原理到实现: http://blog.csdn.net/rubinorth/article/details/52231620

- PageRank:
  - query-independent 与查询无关，用于全局网页的静态排序，可离线完成计算
  - topic-irrelevant 与主题无关，没有区分页面内的导航链接、广告链接和功能链接等，容易对广告页面有过高评价
  - favoring old pages 旧页面等级往往比新页面高，因为新页面的链接较少
- HITS:
  - query-dependent 通常用于用户查询结果的排序，但无法用于大型搜索引擎的实时排序，并不实用
  - topic drift and topic Distillation 主题漂移：扩展集合里可能包含主题无关网页，因此需要主题精化
  - sensitive to spam links 思考：如何通过造假来显著提高某个网页的authority?

1. 用户查询生成根集合
2. 利用链接关系生成扩展集合

A Probabilistic Framework of WPSS [3]
A Single-Surfer Model



1. Jump ($\mathrm{P}(J|i)$)
2. Follow ($\mathrm{P}(L|i)$)
3. Back ($\mathrm{P}(B|i)$)
4. Stay ($\mathrm{P}(S|i)$)

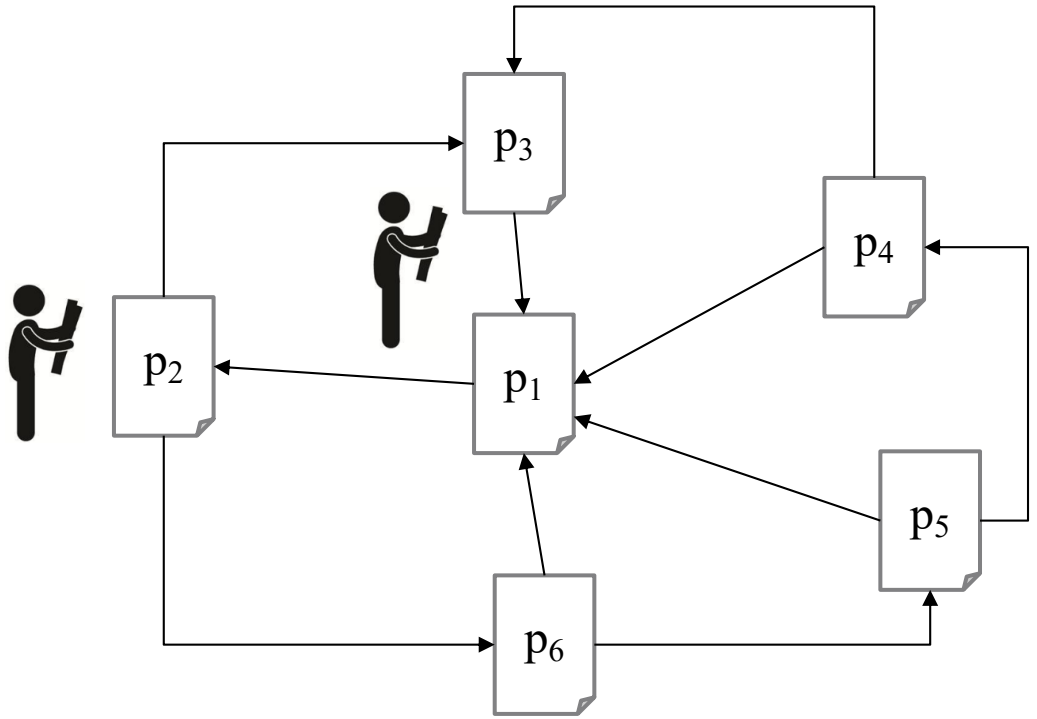
$$\sum \mathrm{P}(J, L, B, S|i) = 1$$

$$\sum_{j \epsilon \mathbf{G}} \mathrm{P}(j|J, i) = 1 \quad \sum_{i \to j} \mathrm{P}(j|F, i) = 1 \quad \sum_{j \to i} \mathrm{P}(j|B, i) = 1$$

The centrality of i is a stationary distribution of the Markov chain

$$c(i) = \mathrm{P}^{(t)}(i)$$

A Multi-Surfer Model



假设有k个surfer同时浏览WWW，则t时刻停留在i上的概率分别为：

$$P_{s1}^{(t)}(i), P_{s2}^{(t)}(i), \ldots, P_{sk}^{(t)}(i)$$

假设每个surfer会以一定的概率听取其他surfer的建议跳转至其他surfer正在浏览的网页

$$P(s_j|s_i)$$

对任意一个s_i，该surfer对所有其他surfer的信任度之和为1

$$\sum_{j=1}^{k} P(s_j|s_i) = 1$$

Horizontal: 仅考虑链接关系

Vertical: 还考虑文本相关性

U(uniform): 不同节点采用统一的转移概率

F(focused): 不同节点根据文本相关性采用不同的转移概率

Main Features of the Proposed Ranking Functions

	Hor./Vert.	Multi/Single	Jump	Back	Forw.
PageRank	H	S	U	-	U
HITS	H	M	-	U	U
Focused PageRank	V	S	U	-	F
Double Focused PR	V	S	F	-	F
Focused HITS	V	M	-	F	F
PageRank-HITS	H	M	U	U	U
Multi-topic	V	M	U/F	-	F

Reference

[1] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. World Wide Web Internet Web Inf. Syst. 54, 1–17 (1998).

[2] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM. 46, 604–632 (1999).

[3] Diligenti M, Gori M, Maggini M: A unified probabilistic framework for Web page scoring systems. IEEE Transactions on Knowledge & Data Engineering. 16(1):4-16 (2004)

接近度中心度

Closeness Centrality

$$l_i = \frac{1}{n}\sum_j d_{ij}$$

该点到所有其他点的平均距离

两点间最短路径长度

图中节点个数

有时会将n改为$(n-1)$，j要求不等于i

$$C_i = \frac{1}{l_i}$$

Problems of Closeness Centrality

最大值与最小值之间跨度太小

- span a small dynamic range from largest to smallest
- in a typical network, $C_{max}/C_{min} \leq 5$
- prone to change with a subtle change of the network structure
- has some problems dealing with disconnected components

In IMDB:

Christopher Lee in the Lord of Rings

C_{max}=0.4143

500,000 movie stars

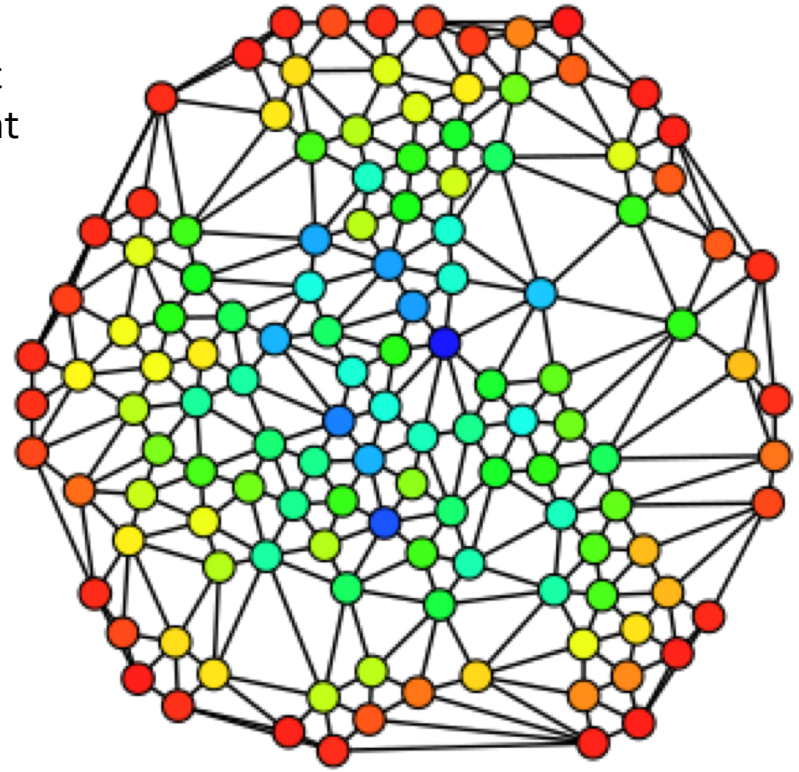
Leia Zanganeh

C_{min}=0.1154

介数中心度

Betweenness Centrality

Red: least important
Blue: most important



Considering any pair of nodes are sending messages to each other, after a long time, how many messages have passed through each node? Given that:

1) Nodes are communicating equally;
2) Messages are send through shortest path;
3) A random path is selected for each message if there are multiple choices.

[qián]
中间人、掮客

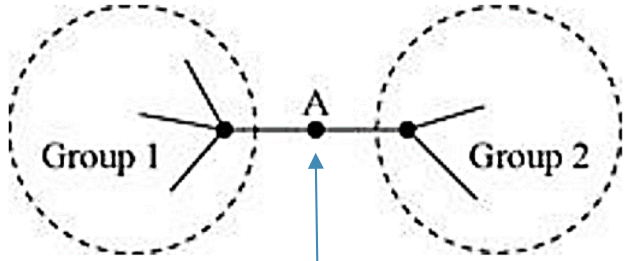
Betweenness: Finding the Broker

s,t间经过i的最短路径数

$$B_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$$

s,t间的总最短路径数

s,t为图中任意两个节点



割集中的节点，通常具有很低的degree，但很高的betweenness

Properties of Betweenness

最大值与最小值之间跨度很大

- span a small dynamic range from largest to smallest

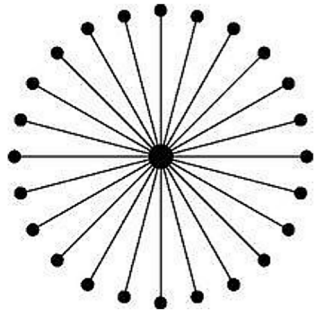
星形图

- considering a star graph

星形图中心点的介数中心度

- $B_{max}/B_{min} = \frac{n^2-n+1}{2n-1} \approx \frac{n}{2}$

星形图边缘点的介数中心度



- stable with the change of network structure

Example of Betweenness

In IMDB:

1st: 7.47×10^8

Fernando Rey

2nd: 6.49×10^8

Christopher Lee

...

least: 8.91×10^5

0-1 Normalization of Betweenness

$$B_i = \frac{1}{n^2} \sum_{st} \frac{n_{st}^i}{g_{st}}$$

$$B_i \in [0,1]$$

流介数

Flow Betweenness

What if the flow of information is not always go through the shortest path?

For Example,

A confirmation of a break-up with you is not directly from your girl,

but indirectly from her roommate…

Calculating Flow Betweenness

$$B_i = \frac{1}{n^2} \sum_{st} \frac{n_{st}^i}{g_{st}}$$

s,t之间的独立路径数

Random-walk Betweenness

- A variant of betweenness
- the flow is a random walk from *s* to *t*

$$B_i = \sum_{s,t} n_{st}^i$$

从*s*到*t*多次游走经过*i*的平均值

A Faster Algorithm for Betweenness Centrality

- $O(mn)$ for unweighted graph
- $O(mn + n\log n)$ for weighted graph
- close to $O(n^2)$ for sparse graph

Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. 25, 163–177 (2001).

ASSIGNMENT

Web Ranking Algorithms

- Goal
  - (mandatory)Baseline: PageRank and HIST
  - (optional)Topic-sensitive PageRank [1]
  - (optional) SALSA and TKC effect [2]
  - (optional) TrustRank [3]
  - (optional) Dynamic Personalized PageRank [4]
- Datasets
  - Google Web graph: http://snap.stanford.edu/data/web-Google.html
  - Social Networks: http://snap.stanford.edu/data/ego-Facebook.html
  - Others: SNAP http://snap.stanford.edu/data/index.html

Web Ranking Algorithms

- Paper writing: a team paper can be collaborated written including:
  - abstract / problem definition / approach overview
  - details of algorithms / experiments / related works / reference
  - LNCS template with latex
  - PPT

- References

[1] Haveliwala, T.H.: Topic-sensitive PageRank. In: Proceedings of the eleventh international conference on World Wide Web - WWW2002. p. 517 (2002).

[2] Lempel, R., Moran, S.: Stochastic approach for link-structure analysis (SALSA) and the TKC effect. Comput. Networks. 33, 387–401 (2000).

[3] Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with TrustRank. Proc. Thirtieth Int. Conf. Very large data bases. 30, 576–587 (2004).

[4] Chakrabarti, S.: Dynamic personalized pagerank in entity-relation graphs. WWW2007. (2007).