データベースと情報検索

情報検索(５)　検索エンジンの仕組み

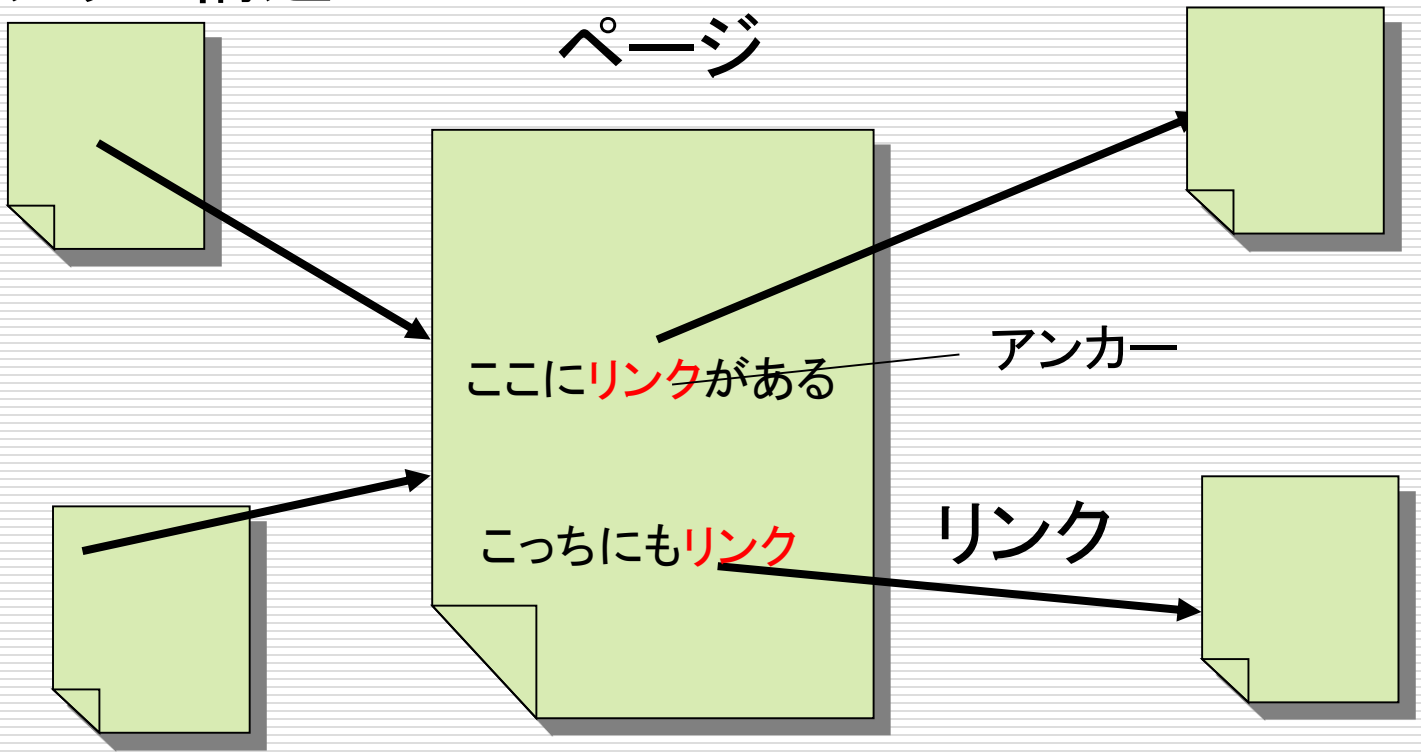
教員　岩村　雅一

日程(情報検索:担当　岩村)

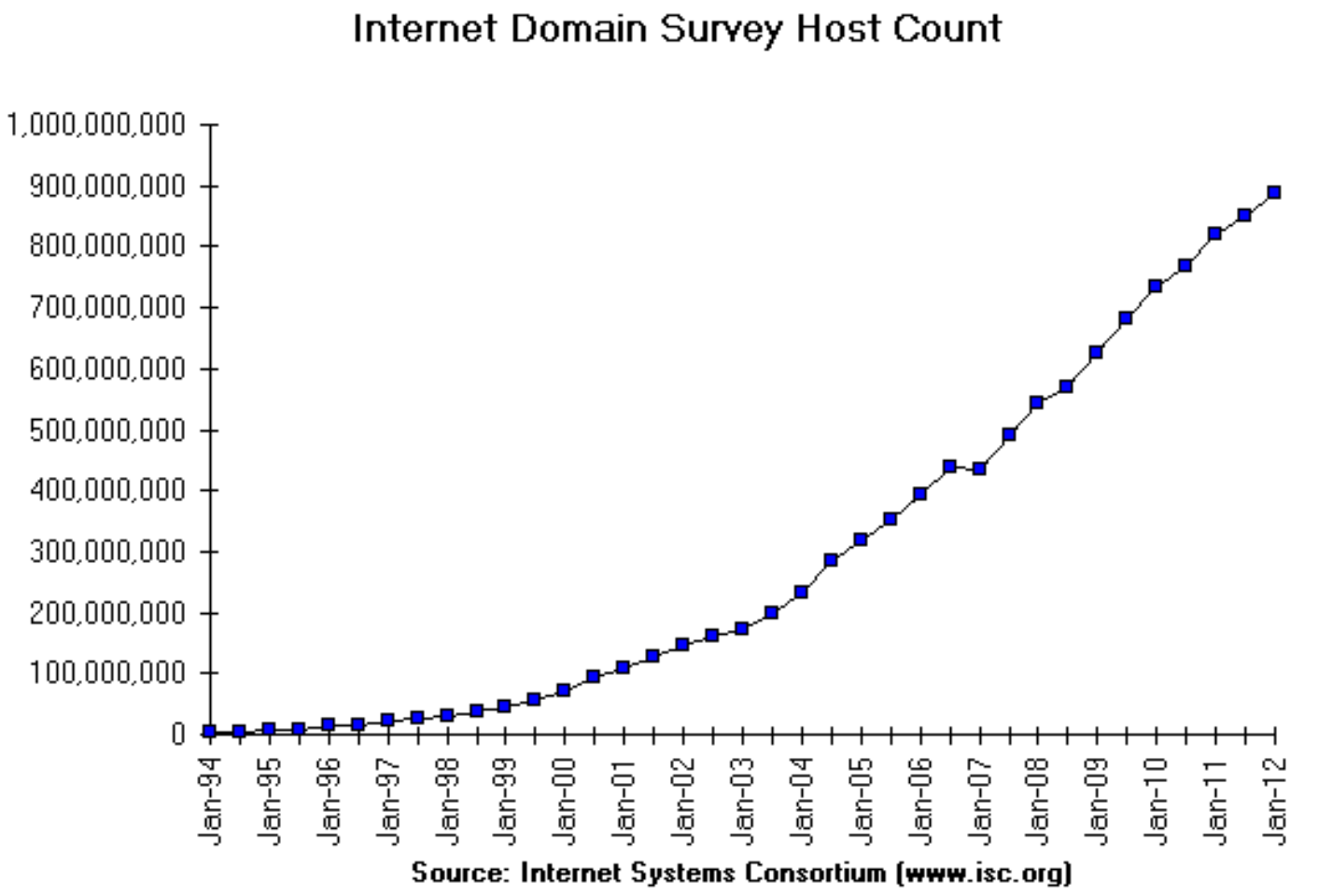
- 12/9　検索エンジンを使ってみる
- 12/16　メディア検索を使ってみる
- 12/25　ウェブアプリケーションを使ってみる
- 1/9　検索エンジンを用いた演習
- 1/20　検索エンジンの仕組み
- 1/27　メディア検索の仕組み
- 2/3　消費者生成メディアの最近

Webの構造

グラフ構造

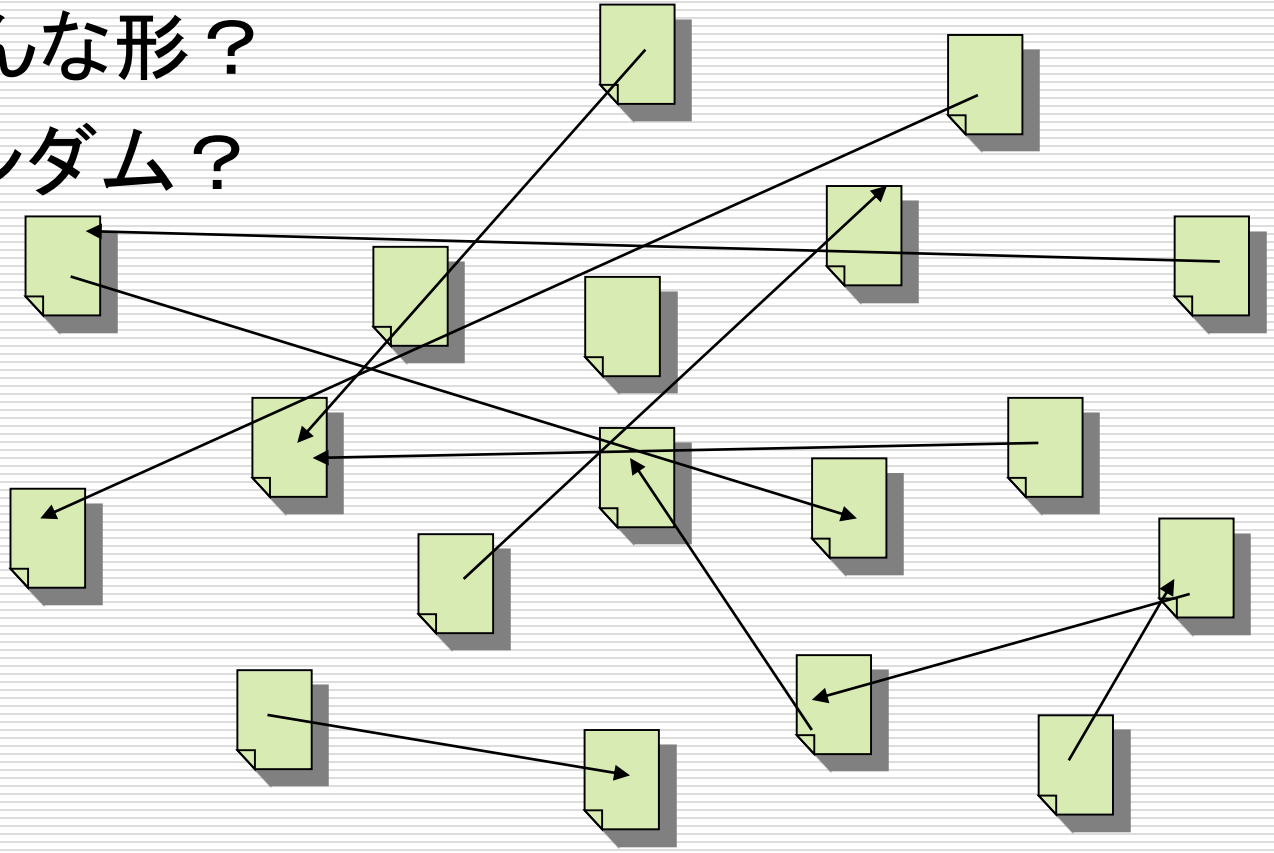


Webのサイズ


Internet Domain Survey Host Count
1,000,000,000
900,000,000
800,000,000
700,000,000
600,000,000
500,000,000
400,000,000
300,000,000
200,000,000
100,000,000
0
Jan-94
Jan-95
Jan-96
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Jan-06
Jan-07
Jan-08
Jan-09
Jan-10
Jan-11
Jan-12
Source: Internet Systems Consortium (www.isc.org)

Webの地図

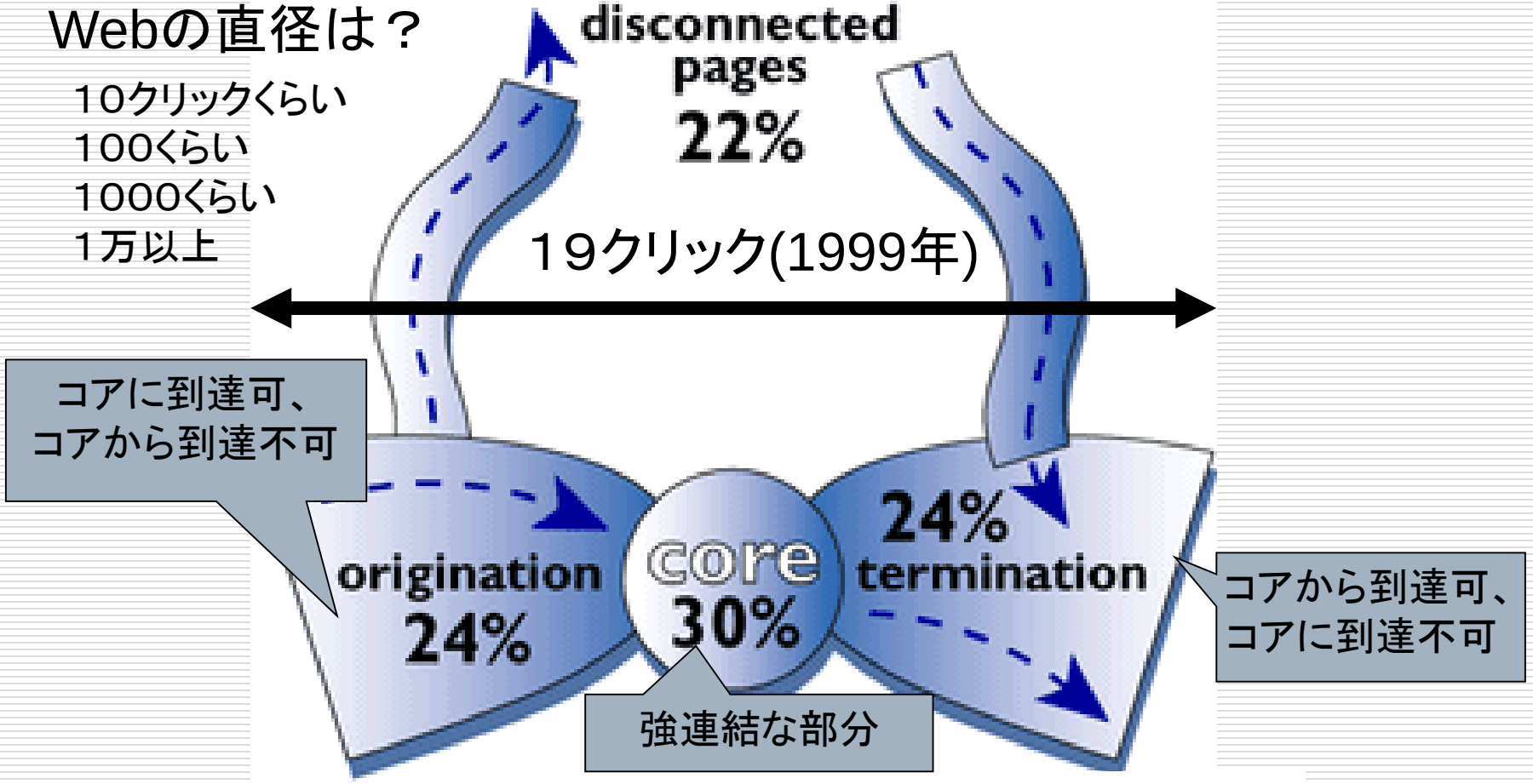
- どんな形？
- ランダム？



Webの地図： 蝶ネクタイ理論

Webの直径は？

- 10クリックくらい
- 100くらい
- 1000くらい
- 1万以上



IBMのHPより

Webの利用（アンケート）

- Webでの調べ物
  - ディレクトリ・サービス主体？
  - 検索エンジン主体？
- 検索エンジンに入れるキーワードの数は？
  - １個
  - ２個
  - ３個
  - ４個
  - ５個
  - それ以上

検索キーワード数

- OneStat.com 調べ（2004年7月）
  1. 2語: 30.09%
  2. 1語: 26.83%
  3. 3語: 16.60%
  4. 4語: 14.83%
  5. 5語: 6.76%
  6. 6語: 2.81%
  7. 7語: 1.13%

簡単な検索

- キーワードの有無
- 100億ものページを、数語で区別可能？
  - 限界あり
- 別の、何か賢い方法が必要？
  - どのような可能性が考えられるか？

参考文献

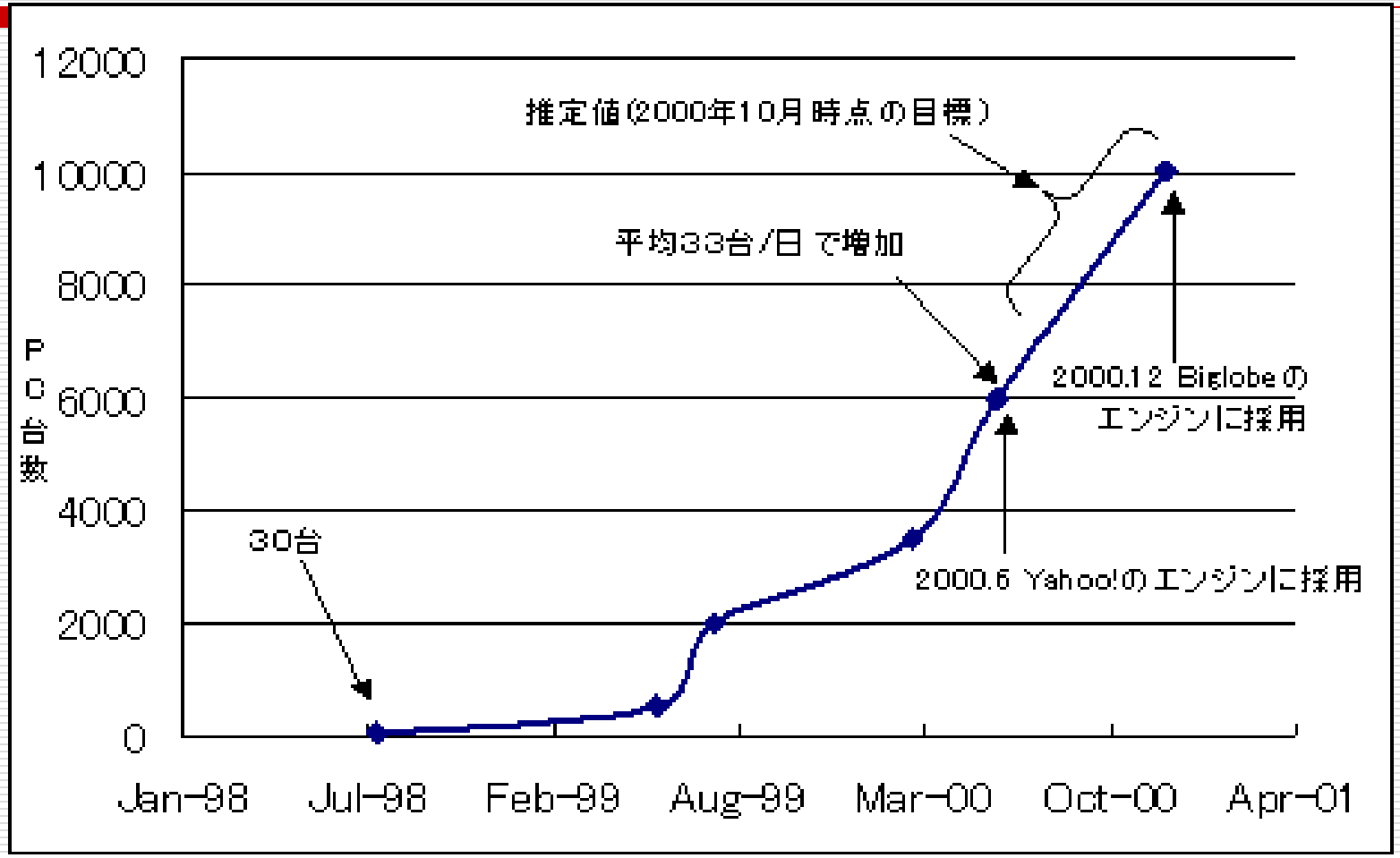
- Google の秘密 - PageRank 徹底解説
  馬場肇
  http://homepage2.nifty.com/baba_hajime/wais/pagerank.html
- サーチエンジンGoogle
  山名早人、近藤秀和
  情報処理, Vol.42, No.8, 2001
- WWWサーチエンジンの作り方
  原田昌紀
  情報処理, Vol.41, No.10, 2000

Google

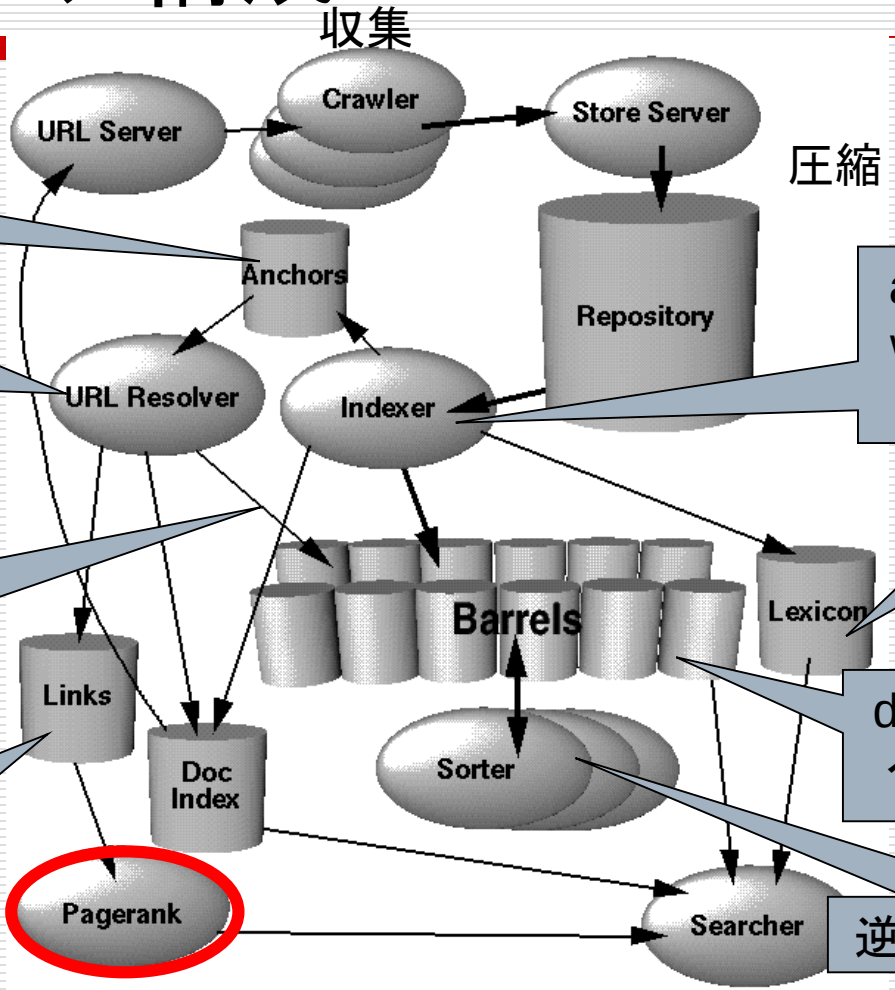
- Page & Brin により設立された（1998）
  - Stanfordの大学院生
  - データマイニングを研究

- 世界最大級の情報を持つ検索エンジン
  - 80億ページ（2005.4現在）
  - クラスタ・コンピューティング
    - PC4．5万台から8万台（CPUは倍；予測値）
    - 2千～6千テラバイト（1テラ＝1,000,000,000,000＝1兆）

PC台数の推移


12000
10000
8000
6000
4000
2000
0
PC台数
推定値(2000年10月時点の目標)
平均33台/日で増加
2000.12 Biglobeの
エンジンに採用
30台
2000.6 Yahoo!のエンジンに採用
Jan-98
Jul-98
Feb-99
Aug-99
Mar-00
Oct-00
Apr-01

ソフトウェア構成



アンカーの情報

相対URLを絶対URLに変更

アンカー部分のテキスト情報

webページの相互リンク情報

anchor, word, word位置などの抽出

wordからword-IDへのハッシュ

doc-IDからword-IDへの索引とその逆

逆向きの索引を作成

Mining＝採鉱（鉱石を採掘すること）

- Data Mining
  - データ＝鉱山
  - 埋もれた有益な情報＝鉱石
- Text Mining
  - データがテキストとして与えられたもの
  - IBMの事例が有名
- Web Mining
  - Mining の対象がweb
  - PageRankは Web Mining の一種

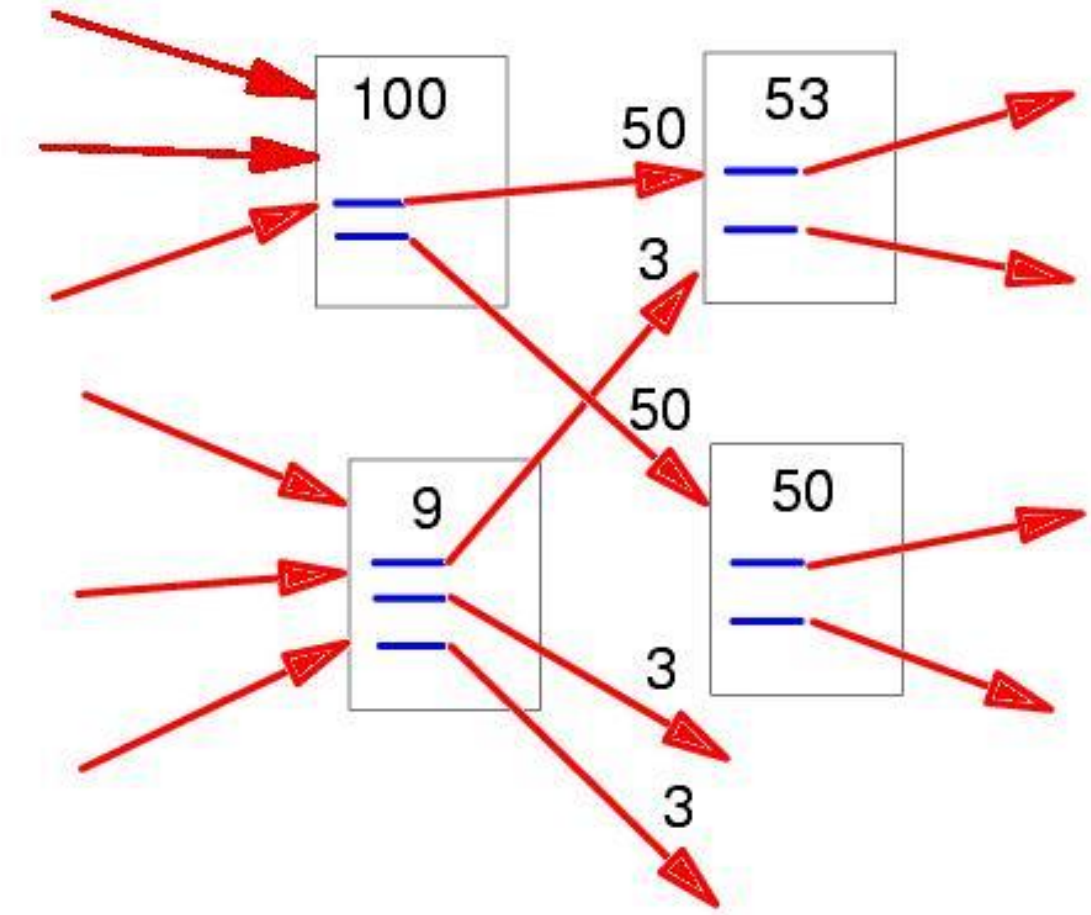
Web Mining

- Web Contents Mining
  - Webからの情報抽出やテキストマイニング
- Web Usage Mining
  - ログやクリック履歴を解析してアクセスパターンを分析
- Web Structure Mining
  - リンク構造に基づくマイニング
  - PageRankはこの一種

PageRank

- 基本的な考え方
  - 「多くの重要なページからリンクされているページは、やはり重要なページである。」
- リンク＝投票
  - ただし、1ページが1票持っているのではない
  - ページの「重要度」に応じた票数

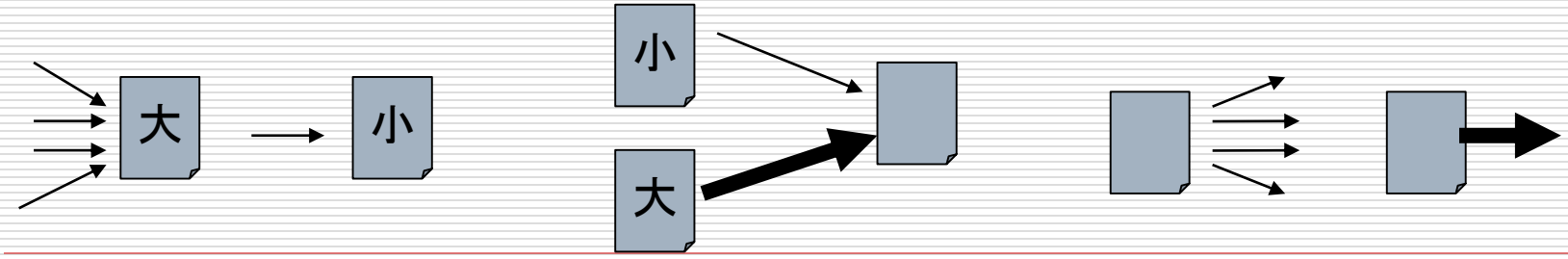
重要度



Google の秘密 - PageRank 徹底解説　**馬場肇　より引用**

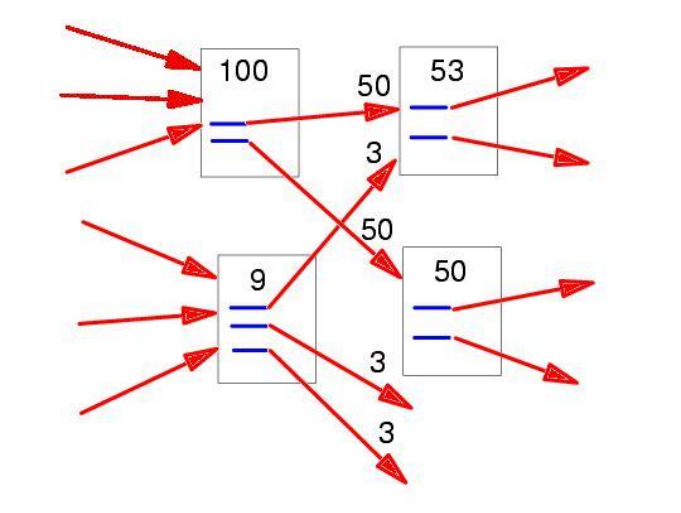
重要度の意味

- 被リンク数
  - リンクされていれば、それだけ重要度は大
- リンク元の重要度
  - 重要度が高いページからのリンクは高く評価
- リンク元のリンク数
  - 選び抜かれたリンクならば重要視

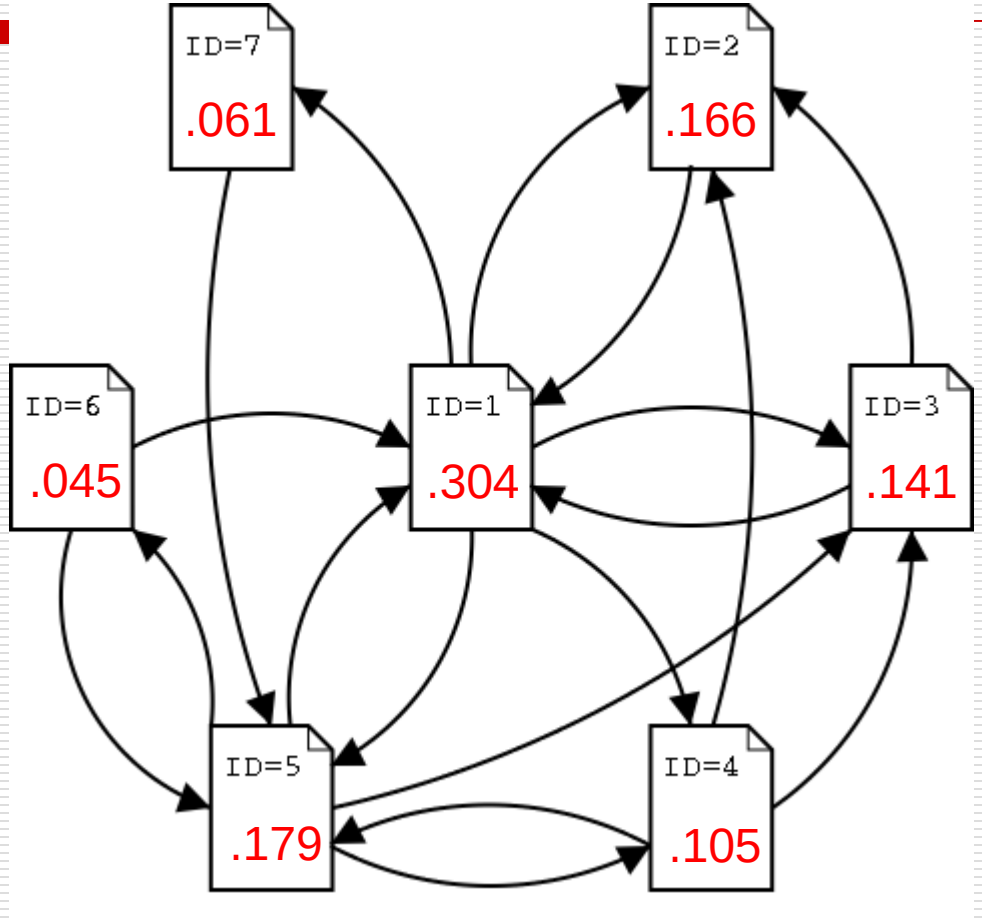


PageRankの計算

- 重要度の初期値を定める
- 推移確率に従って重要度を伝播
- 収束した結果をPageRankとする



小規模な例に対するPageRank



PageRankの値が
最大のページは？

Google の秘密 - PageRank 徹底解説 **馬場肇** より引用

PageRankの評価

順位	PageRank	文書ID	発リンクID	被リンクID
1	0.304	1	2,3,4,5,7	2,3,5,6
2	0.179	5	1,3,4,6	1,4,6,7
3	0.166	2	1	1,3,4
4	0.141	3	1,2	1,4,5
5	0.105	4	2,3,5	1,5
6	0.061	7	5	1
7	0.045	6	1,5	5

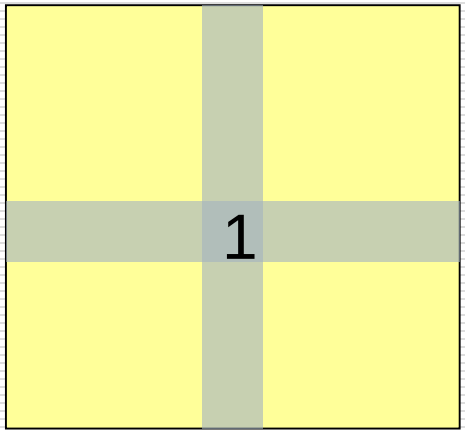
PageRankの意味と計算

- ランダムにリンクを辿るユーザが、
- 一定時間に、各ページを訪問する確率
- ちょっと高度な内容
  - 推移確率を行列で表したとき最大固有値に対する固有ベクトルがPageRankとなる
  - 詳しいことは、Googleで「PageRank」を検索して出てくる「 Google の秘密 - PageRank 徹底解説」を見て！

リンク構造の表現

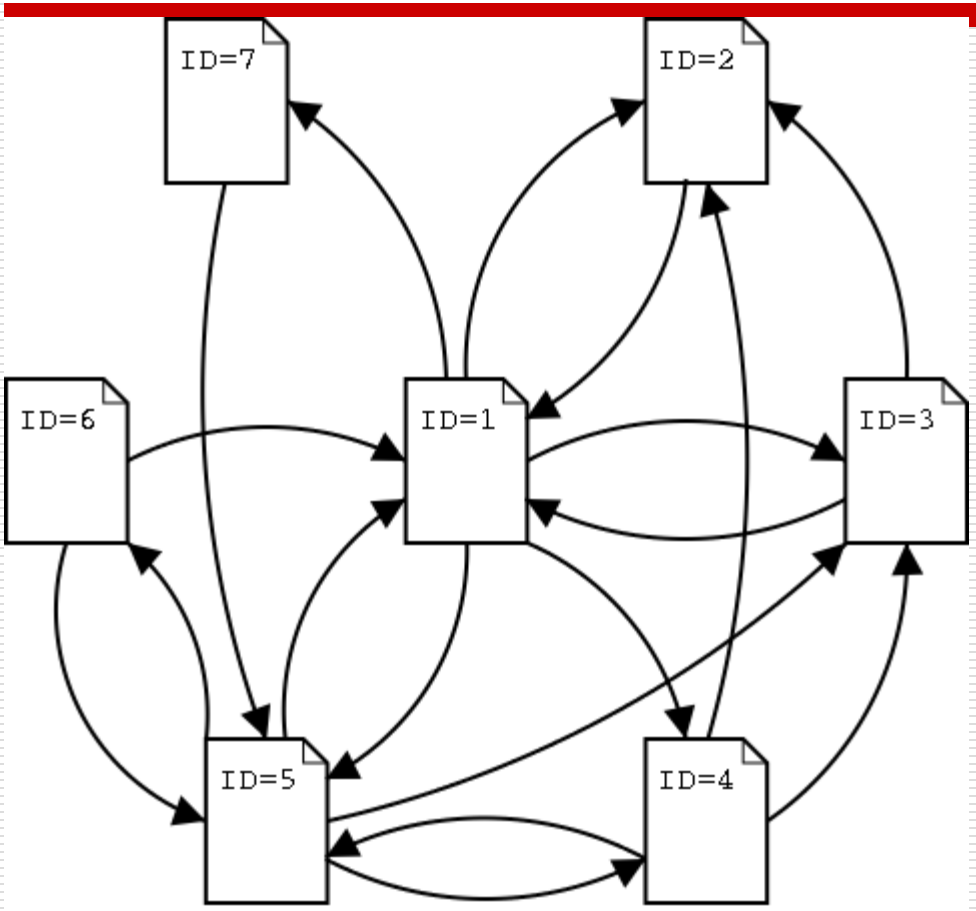
- 隣接行列で表す

A=



ページ i から j にリンクがあれば $a_{ij}=1$

小規模な例



$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{matrix} \mathcal{TO} \\ \mathcal{F} \\ \mathcal{R} \\ \mathcal{O} \\ \mathcal{M} \end{matrix}$$

推移確率行列

- 推移確率行列M

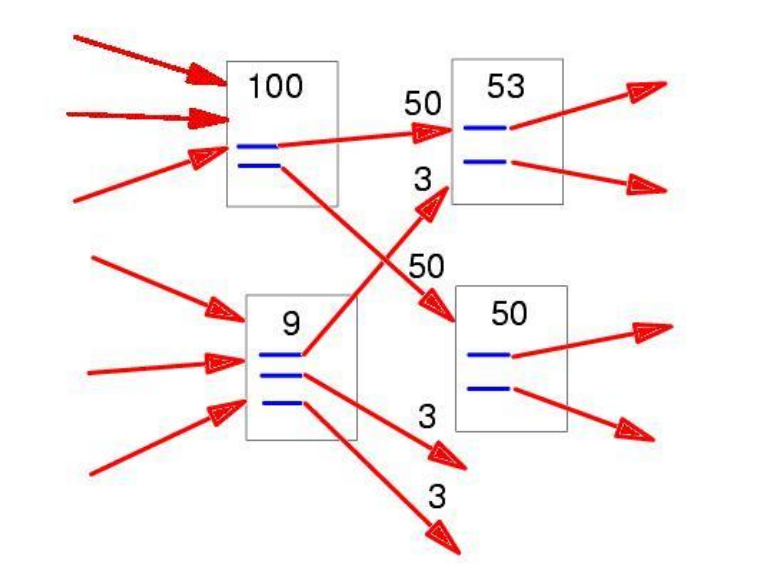
$$A^{T} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{matrix} \mathcal{FROM} \\ \mathcal{T} \\ \mathcal{O} \end{matrix}$$

$$M = \begin{pmatrix} 0 & 1 & 1/2 & 0 & 1/4 & 1/2 & 0 \\ 1/5 & 0 & 1/2 & 1/3 & 0 & 0 & 0 \\ 1/5 & 0 & 0 & 1/3 & 1/4 & 0 & 0 \\ 1/5 & 0 & 0 & 0 & 1/4 & 0 & 0 \\ 1/5 & 0 & 0 & 1/3 & 0 & 1/2 & 1 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 0 \\ 1/5 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

和が1

PageRankの計算

- 重要度の初期値を定める
- 推移確率行列に従って重要度を伝播
- 収束した結果をPageRankとする



PageRankの計算

- 収束したときのPageRankをR(ベクトル)とすると

$$R = cMR$$

- これは良く見ると

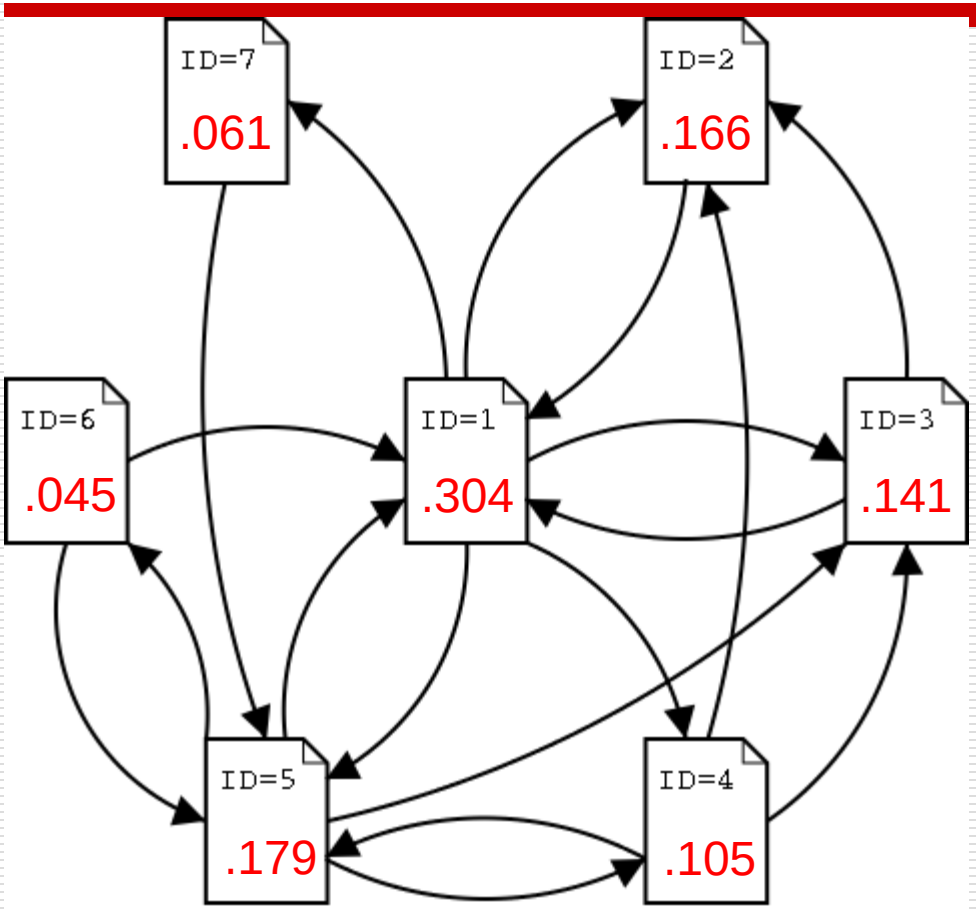
$$MR = \lambda R$$

において$\lambda = 1/c$としたもの

PageRankの計算

- 要するに、Mの固有値と固有ベクトルを求めればよい。
- Rは、絶対値最大の固有値に対する固有ベクトル（優固有ベクトル）

小規模な例に対するPageRank



R=

0.304	1
0.166	2
0.141	3
0.105	4
0.179	5
0.045	6
0.061	7

現実の問題への適用

1. 数学用語
2. 現実世界との相違
3. 数値計算の方法

数学用語(1)

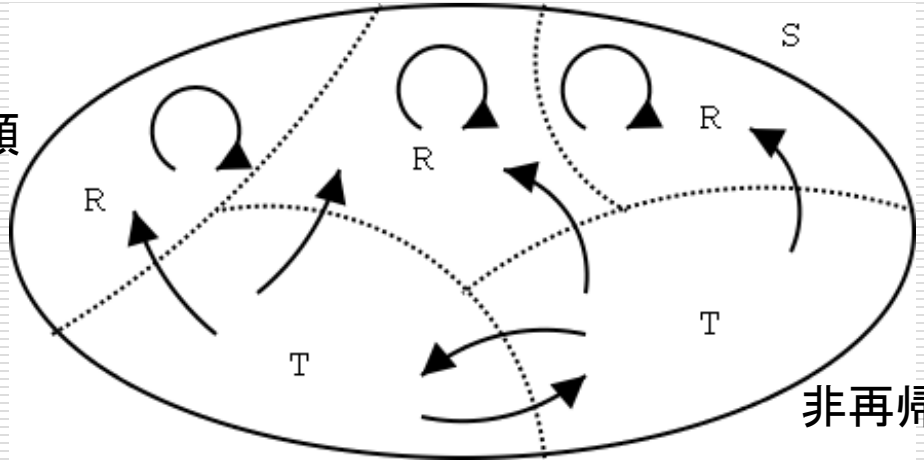
- PageRankはマルコフ過程と関連している
- PageRankが表す量
  - ランダムにリンクを辿って動くユーザが、一定の時間のうちにそれぞれのページを訪問する定常分布
  - ただし、推移確率行列が既約であることが条件

数学用語（2）

- 再帰
  - 状態iから出発していつかはiに戻る確率が1のとき、状態iは再帰的という
- 強連結
  - 任意の頂点から出発して、他の任意の頂点へ到達できること

数学用語（3）

- 再帰類
  - リンクをたどっていける範囲



- 既約
  - ただ一つの再帰類しかできないこと
  - 強連結なら既約

現実世界との相違(1):問題点

- 理論では既約(強連結)を仮定
- 実際にはこの仮定は成り立たない
  - リンクが出ていないページ
  - リンクされていないページ
- 推移確率行列が既約でないとどうなるか
  - 優固有ベクトルが複数存在
  - PageRankが一意に定まらない

現実世界との相違(2):解決策

- 推移確率行列を既約にする

$$M' = \mu M + (1-\mu)\left[\frac{1}{N}\right]$$

0.85

すべての要素が1/NであるN次正方行列

- 意味
  - ユーザは時々(確率1-μで)、全く無関係なページにジャンプする

数値計算の方法

- 大規模疎行列の計算
  - メモリの問題は出てこない
- 優固有ベクトルの計算
  - 固有値をすべて求めるのは計算量が多い
  - べき乗法で求める

PageRankの使い方

- PageRankの値
  - 検索質問(入力されるキーワード)に依存しない
- 検索質問に対する回答
  - PageRankでランキングされたページの中から、類似ページを探し出す処理が必要

試してみよう

- ページランクが分かるページ
  - http://pagerank.bookstudio.com/
- ページランクの計算
  - http://www.webworkshop.net/pagerank_calculator.php
  - http://www.markhorrell.com/seo/pagerank.asp

など

レポート課題

- PageRankを調べてみよ
  - pagerankを調べることができるサイトがある
  - それを使って、いくつかサイトのランクを調べる
  - 妥当性を論じる
- 適当に設定した小規模なグラフに対して、PageRankを求めてみよ
  - グラフの構造と値を見比べて考察
  - 妥当な値かどうか