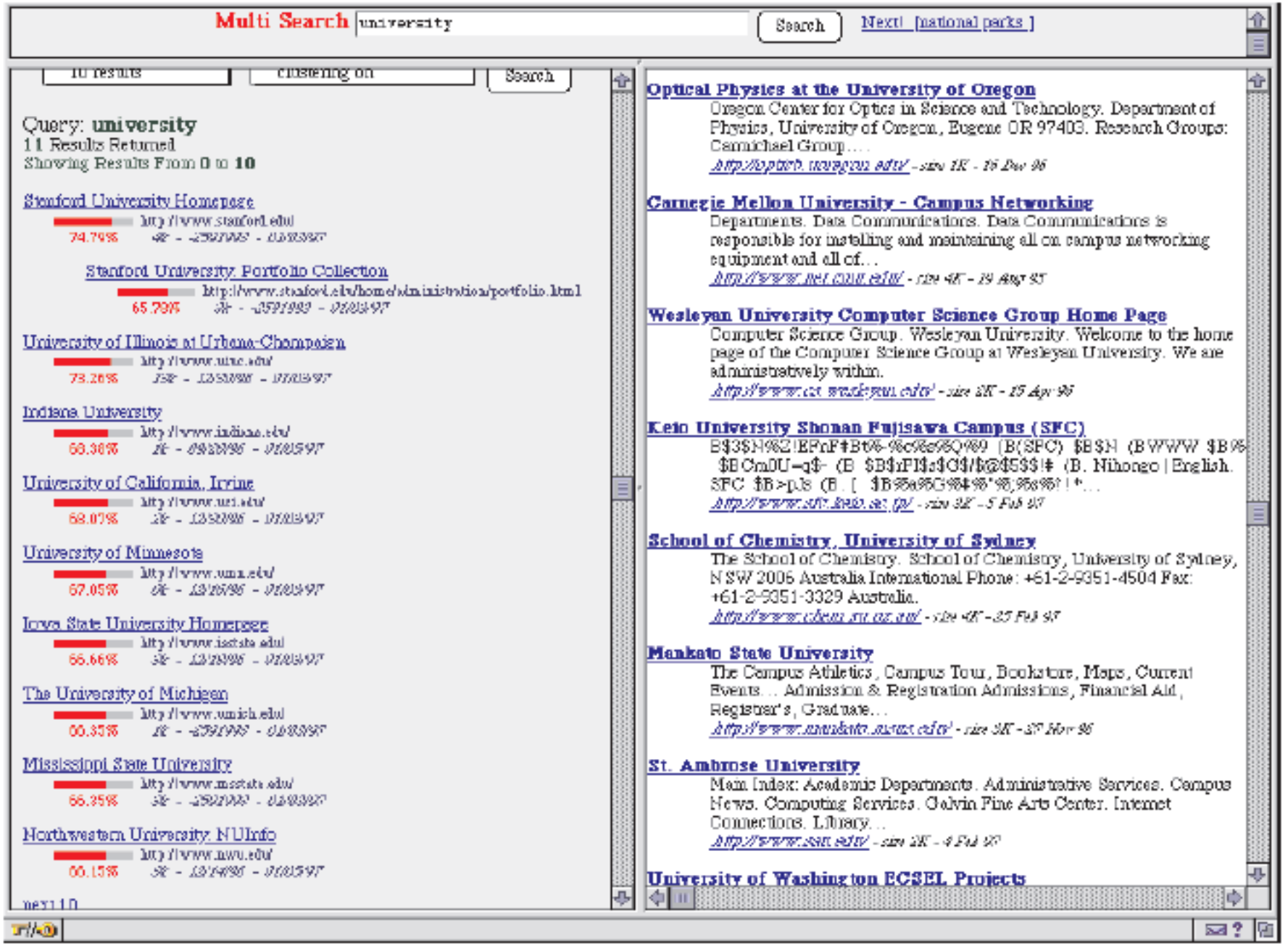
PageRank

教科書 2 章 23-30 ページ

左 :Google　右 :Altavista



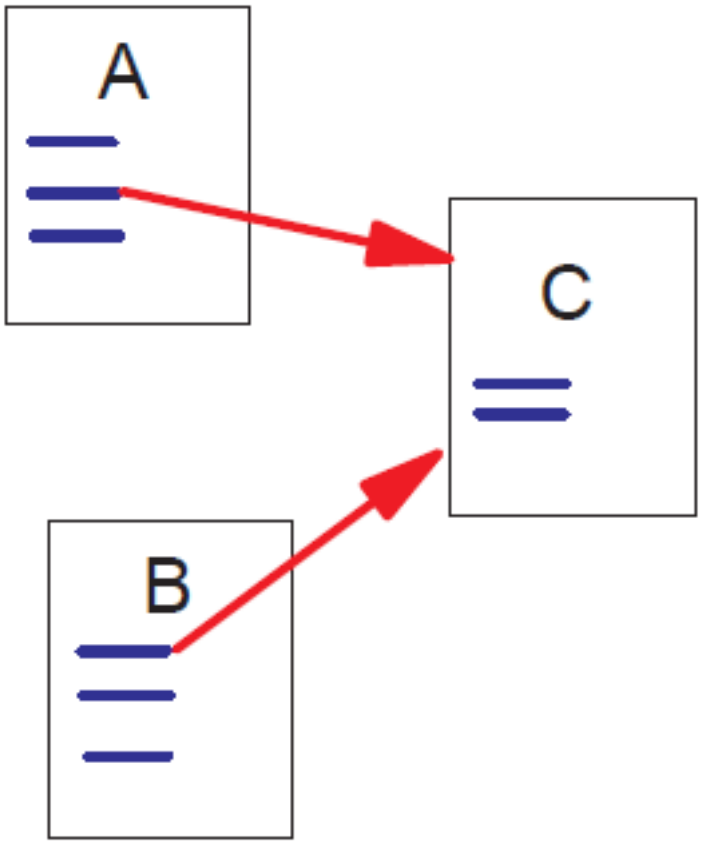
PageRank

- 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

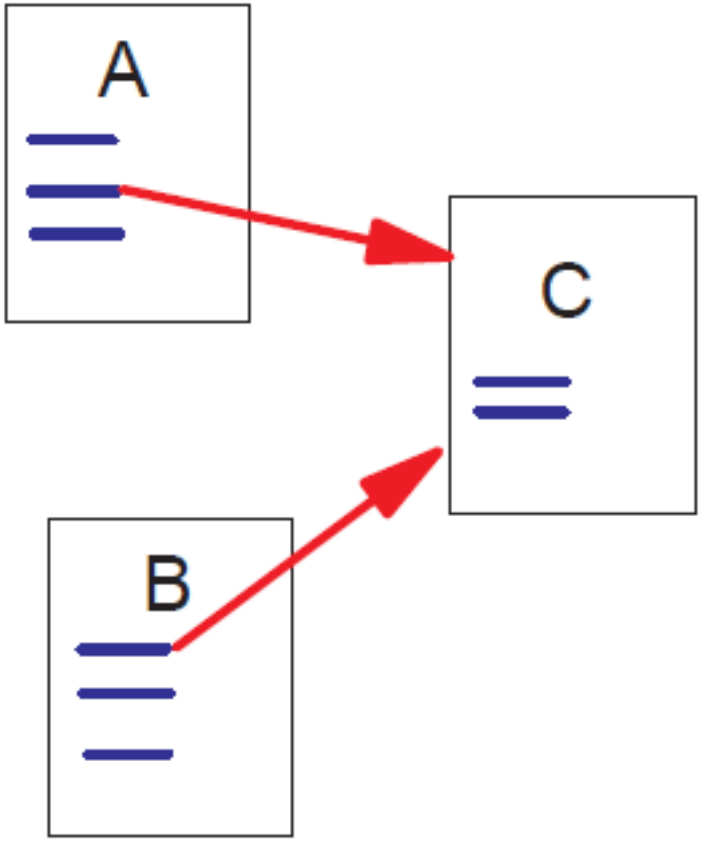
- 重要なサイトにリンクされているサイトは重要なサイト。
- 多くのサイトにリンクされているサイトは重要なサイト。
- リンク集のように多くのページへのリンクをもつサイトは重要でないサイト。

簡単な PageRank(1)



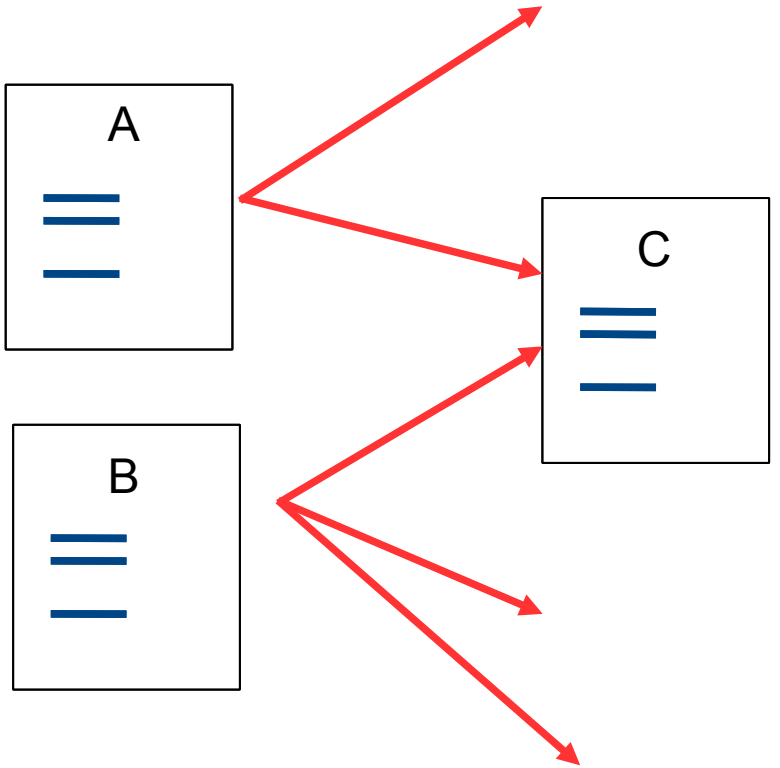
- ページ A とページ B がページ C を指しているとする。

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。

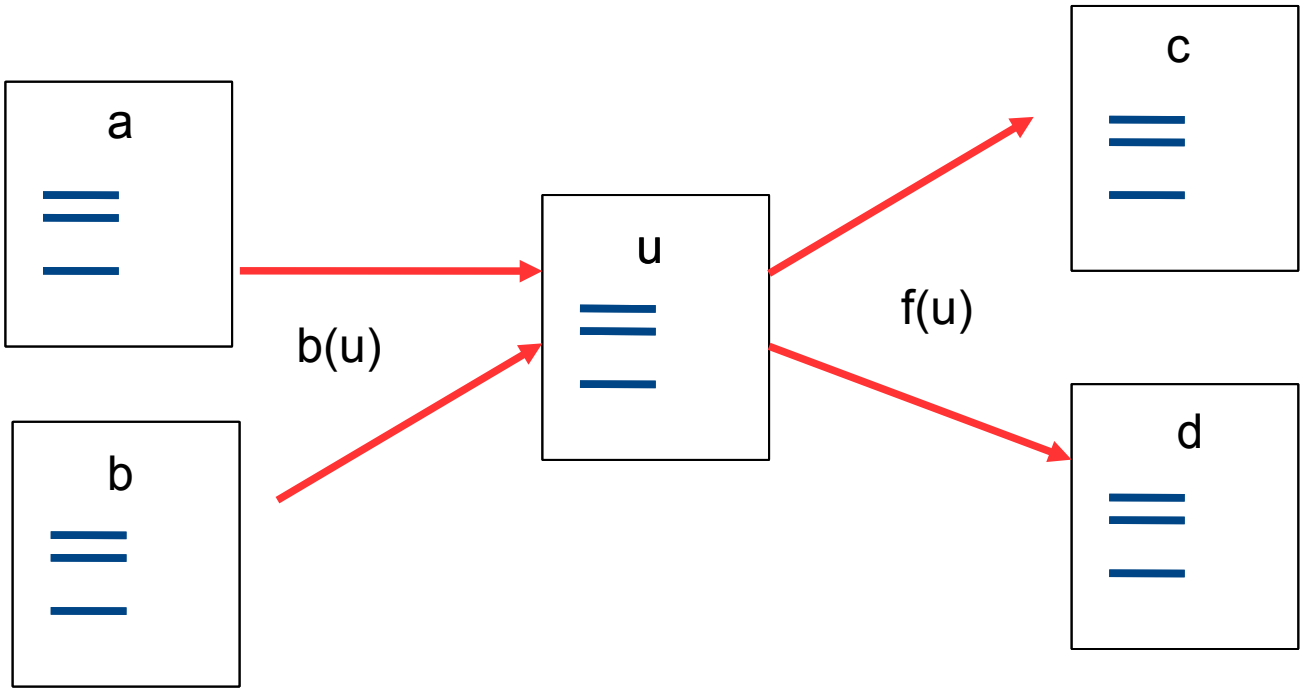
PageRank の簡単な例 (2)



- そこで、A や B から出ているリンクの数を考慮する。

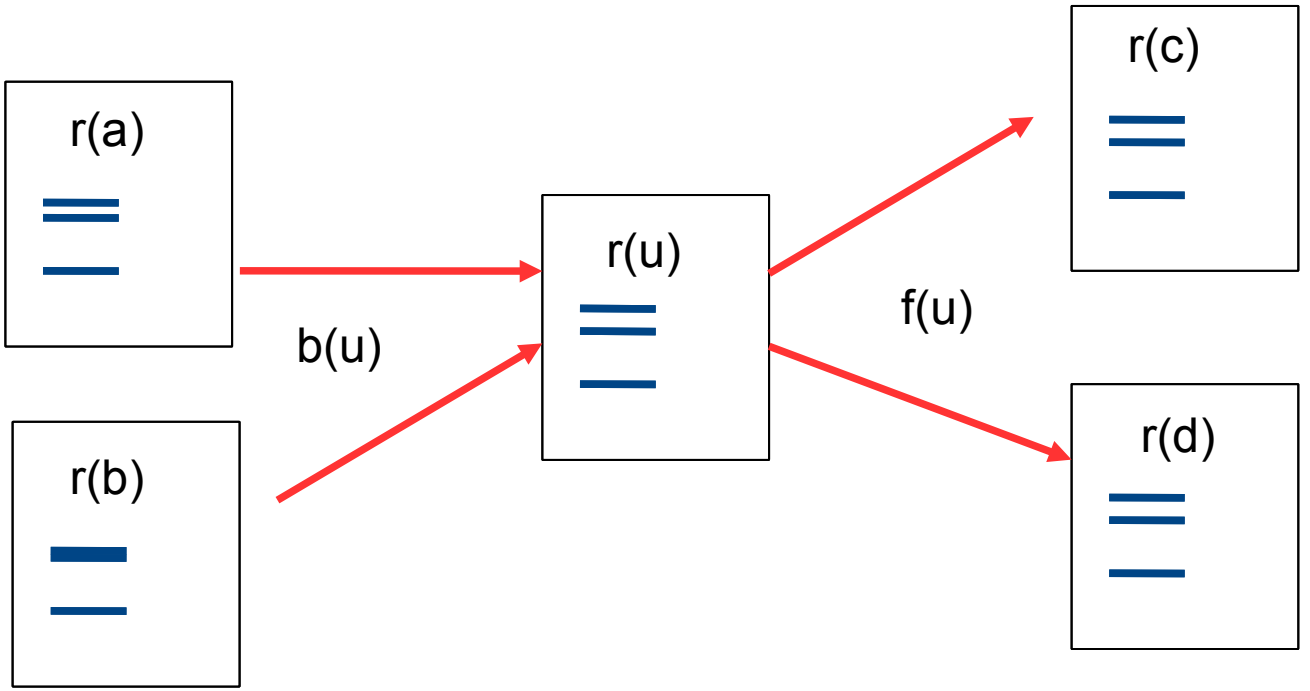
  r(C) = r(A) / 2 + r(B) / 3

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

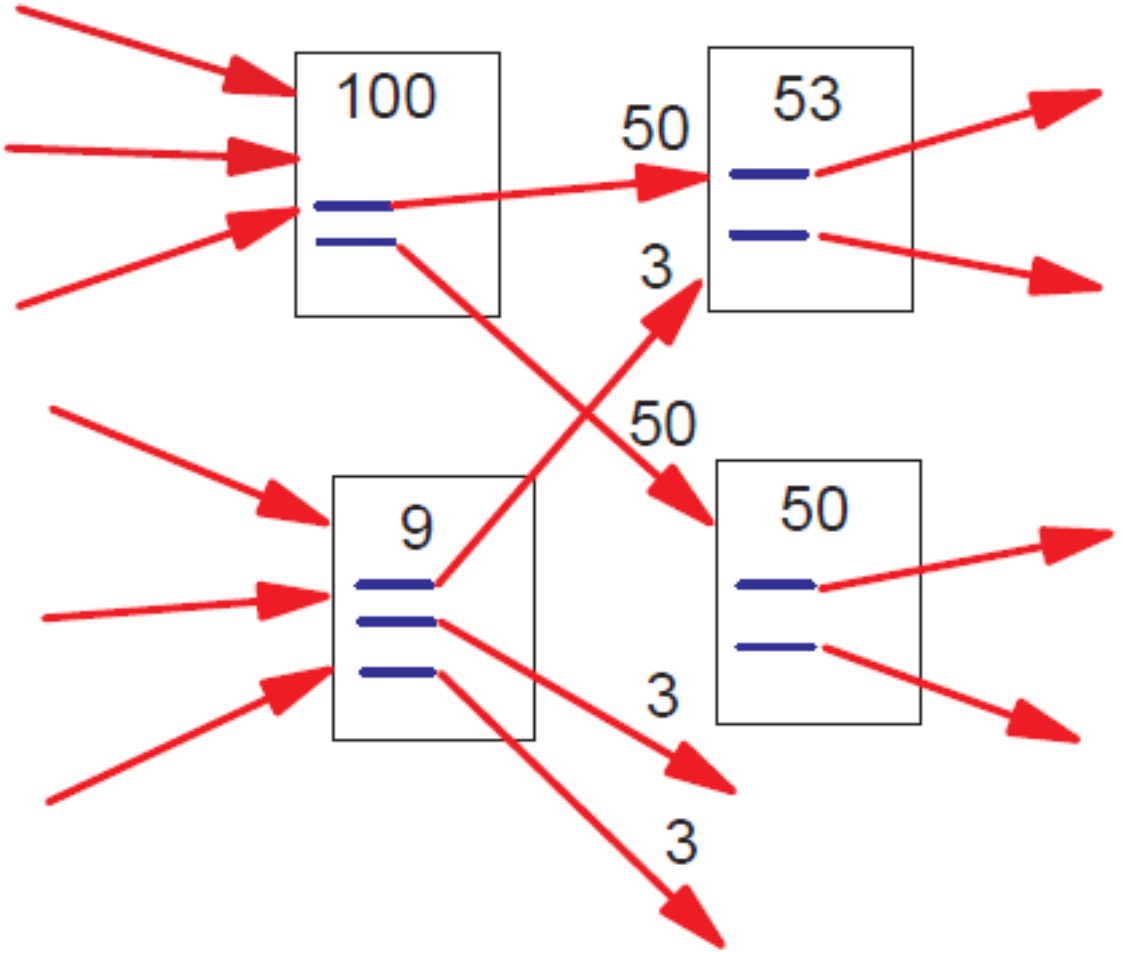


今、ページ u のランクを r(u) とすると、

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

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

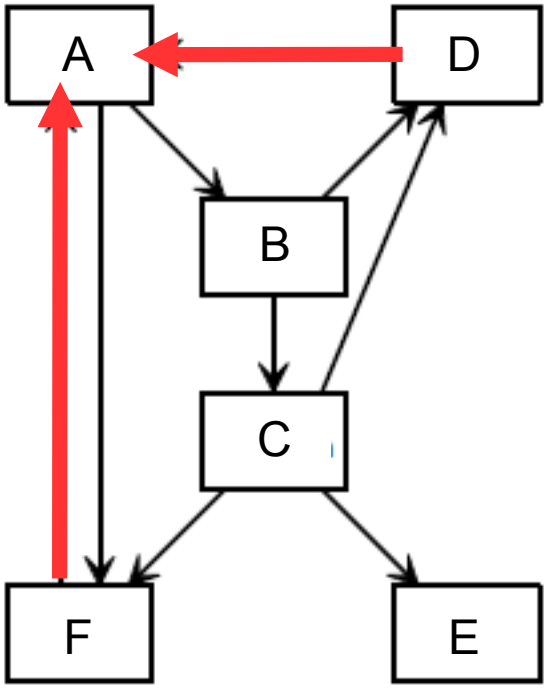


$$r(u) = c \sum_{v \in b(u)} r(v) / |f_v|$$

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行列で表すことができる。この行列を隣接行列という。



$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}
\begin{array}{c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}
\end{array}
$$

- 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となるように正則化した行列を **A** とする。

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

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 & 0 & 0 \end{pmatrix}$$

```
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となるように正則化

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

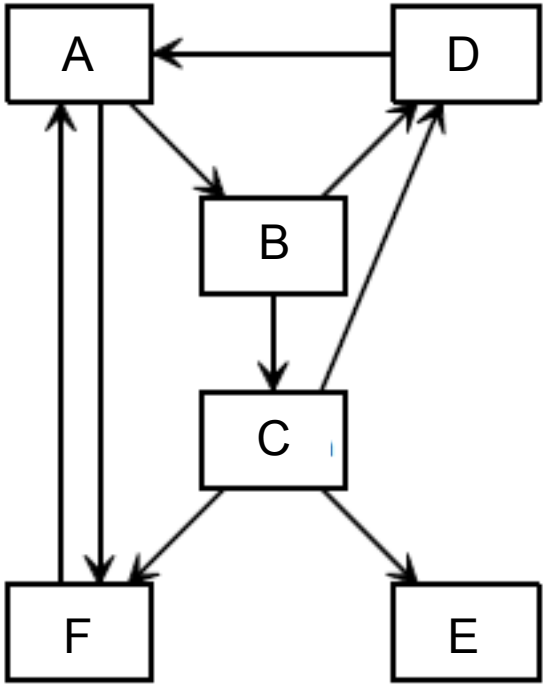
すると、先程の式は、行列 **A** の行成分を足し合わせるだけです。行列 **A** の成分は 1/| f |

A

	A	B	C	D	E	F
A	0	0	0	1	0	1
B	1/2	0	0	0	0	0
C	0	1/2	0	0	0	0
D	0	1/2	1/3	0	0	0
E	0	0	1/3	0	0	0
F	1/2	0	1/3	0	0	0

各行の和

2
1/2
1/2
5/6
1/3
5/6



r = sum(A,2)

行列 A に対して sum(A,1) で A の列成分の和を計算。 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　$\boldsymbol{r} = c \times \boldsymbol{A} \times \boldsymbol{r}$　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
```

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

- この式は固有方程式。1/c が固有値、**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

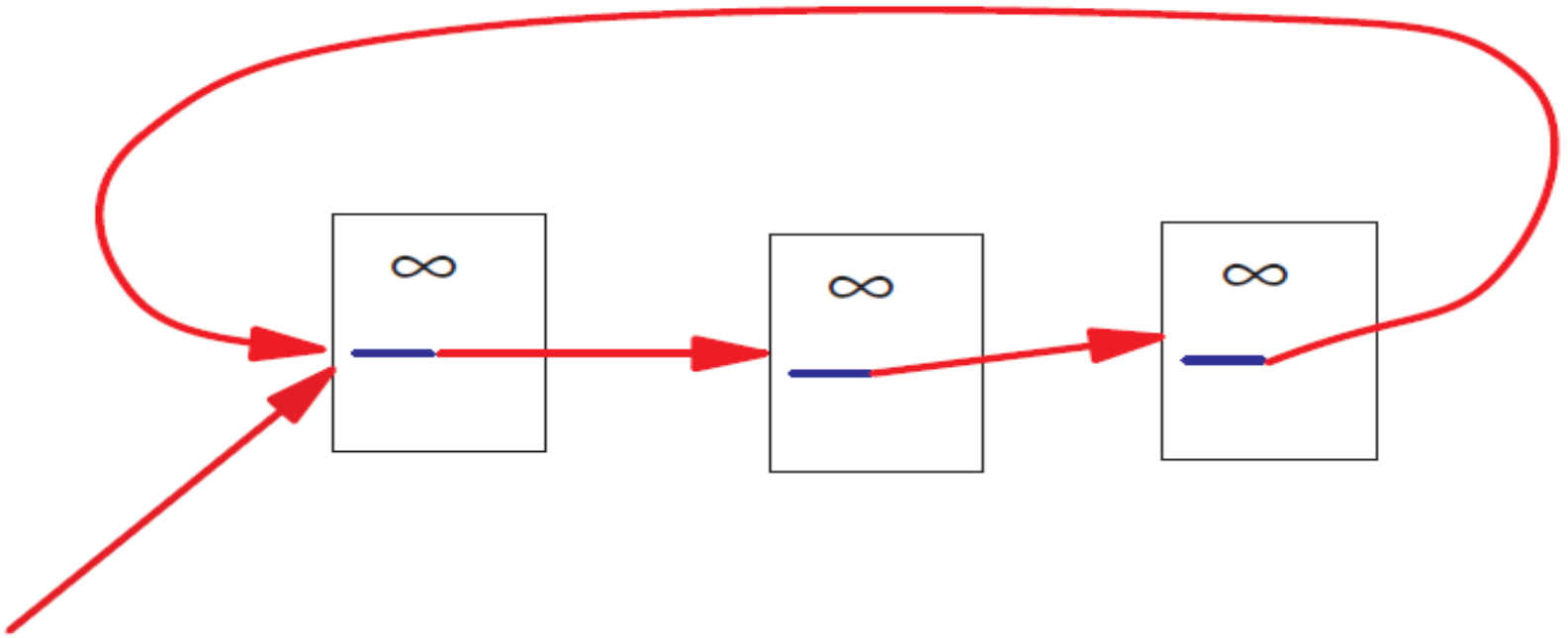
```
function [r] = power(A)

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 $\boldsymbol{r} = c\left(\boldsymbol{A} \times \boldsymbol{r} + \boldsymbol{e}\right)$

  もしくは **E** を 単位対角行列として

$$\boldsymbol{r} = \left(\boldsymbol{E} - c\,\boldsymbol{A}\right)^{-1} c\,\boldsymbol{e}$$

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