

PageRank

教科書 2 章 23-30 ページ

左:Google 右:Altavista

The image shows a side-by-side comparison of search results from Google and Altavista. The browser window has a single address bar at the top with the text "Multi Search" and "university". Below the address bar, the left pane (Google) displays search results for the query "university", showing 11 results. The right pane (Altavista) displays search results for the query "national parks".

Google Search Results (Left Pane):

- Query: **university**
11 Results Returned
Showing Results From 0 to 10
- Stanford University Homepage**
http://www.stanford.edu/ 74.74% 4K - 12/19/96 - 01/03/97
- Stanford University Portfolio Collection**
http://www.stanford.edu/home/administration/portfolio.html 65.79% 3K - 12/19/96 - 01/03/97
- University of Illinois at Urbana-Champaign**
http://www.uiuc.edu/ 73.26% 15K - 12/19/96 - 01/03/97
- Indiana University**
http://www.indiana.edu/ 60.30% 2K - 09/20/96 - 01/03/97
- University of California, Irvine**
http://www.uci.edu/ 68.07% 2K - 12/19/96 - 01/03/97
- University of Minnesota**
http://www.umn.edu/ 67.05% 4K - 12/19/96 - 01/03/97
- Iowa State University Homepage**
http://www.iastate.edu/ 66.66% 3K - 12/19/96 - 01/03/97
- The University of Michigan**
http://www.umich.edu/ 66.33% 2K - 12/19/96 - 01/03/97
- Mississippi State University**
http://www.msstate.edu/ 66.35% 3K - 12/19/96 - 01/03/97
- Northwestern University NUInfo**
http://www.nwu.edu/ 66.15% 3K - 12/19/96 - 01/03/97

Altavista Search Results (Right Pane):

- Optical Physics at the University of Oregon**
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group...
<http://optics.uoregon.edu/> - size 1K - 16 Dec 96
- Carnegie Mellon University - Campus Networking**
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
<http://www.cmu.edu/cnm/> - size 4K - 19 Aug 97
- Wesleyan University Computer Science Group Home Page**
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
<http://www.cs.wesleyan.edu/> - size 5K - 15 Apr 96
- Keto University Shonan Fujisawa Campus (SFC)**
B335N%2IEPnF+Bt%-%c%e%Q%9 (B(SFC) B\$N (BWWW B\$% B\$CmOU-q\$- (B \$B4rF1s4Q4/b@455\$# (B. Nihongo | English. SFC B\$>p.s (B. [B\$%a%G%4%*%9%\$%? | *...
<http://www.sfc.kyoto.ac.jp/> - size 3K - 5 Feb 97
- School of Chemistry, University of Sydney**
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
<http://www.chem.usy.ac.au/> - size 4K - 25 Feb 97
- Mankato State University**
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events, Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
<http://www.mankato.msut.edu/> - size 3K - 27 Nov 96
- St. Ambrose University**
Main Index: Academic Departments, Administrative Services, Campus News, Computing Services, Galvin Fine Arts Center, Internet Connections, Library...
<http://www.sau.edu/> - size 2K - 4 Feb 97
- University of Washington ECSEL Projects**

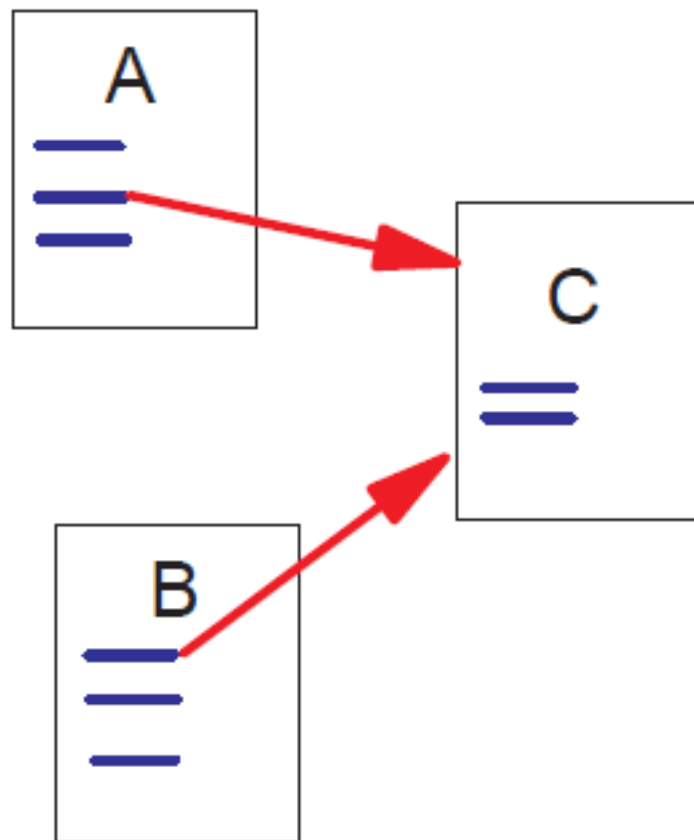
PageRank

- Google の検索エンジン
- 1998 年にスタンフォード大学の学生だった Brin と Page により提案された。2 人は現在の Google の社長

PageRank の基本的なアイデア

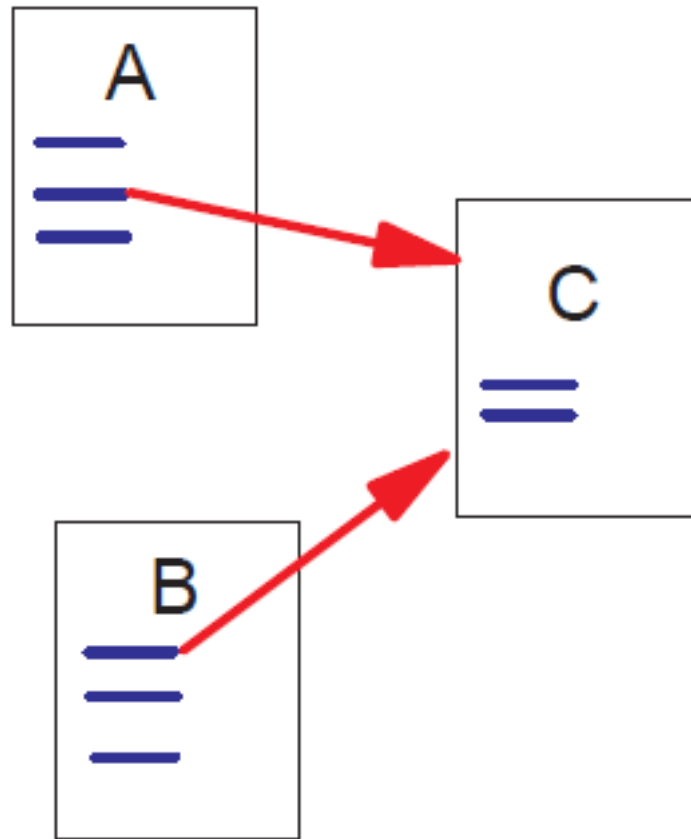
- 重要なサイトにリンクされているサイトは重要なサイト。
- 多くのサイトにリンクされているサイトは重要なサイト。
- リンク集のように多くのページへのリンクをもつサイトは重要でないサイト。

簡単な PageRank(1)



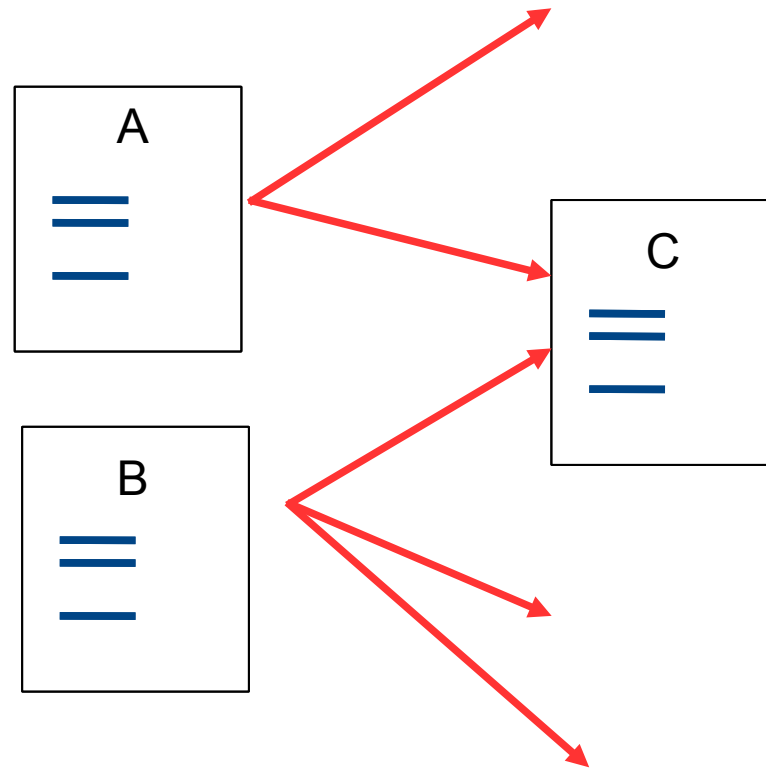
- ページ A とページ B がページ C を指しているとする。

簡単な PageRank(1)



- ページ C のランクを $x(C)$ とすると、
 - $r(C) = r(A) + r(B)$
 - A や B が重要なページであるほど PageRank 上昇
- しかし、A や B はただのリンク集かもしれない。

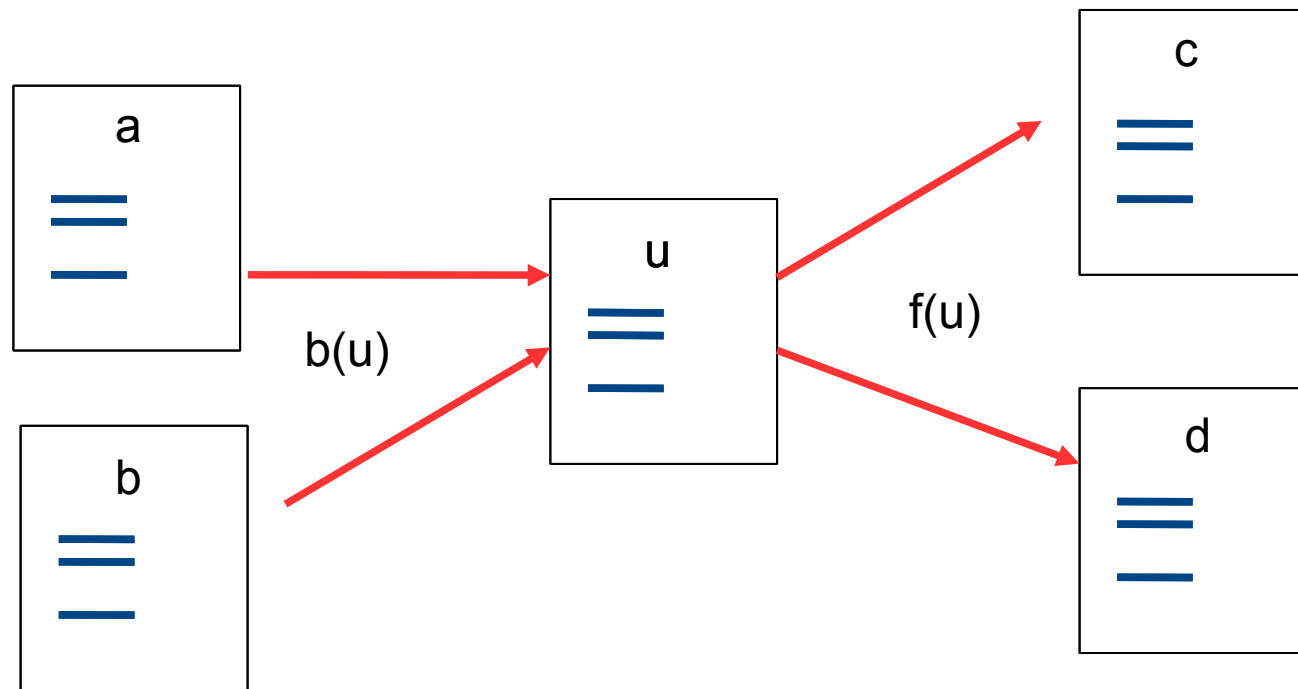
PageRank の簡単な例 (2)



- そこで、A や B から出ているリンクの数を考慮する。

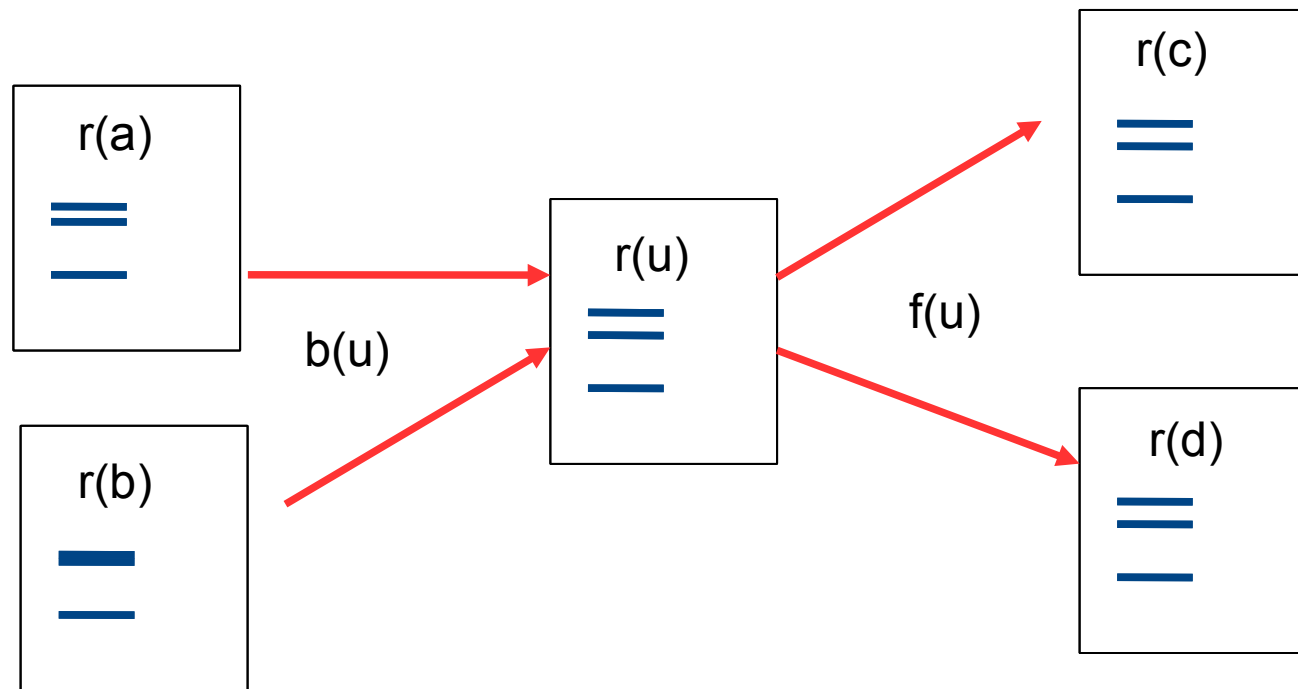
$$r(C) = r(A) / 2 + r(B) / 3$$

PageRank の簡単な例 (2) の一般化



あるウェブページ u に関して、 u を指すページによるリンク $b(u)$ と、 u 自身が指すリンク $f(u)$ の 2 種類がある。

PageRank の簡単な例 (2) の一般化

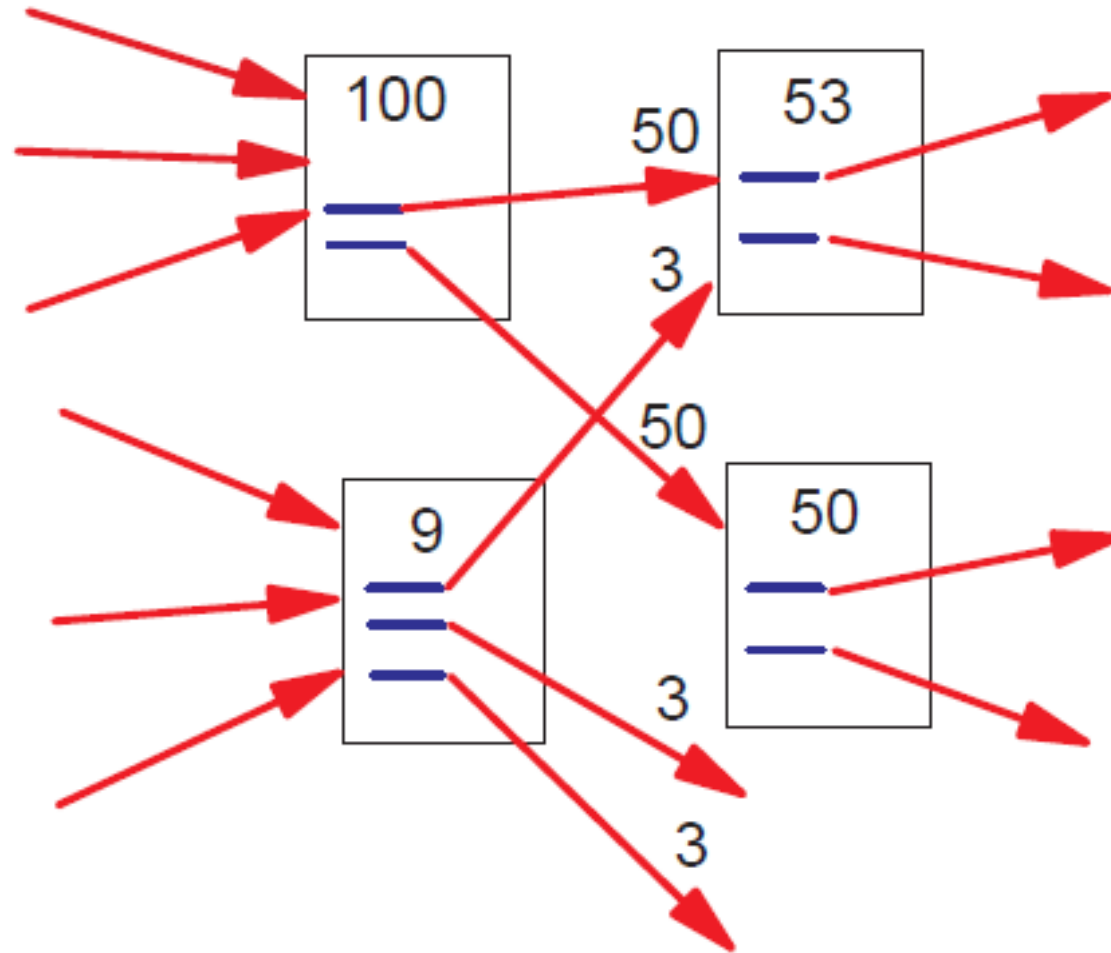


今、ページ u のランクを $r(u)$ とすると、

$$r(u) = c \sum_{v \in b(u)} r(v) / |f_v|$$

c は 1 より小さい定数 (後で解説)
 $|f_v|$ は v の forward リンクの数

簡単な PageRank(2) の具体例

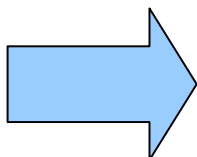
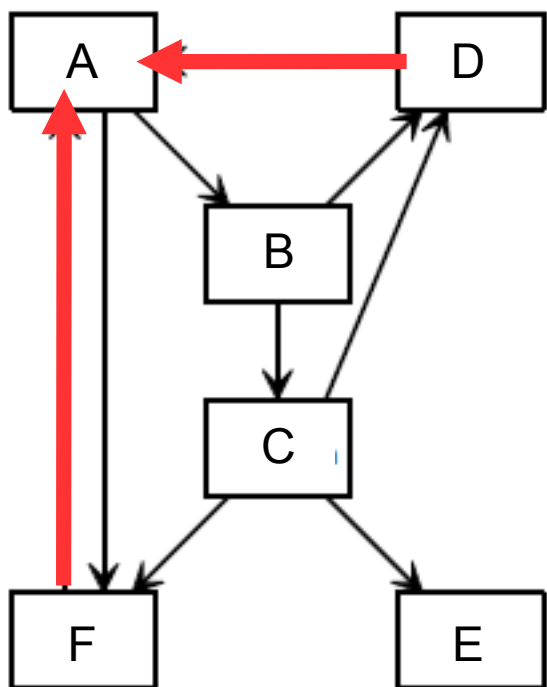


$$r(u) = c \sum_{v \in b(u)} r(v) / |f_v|$$

PageRank を行列表現で
計算してみよう

リンクの隣接行列表現

- 今、教科書 26 ページの例のリンク関係は次の正方形行列で表すことができる。この行列を隣接行列という。



	A	B	C	D	E	F
A	0	0	0	1	0	1
B	1	0	0	0	0	0
C	0	1	0	0	0	0
D	0	1	1	0	0	0
E	0	0	1	0	0	0
F	1	0	1	0	0	0

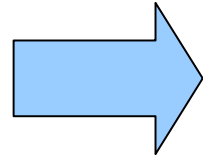
- 隣接行列の行成分が、**被リンク関係**に対応。
PageRank の計算は、**被リンク数**の数え上げです。

リンクの隣接行列表現

- 隣接行列の各列の和が1となるように正規化した行列を **A** とする。

G

0	0	0	1	0	1
1	0	0	0	0	0
0	1	0	0	0	0
0	1	1	0	0	0
0	0	1	0	0	0
1	0	1	0	0	0



A

0	0	0	1	0	1
1/2	0	0	0	0	0
0	1/2	0	0	0	0
0	1/2	1/3	0	0	0
0	0	1/3	0	0	0
1/2	0	1/3	0	0	0

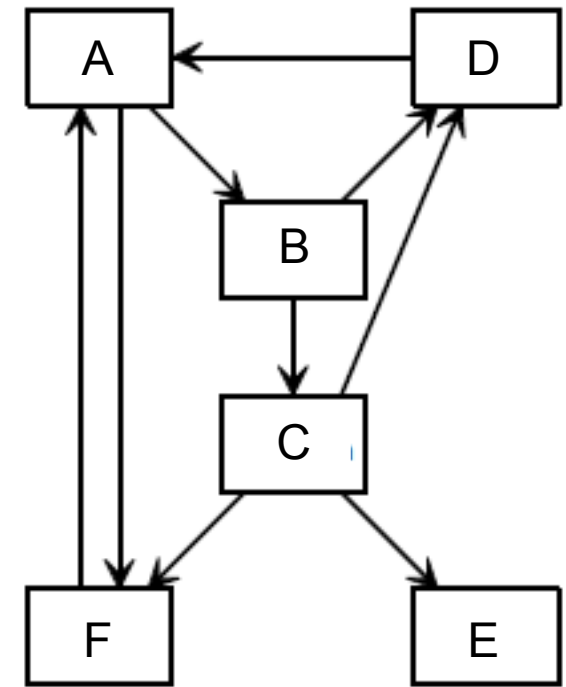
```
A = zeros(6,6)
for i = 1:6
    if sum(G(:,i)) ~= 0
        A(:,i) = G(:,i)./sum(G(:,i));
    end
end
```

各列の和が1となるように正規化

$$r(u) = c \sum_{v \in b(u)} r(v) / |f_v|$$

すると、先程の式は、行列 **A** の行成分を足し合わせるだけです。**行列 A の成分は $1/|f|$**

	A						各行の和
A	0	0	0	1	0	1	2
B	1/2	0	0	0	0	0	1/2
C	0	1/2	0	0	0	0	1/2
D	0	1/2	1/3	0	0	0	5/6
E	0	0	1/3	0	0	0	1/3
F	1/2	0	1/3	0	0	0	5/6
	A	B	C	D	E	F	



$r = \text{sum}(A, 2)$

行列 A に対して $\text{sum}(A, 1)$ で A の列成分の和を計算。 $\text{sum}(A, 2)$ で A の行成分の和を計算。

$$r(u) = c \sum_{v \in b(u)} r(v) / |f_v|$$

- PageRank の式は r が 左辺と右辺に出てくる再帰的な式なので、収束するまで何度も r を再計算する

- 行列 A が正規化されているとき、その成分は $1/|f|$ なので、PageRank の式

$$r(u) = c \sum_{v \in b(u)} r(v) / |f_v|$$

は $\mathbf{r} = c \times \mathbf{A} \times \mathbf{r}$ という行列式
と等価です。

Octave を用いた PageRank の計算

- 次の rec は、行列 **A** を引数(入力)として、ページランク **r** を出力とする関数です。
- **r** の再計算を 100 回行います。

```
rec.m  
function [r] = rec(A)  
  
r = ones(6,1);  
  
for i = 1:100  
    r = A*r  
end
```

初期値はゼロベクトルでなければ何でも
いいです。無限回繰
り返せば同じ結果に
収束します。

計算結果

r= 0.0657717
0.0339851
0.0175606
0.0236098
0.0060492
0.0400344

$$\mathbf{r} = \mathbf{C} \times \mathbf{A} \times \mathbf{r} \quad \left(\mathbf{A} \mathbf{r} = \frac{1}{c} \mathbf{r} \right)$$

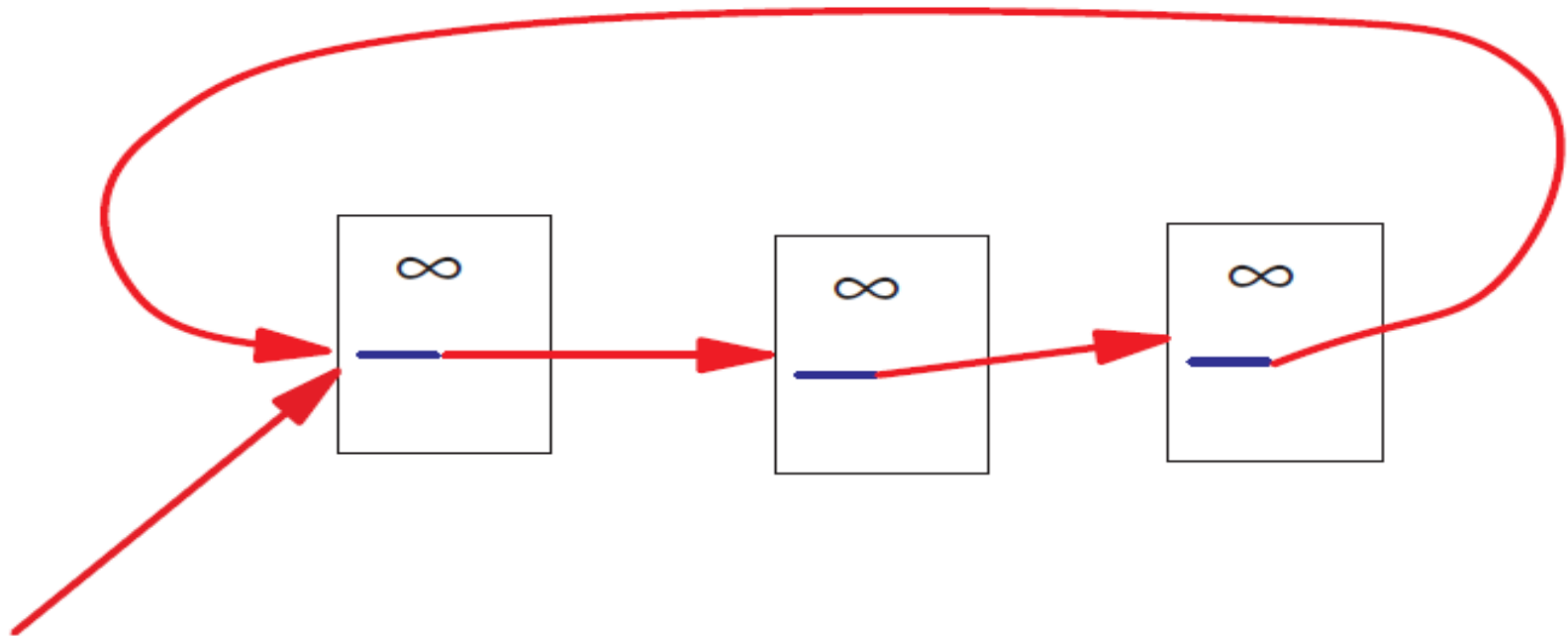
- この式は固有方程式。1/c が固有値、 \mathbf{r} が固有ベクトルです。
- また、先程の繰り返し関数は、行列の最大固有値を見つけるためのべき乗法 (power method) と同じです。

```
power.m  
function [r] = power(A)  
  
r = ones(6,1);  
for i = 1:100  
    r = A*r  
    r = r./norm(r)  
end
```

もし A も r も正規化
されていないと、不安
定になって桁落ちし
やすいので注意

これまでの PageRank の問題

- もし、次のような閉じたループがある時は、ランクは外に出ていくことなく、蓄積されてしまう。
- 実際の行動でいえば、ウェブサーファースは同じページを巡回することになり、不自然。
- そこで、実際のウェブサーファースはある程度ランダムにこの閉じたループから飛び出すと仮定。



改良 PageRank

- 全てのページ u に対して、外からランダムに人が訪れると考え、これを単位ベクトル e であらわす。

$$r(u) = c \left(\sum_{v \in b(u)} r(v) / |f_v| + e \right)$$

- 行列で書くと $\mathbf{r} = c (\mathbf{A} \times \mathbf{r} + \mathbf{e})$
もしくは \mathbf{E} を単位対角行列として

$$\mathbf{r} = (\mathbf{E} - c \mathbf{A})^{-1} c \mathbf{e}$$

pagerank.m

```
function [r] = pagerank(A)
c = 0.85;
n = size(A,1);
r = (eye(n)-c*A)\(c*ones(n,1));
```

1996 年 1 月の Top15 ウェブサイト

Web Page	PageRank (average is 1.0)
Download Netscape Software	11589.00
http://www.w3.org/	10717.70
Welcome to Netscape	8673.51
Point: It's What You're Searching For	7930.92
Web-Counter Home Page	7254.97
The Blue Ribbon Campaign for Online Free Speech	7010.39
CERN Welcome	6562.49
Yahoo!	6561.80
Welcome to Netscape	6203.47
Wusage 4.1: A Usage Statistics System For Web Servers	5963.27
The World Wide Web Consortium (W3C)	5672.21
Lycos, Inc. Home Page	4683.31
Starting Point	4501.98
Welcome to Magellan!	3866.82
Oracle Corporation	3587.63

今日のまとめ

- PageRank のしくみ
- 行列を用いた PageRank の計算法

演習

- PageRank の定数 c は 1 より小さい値とするべきである。なぜか？

- 答えは教科書中

- 一般的な検索結果とパーソナライズド検索の結果は異なる。どのように個人情報を使って検索結果の精度を向上できるか？

- 答えは原著論文中

<http://www.cs.umd.edu/areas/db/dbchat/papers/pageranksub.pdf>

Reference

The PageRank Citation Ranking: Bringing Order to the Web, Lawrence Page and Sergey Brin, Technical Report, Stanford University 1999