

Google ページランク

使うのは 250 億元連立一次方程式

non rel QFT

SIMESABA

July 18, 2009

“ How Google Finds Your Needle in the Web ’s
Haystack, ” by David Austin
is the December 2006 Feature Column from the
American Mathematical Society,
<http://www.ams.org/featurecolumn/>.

- ページランク .
- Google という「身近なところ」にある役に立つ技術 .

- なぜページランクが必要か
- 算数レベルで分かる問題点
- 問題設定と数学
- 確率論の導入
- ベキ乗法
- 問題点とモデルの修正
- ページランクの計算：精度の問題
- まとめ

- 250 億種類の資料がある図書館 .

- 250 億種類の資料がある図書館 .
- 管理者はいない .

- 250 億種類の資料がある図書館 .
- 管理者はいない .
- いつでもだれでも好き勝手に資料を足せる .

- 250 億種類の資料がある図書館 .
- 管理者はいない .
- いつでもだれでも好き勝手に資料を足せる .
- 「自分にとっても役に立つ資料があるかも？」

- 250 億種類の資料がある図書館 .
- 管理者はいない .
- いつでもだれでも好き勝手に資料を足せる .
- 「自分にとっても役に立つ資料があるかも？」
- 「面倒だしさっさと必要な資料を見つけない」

- 250 億種類の資料がある図書館 .
- 管理者はいない .
- いつでもだれでも好き勝手に資料を足せる .
- 「自分にとっても役に立つ資料があるかも？」
- 「面倒だしさっさと必要な資料を見つけない」
- どうやって探そうか？

- インターネットはこんな感じ：
- たくさんの資料が未整理のまま，様々なフォーマットで置いてある．
- e.g. 文章，図，動画，パワーポイント，pdf．
- サーチエンジンの存在：何かいい探し方があるらしい．
- **Google** は100種ほどの判定基準を持つらしい：そのうちの1つが**ページランク**．

- ウェブからページを引き出す .
- 各文章の単語に索引を付けておく .
- あとで便利に使える形で情報をためておく .
- 誰かが適当なキーワードで検索をかける :
- サーチエンジンは同じキーワードを含むページを全て取ってきてくれる .

- 2006 年時点で 250 億のウェブページがある .
- ウェブ中の文章の 95% は 10,000 語程度 .
- 検索用のキーワードを含むページは山ほどあるはず .
- 各ページに重要度の高さに応じたランクをつけたい :
- 重要度の高いページほどリストのトップに並んでくれるような判定基準を探せ !

- 例：ある分野（たとえば激安店検索）で，役に立つサイトへのリンク集からなるページ．
- 管理者に信頼がおけるのなら，リンク先のページは役に立つ可能性が高い．
- **問題点**：

- 例：ある分野（たとえば激安店検索）で，役に立つサイトへのリンク集からなるページ．
- 管理者に信頼がおけるのなら，リンク先のページは役に立つ可能性が高い．
- **問題点**：
- すぐに内容が古くなってしまう．

- 例：ある分野（たとえば激安店検索）で，役に立つサイトへのリンク集からなるページ．
- 管理者に信頼がおけるのなら，リンク先のページは役に立つ可能性が高い．
- **問題点：**
- すぐに内容が古くなってしまう．
- 管理者が重要なページを見落としているかも．

- 例：ある分野（たとえば激安店検索）で，役に立つサイトへのリンク集からなるページ．
- 管理者に信頼がおけるのなら，リンク先のページは役に立つ可能性が高い．
- **問題点：**
- すぐに内容が古くなってしまう．
- 管理者が重要なページを見落としているかも．
- わざとリンクを外しているかもしれない（たとえば競合店）．

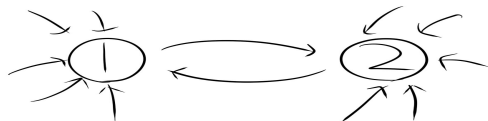
- 例：ある分野（たとえば激安店検索）で，役に立つサイトへのリンク集からなるページ．
- 管理者に信頼がおけるのなら，リンク先のページは役に立つ可能性が高い．
- **問題点**：
 - すぐに内容が古くなってしまう．
 - 管理者が重要なページを見落としているかも．
 - わざとリンクを外しているかもしれない（たとえば競合店）．
- **人が作った判定基準を信頼できる？**

- ページランクアルゴリズム：
- 人間の価値判断なしにウェブページの重要性を判断する．
- Google：「自分たちのサービスの価値は，調べたいことに対して偏見のない公正な情報を提供できることにある．」
- ページランクがその核心．
- ウェブ自体に各ページの重要性を教えてもらおう！

- ページランクは順序付け：数で表せるといいことがある **かもしれない** .
- 数に関する話なら数学がつかえるかもしれない , という安直な発想 .
- 必ずしも数学がつかえるわけでもなく , 使いやすいわけでもない .
- 今回はたまたまうまく使えた .
- 最近の金融のように数学に遊ばれていてはいけない .

算数レベルで分かる 問題点

- ウェブページにリンクを張るというのはどういうことか？
- → リンク先のページは重要だ．
- Google のページランク：月に一度の人気投票会．
- 仮定：あるページの重要度は，そこにリンクしているページ数と各ページの重要度で決まる．
- 注意：話が堂々めぐりしている．



Assumption

1. あるページのページランクは被リンク数と，リンク元のページランクで決まる．
2. ページランクは全リンク先に等分配する．

- 各ウェブページに名前をつける：一般に P とする．
- 各ページ P に対してその重要度を $I(P)$ とする：
ページランクと呼ぶ．
- ランキングを操作したがる人たちがいるので，実際のページランクは公開されていない．

- 全ページ数は $N = 2.5 \times 10^{10}$ で各ページの持ち点は 1 .
- ページ P_j が l_j 個リンクを張っているとする .
- ページ P_j はページ P_i にリンクを張っているとする .
- **仮定** : ページ P_j は自分の持ち点をリンク数で割って , リンク先に均等に配分する .
- **仮定** : 各ページのページランクは , リンクしてくれているページからもらえる点の単純な和 .

- ページ P_i にリンクを張っているページの集合を B_i として決定方程式を書く：

$$I(P_i) = \sum_{P_j \in B_i} \frac{1}{l_j} I(P_j) \quad (1)$$

$$= \sum_{j: P_j \text{ は } P_i \text{ にリンクを張っている}} \frac{1}{l_j} I(P_j). \quad (2)$$

- 問題勃発**：あるページのページランクを知りたいなら，そこにリンクを張っている全てのページのページランクを決定せよ．
- 自縄自縛．

- とりあえず数学的に問題を整理することから始める .
- ハイパーリンク行列 $H = [H_{i,j}]$ を次のように定義する :

$$H_{i,j} = \begin{cases} 1/l_j, & \text{if } P_j \in B_i, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

- $I = [I(P_i)]$ は N ($=250$ 億) 次元ベクトル：各要素はページランク．
- ページランク決定方程式を行列形式で書くと

$$I = HI. \quad (4)$$

- ページランクベクトル I は H の固有値 1 の固有ベクトル： H の定常ベクトルという．
- すっきり書けるので，考えやすくなることもある．

Theorem

確率行列は必ず定常ベクトルを持つ .

Remark

一般の文脈では I は H の不動点ともいう .

- ここでは概略だけ：

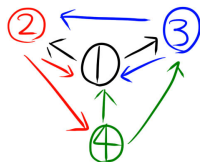
$$I^{k+1} = HI^k. \quad (5)$$

- I^0 を適当に決めて帰納的にベクトル列 (I^k) を定義する．
- $I = \lim_{k \rightarrow \infty} I^k$ が求めるベクトルになる．

- 決定方程式の解の存在 \leftrightarrow 自縄自縛 .
- 定常ベクトルの要素は全て非負か？
- 決定方程式の解の一意性 \leftrightarrow どれを選べばいい？
- 数値計算にのる？
- 250 億次 \leftrightarrow 計算には時間がかかる .
- 高速のアルゴリズムを実装するだけでどうにかなるのか？
- I^k の収束が遅いと現実的な時間内に計算が終わらない .
- 近似列の収束のスピードの保証 .

- 有理数係数の連立 1 次方程式なので，解も有理数．
- コンピュータでも厳密解が出せる．
- 別に厳密解はいらない：精度のいい近似解で十分．
- 数値計算，高速なアルゴリズム．
- モデルで出てくる行列の性質からベキ乗法が良いらしい．
- ベキ乗法を使うことを前提とした議論．

- 各ページは人（有向）リンクは得点の等分配．
- 初期得点分布を決める．
- 全得点は固定（されてほしい）．
- リンクに沿って点を等分配．
- 何度も繰り返して得点分布をみる．
- この終着点が欲しい得点分布 = 定常分布 = ページランク．

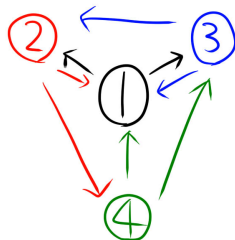


ページランク決定方程式:

$$\begin{cases} x_1 = 0x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3 + \frac{1}{2}x_4 \\ x_2 = \frac{1}{2}x_1 + 0x_2 + \frac{1}{2}x_3 + 0 \\ x_3 = \frac{1}{2}x_1 + 0x_2 + 0 + \frac{1}{2}x_4 \\ x_4 = 0x_1 + \frac{1}{2}x_2 + 0 + 0 \end{cases}$$

$$\Rightarrow x := \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, H = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 \end{bmatrix}, x = Hx$$

ゲームをやってみた

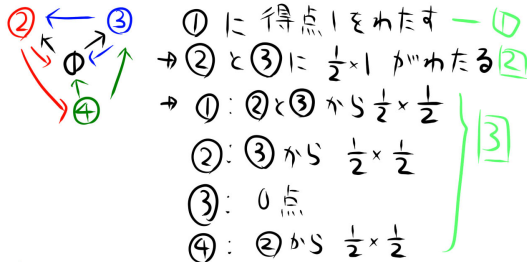


"ページランク"は

$$\begin{aligned} & 0 : 2 : 3 : 4 \\ & = 7 : 6 : 5 : 3 \\ & = 14 : 12 : 10 : 6 \end{aligned}$$

$$\left. \begin{array}{l} \textcircled{1} : 0 + \frac{1}{2}12 + \frac{1}{2}10 + \frac{1}{2}6 = 14 \\ \textcircled{2} : \frac{1}{2}14 + 0 + \frac{1}{2}10 + 0 = 12 \\ \textcircled{3} : \frac{1}{2}14 + 0 + 0 + \frac{1}{2}6 = 10 \\ \textcircled{4} : 0 + \frac{1}{2}12 + 0 + 0 = 6 \end{array} \right\}$$

ゲームをやってみた



時点 ① → ② → ③

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix} =: \frac{1}{2} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \quad \frac{1}{2^2} \begin{bmatrix} 0 + \text{red} + \text{blue} + 0 \\ 0 + \text{red} + \text{blue} + 0 \\ 0 + \text{red} + \text{blue} + 0 \\ 0 + \text{red} + 0 + 0 \end{bmatrix} =: \frac{1}{2^2} \begin{bmatrix} 2 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

ゲームをやってみた



時点③での得点分布

$$\frac{1}{2^2} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

足す1

次の分配:

$$\frac{1}{2^3} \begin{bmatrix} 0 + 1 + 0 + 1 \\ 2 + 0 + 0 + 0 \\ 2 + 0 + 0 + 1 \\ 0 + 1 + 0 + 0 \end{bmatrix} = \frac{1}{2^3} \begin{bmatrix} 2 \\ 2 \\ 3 \\ 1 \end{bmatrix}$$

次:

$$\frac{1}{2^4} \begin{bmatrix} 0 + 2 + 3 + 1 \\ 2 + 0 + 3 + 0 \\ 2 + 0 + 0 + 1 \\ 0 + 2 + 0 + 0 \end{bmatrix} = \frac{1}{2^4} \begin{bmatrix} 6 \\ 5 \\ 3 \\ 2 \end{bmatrix}$$

次:

$$\frac{1}{2^5} \begin{bmatrix} 0 + 5 + 3 + 2 \\ 6 + 0 + 3 + 0 \\ 6 + 0 + 0 + 2 \\ 0 + 5 + 0 + 0 \end{bmatrix} = \frac{1}{2^5} \begin{bmatrix} 10 \\ 9 \\ 8 \\ 5 \end{bmatrix}$$

ゲームをやってみた



前時点での得点分布

$$\frac{1}{2^5} \begin{bmatrix} 10 \\ 9 \\ 8 \\ 5 \end{bmatrix}$$

足す1

次の分配:

$$\frac{1}{2^6} \begin{bmatrix} 0 + 9 + 8 + 5 \\ 10 + 0 + 8 + 0 \\ 10 + 0 + 0 + 5 \\ 0 + 9 + 0 + 0 \end{bmatrix} = \frac{1}{2^6} \begin{bmatrix} 22 \\ 18 \\ 15 \\ 9 \end{bmatrix}$$

3でわる

次

$$\frac{1}{2^7} \begin{bmatrix} 0 & 18 & 15 & 9 \\ 22 & 0 & 15 & 0 \\ 22 & 0 & 0 & 9 \\ 0 & 18 & 0 & 0 \end{bmatrix} = \frac{1}{2^7} \begin{bmatrix} 42 \\ 37 \\ 31 \\ 18 \end{bmatrix}$$

6でわる

次

$$\frac{1}{2^8} \begin{bmatrix} 0 & 37 & 31 & 18 \\ 42 & 0 & 31 & 0 \\ 42 & 0 & 0 & 18 \\ 0 & 37 & 0 & 0 \end{bmatrix} = \frac{1}{2^8} \begin{bmatrix} 86 \\ 73 \\ 60 \\ 37 \end{bmatrix}$$

12でわる



- **Dangling node**（ぶらさがり結節）の存在．
- 渡す人がいない：今の設定では，この人は点を捨てなければならない．
- 定常分布は0しかない：無意味．
- どのページにも点を分配しない＝ページランクに寄与しない．
- **全てのページに均等に持ち点を分配する**と解釈する．

① → ②

$$\begin{cases} a, b \geq 0 \\ a+b=1 \end{cases}$$

初期分布: $I_1^0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $I_2^0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $I_3^0 = \begin{bmatrix} a \\ b \end{bmatrix}$

$$I_1^0: \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$I_2^0: \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$I_3^0: \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ a \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

注: $I = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ は "定常分布."

$$\textcircled{1} \rightarrow \textcircled{2} \xrightarrow{\text{修正}} \textcircled{1} \rightleftharpoons \textcircled{2} \curvearrowright$$

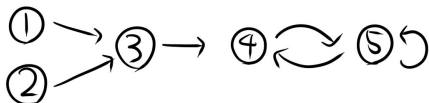
定常分布 $\textcircled{1} : \textcircled{2} = 1 : 2$
"②の方が2倍大事"

実際に計算:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \rightarrow \begin{bmatrix} 0+1 \\ 1+1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

- Dangling nodes が出てきたら , そこから全てのページにリンクを張れ .
- 点を捨てるぐらいなら , 皆に分配する .
- 全得点の保存 .

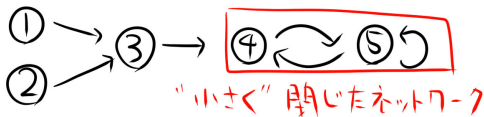
- リンクが張られていれば，そのページは重要なページ．
- ページランクは厳密に正 (> 0) になってもらいたい．
- **Importance sink** の存在．
- 小規模の**閉じた**ネットワーク．



☆ Dangling nodes ない

☆ 定常分布 (の1つ): $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 2 \end{bmatrix}$

☆ ③に張られたリンクは2つ
重要度はあるはず
⇔ でもページランク 0.



☆問題

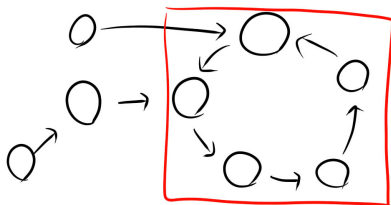
①, ②, ③ のもっている得点が全て

 に流れこんだまま戻らない

→ ① ② ③ は “最終的に餓え死ぬ。”

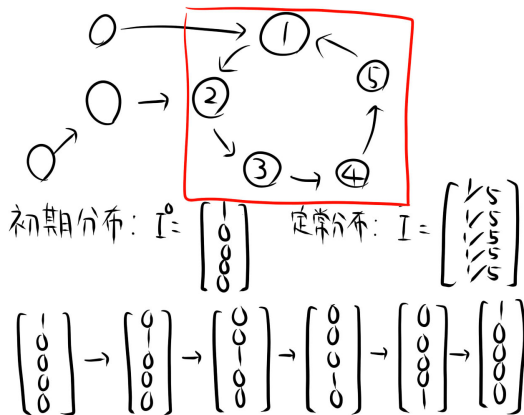
- Importance sinks が出てくると，点が分配されてほしい所に点が行き渡らない．
- Importance sinks は現実には存在する．
- どうごまかすか．
- 嫌ならページランクというアイデアを捨てる．

- ベキ乗法がつかえないということ : ベキ乗法固有の問題 .
- 巡回的なネットワークが問題 .
- 点の持ち回り : 収束しない .
- 数学的には「ハイパーリンク行列の第 2 固有値 $|\lambda_2| < 1$ となるか」 .



Importance sink
&
"巡回ネットワーク"

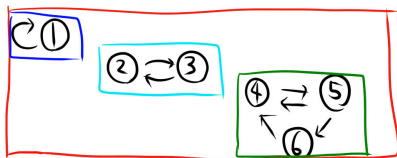
ゲームをやってみた



- 巡回ネットワークが出てくるとベキ乗法は使えない．
- 別の近似計算法を使う？
- 計算スピードの問題．

- 孤立したネットワークがたくさんあるときが問題 .
- 現実にかかる .
- どの定常分布が“いい”のか , 判断基準がない .

ゲームをやってみた

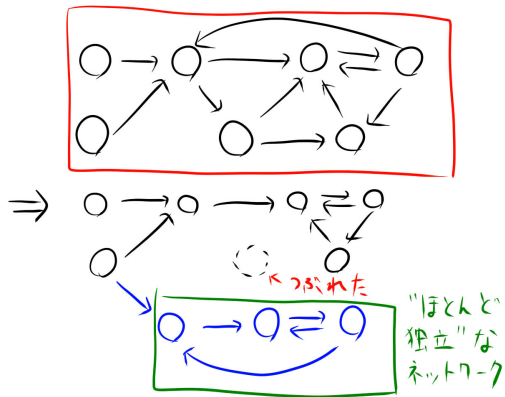


定常分布

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 2/5 \\ 2/5 \\ 1/5 \end{bmatrix}, \frac{1}{30} \begin{bmatrix} 10 \\ 5 \\ 5 \\ 4 \\ 4 \\ 2 \end{bmatrix}, \frac{1}{1000} \begin{bmatrix} 10 \\ 15 \\ 15 \\ 384 \\ 384 \\ 192 \end{bmatrix}$$

- 前の問題を言い換えてみた．
- 数値計算結果の“引き継ぎ”ができない．
- ウェブ上の情報は新鮮さが命：頻繁に計算しなおしたい．
- 新規計算時，ネットワーク構造は変わっているはず．
- しかし前とそれほど大きくは変わらないはず．
- 前の得点分布がよい近似になるのでは？
- この目論見が潰える．

ゲームをやってみた



依存あり：玉ころかし

$t=0$

→ ①

同じ"カ"
で"
つづく

→

$t>0$

○-----○

場
↓
所
も
違
う

→ ②

○-----○

依存なし：コーヒーにクリーム



→



- 盲目的に数値計算に頼るのは危険．
- シミュレーションなども何のシミュレーションをしているのか，本当にシミュレートできているのかきちんと考えないと意外と危険．
- 数学に「何とかしろ」といわれても困る．
- ある数学的設定下でそれなりの結果は出せる（ことはある）．
- ただ，その条件が外れた時にどうなるかは分からない：例：直線と曲線．
- “美しい一般論”から外れたゲテモノ．

- せっかく高速のアルゴリズムを用意しても，そのアルゴリズムを 10^{10000} 回も回さなければならなかったりしたら問題．
- 遅くとも回転数が少なくて済む方がいい？
- 回転数に初期分布依存があったりしたらどうする？
- ソフトウェアのテストなどでも同じだが，「こんな使い方はありえない」という使い方をされるかもしれない：きちんとチェック．

- Dangling nodes : 何とかなりそう？

- Dangling nodes : 何とかなりそう？
- Importance sinks : ネットワーク構造 : いじれない？

- Dangling nodes : 何とかなりそう？
- Importance sinks : ネットワーク構造 : いじれない？
- 定常分布が非一意 : どれを選べばいい？

- Dangling nodes : 何とかなりそう？
- Importance sinks : ネットワーク構造 : いじれない？
- 定常分布が非一意 : どれを選べばいい？
- 定常分布が初期分布に依存 : どの初期分布を選べばいい？

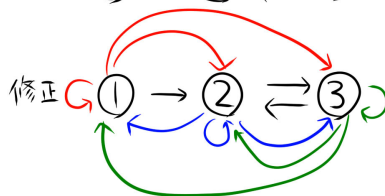
- Dangling nodes : 何とかなりそう？
- Importance sinks : ネットワーク構造 : いじれない？
- 定常分布が非一意 : どれを選べばいい？
- 定常分布が初期分布に依存 : どの初期分布を選べばいい？
- 同じ初期分布に対して定常分布が一意か？

- Dangling nodes : 何とかなりそう？
- Importance sinks : ネットワーク構造 : いじれない？
- 定常分布が非一意 : どれを選べばいい？
- 定常分布が初期分布に依存 : どの初期分布を選べばいい？
- 同じ初期分布に対して定常分布が一意的か？
- 収束のスピード : アルゴリズムのスピードとは無関係 .

- Dangling nodes : 何とかなりそう？
- Importance sinks : ネットワーク構造 : いじれない？
- 定常分布が非一意 : どれを選べばいい？
- 定常分布が初期分布に依存 : どの初期分布を選べばいい？
- 同じ初期分布に対して定常分布が一意的か？
- 収束のスピード : アルゴリズムのスピードとは無関係 .
- そもそも定常分布は存在するのか？

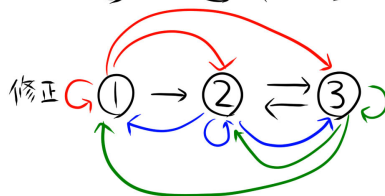
- 無理やり埒をあける．
- 今までの例を見てみると，**孤立点**が問題を起こす．
- Dangling nodes, importance sinks, 孤立 / ほぼ独立した小規模ネットワーク，
- モデルを人為的に変更：孤立させない．
- 元の構造をなるべく保ちたい．
- どうやるのか：**二重構造**．
- 欲しい性質を**ほぼ**すべて持つ：一意性存在，初期分布への非依存，定常分布への収束．

もと $\textcircled{1} \rightarrow \textcircled{2} \rightleftharpoons \textcircled{3}$

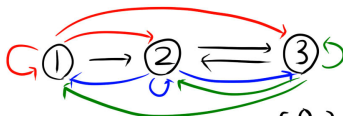


もとの構造 (黒線) には (例えば) 8割
色つきは残りの2割を等分配

もと $\textcircled{1} \rightarrow \textcircled{2} \rightleftharpoons \textcircled{3}$



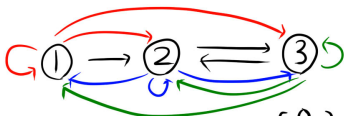
もとの構造 (黒線) には (例えば) 8割
色つきは残りの2割を等分配



初期分布: $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ (定) $\frac{\text{修正}}{\text{正規化}} = \begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$ 後 $= \frac{1}{945} \begin{bmatrix} 63 \\ 455 \\ 427 \end{bmatrix}$

$$\text{①} \begin{bmatrix} 0 + 1 \times \frac{1}{5} \times \frac{1}{3} + 0 + 0 + 0 + 0 \\ 1 \times \frac{4}{5} + 1 \times \frac{1}{5} \times \frac{1}{3} + 0 + 0 + 0 + 0 \\ 0 + 1 \times \frac{1}{5} \times \frac{1}{3} + 0 + 0 + 0 + 0 \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 1 \\ 13 \\ 1 \end{bmatrix}$$

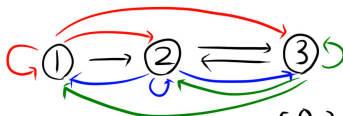
$$\text{②} \frac{1}{15} \begin{bmatrix} 0 + 1 \times \frac{1}{5} \times \frac{1}{3} + 0 + 13 \times \frac{1}{5} \times \frac{1}{3} + 0 + 1 \times \frac{1}{15} \\ 1 \times \frac{4}{5} + 1 \times \frac{1}{5} \times \frac{1}{3} + 0 + 13 \times \frac{1}{5} \times \frac{1}{3} + 1 \times \frac{4}{5} + 1 \times \frac{1}{15} \\ 0 + 1 \times \frac{1}{5} \times \frac{1}{3} + 13 \times \frac{4}{5} + 13 \times \frac{1}{5} \times \frac{1}{3} + 0 + 1 \times \frac{1}{15} \end{bmatrix} = \frac{1}{15^2} \begin{bmatrix} 15 \\ 39 \\ 171 \end{bmatrix}$$



初期分布: $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ (定) $\frac{\text{修正}}{\text{正規化}} = \begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$ $\text{後} = \frac{1}{945} \begin{bmatrix} 63 \\ 455 \\ 427 \end{bmatrix}$

[3] $\frac{1}{15^2} \begin{bmatrix} 0 & 15 \times \frac{1}{15} & 0 & 39 \times \frac{1}{15} & 0 & 171 \times \frac{1}{15} \\ 15 \times \frac{4}{5} & 15 \times \frac{1}{15} & 0 & 39 \times \frac{1}{15} & 171 \times \frac{4}{5} & 171 \times \frac{1}{15} \\ 0 & 15 \times \frac{1}{15} & 39 \times \frac{4}{5} & 39 \times \frac{1}{15} & 0 & 171 \times \frac{1}{15} \end{bmatrix} = \frac{1}{15^3} \begin{bmatrix} 225 \\ 2457 \\ 693 \end{bmatrix}$

[4] $\frac{1}{15^3} \begin{bmatrix} 225 \times \frac{1}{15} & 2457 \times \frac{1}{15} & 693 \times \frac{1}{15} \\ 225 \times \frac{13}{15} & 2457 \times \frac{1}{15} & 693 \times \frac{13}{15} \\ 225 \times \frac{1}{15} & 2457 \times \frac{13}{15} & 693 \times \frac{1}{15} \end{bmatrix} = \frac{1}{15^4} \begin{bmatrix} 3315 \\ 14391 \\ 32859 \end{bmatrix}$

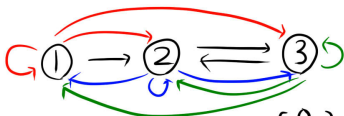


初期分布: $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ (定) 修正 $= \begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$ 後 $= \frac{1}{945} \begin{bmatrix} 63 \\ 455 \\ 427 \end{bmatrix}$

⑤ $\frac{1}{156} \begin{bmatrix} 3375 & 14391 & 32859 \\ 3375 \times 13 & 14391 & 32859 \times 13 \\ 3375 & 14391 \times 13 & 32859 \end{bmatrix} = \frac{1}{156} \begin{bmatrix} 50625 \\ 485433 \\ 223317 \end{bmatrix}$

⑤ 例外的なところ:

- ① 加値をもつ
- ② $> ③$
- 50回くさい計算するといい値になる



初期分布: $\begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$ (定) $\frac{\text{修正}}{\text{正規化}} = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$ $\vec{x} = \frac{1}{945} \begin{bmatrix} 63 \\ 455 \\ 427 \end{bmatrix}$

$$\textcircled{1} \begin{bmatrix} 0 + 0 + 0 + \frac{1}{2} \times \frac{1}{5} \times \frac{1}{3} + 0 + \frac{1}{2} \times \frac{1}{5} \times \frac{1}{3} \\ 0 + 0 + 0 + \frac{1}{2} \times \frac{1}{5} \times \frac{1}{3} + \frac{1}{2} \times \frac{4}{5} + \frac{1}{2} \times \frac{1}{5} \times \frac{1}{3} \\ 0 + 0 + \frac{1}{2} \times \frac{4}{5} + \frac{1}{2} \times \frac{1}{5} \times \frac{1}{3} + 0 + \frac{1}{2} \times \frac{1}{5} \times \frac{1}{3} \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 1 \\ 7 \\ 7 \end{bmatrix}$$

$$\textcircled{2} \frac{1}{15} \begin{bmatrix} 0 + 1 \times \frac{1}{5} \times \frac{1}{3} + 0 + 7 \times \frac{1}{5} \times \frac{1}{3} + 0 + 7 \times \frac{1}{5} \\ 1 \times \frac{4}{5} + 1 \times \frac{1}{5} \times \frac{1}{3} + 0 + 7 \times \frac{1}{5} \times \frac{1}{3} + 7 \times \frac{4}{5} + 7 \times \frac{1}{5} \\ 0 + 1 \times \frac{1}{5} \times \frac{1}{3} + 7 \times \frac{4}{5} + 7 \times \frac{1}{5} \times \frac{1}{3} + 0 + 7 \times \frac{1}{5} \end{bmatrix} = \frac{1}{15^2} \begin{bmatrix} 15 \\ 111 \\ 99 \end{bmatrix}$$

収束しなかった構造でも収束するようになる

$$t \quad \textcircled{1} \rightleftharpoons \textcircled{2} \quad \boxed{\text{E}}: \frac{1}{2} [1]$$

不修正 $G \begin{matrix} \textcircled{1} \\ \rightleftharpoons \\ \textcircled{2} \end{matrix} \textcircled{5}$ 定 $\frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

★ ①の構造(黒線)に8割

☆裏構造(色付き)に2割

初 $I^0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\boxed{II} \begin{bmatrix} 0 & 1 \times \frac{1}{5} \times \frac{1}{2} & 0 & 0 \\ 1 \times \frac{4}{5} & 1 \times \frac{1}{5} \times \frac{1}{2} & 0 & 0 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} 1 \\ 4 \\ 9 \end{bmatrix}$$

$$\boxed{2} \cdot \frac{1}{10} \begin{bmatrix} 1 \times \frac{1}{10} & 9 \times \frac{9}{10} \\ 1 \times \frac{9}{10} & 9 \times \frac{1}{10} \end{bmatrix} = \frac{1}{10^2} \begin{bmatrix} 82 & 82 \\ 82 & 18 \end{bmatrix} \quad \begin{array}{l} \text{よって} \\ \text{"シ" - "1" - "7" - "4"} \end{array}$$

- どれだけ小さい割合でも得点のリークさえあれば，定常分布の一意性存在，初期分布への非依存，定常分布への収束の全てが成り立つ．

- どれだけ小さい割合でも得点のリークさえあれば，定常分布の一意性存在，初期分布への非依存，定常分布への収束の全てが成り立つ．
- リークが少ないほど，もとの構造がはっきり出る．

- どれだけ小さい割合でも得点のリークさえあれば，定常分布の一意性存在，初期分布への非依存，定常分布への収束の全てが成り立つ．
- リークが少ないほど，もとの構造がはっきり出る．
- リークが小さすぎると，なかなか収束しない．

① \rightleftharpoons ② 固: $\frac{1}{2}[1]$

不修正 $G \textcircled{1} \rightleftharpoons \textcircled{2} G$ 定: $\frac{1}{2} [1]$

★ 土の構造 (黒線) に $(1-10^{-10})$

$$\star \text{裏構造 (色) } \pm = 10^{-10} \quad \left(= \frac{1}{4} \begin{bmatrix} 4 - 4 \times 10^{-10} + 2 \times 10^{-20} \\ 4 \times 10^{-10} - 2 \times 10^{-20} \end{bmatrix} \right)$$

初 $I^0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\boxed{11} \begin{bmatrix} 0 & 1 \times 10^{-10} \times \frac{1}{2} \\ 1 \times (1 - 10^{-10}) & 1 \times 10^{-10} \times \frac{1}{2} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 10^{-10} \\ 2 - 10^{-10} \end{bmatrix}$$

$$\boxed{2} \quad \frac{1}{2} \begin{bmatrix} 0 & \cancel{10^{-10}} \cdot \cancel{10^{-10}} \cdot \frac{1}{2} & (2 \cdot \cancel{10^{-10}}) \cdot (1 \cdot \cancel{10^{-10}}) & (2 \cdot \cancel{10^{-10}}) \cdot \cancel{10^{-10}} \cdot \frac{1}{2} \\ \cancel{10^{-10}} \cdot (1 \cdot \cancel{10^{-10}}) & \cancel{10^{-10}} \cdot \cancel{10^{-10}} \cdot \frac{1}{2} & 0 & (2 \cdot \cancel{10^{-10}}) \cdot \cancel{10^{-10}} \cdot \frac{1}{2} \end{bmatrix}$$

- リーク率を $\alpha \in [0, 1]$ とする .
- 収束の**最悪**のスピードは $(1 - \alpha)$ で制御される .
- 第 k ステップでの“誤差”は $(1 - \alpha)^k$ でおさえられる .
- 実際には $\alpha = 0.15$ でやっている .

$$(1 - 10^{-10})^{50} = 0.9999999995, \quad (6)$$

$$(1 - 10^{-10})^{10^{10}} = 0.36787941, \quad (7)$$

$$0.99^{50} = 0.60500606, \quad (8)$$

$$0.9^{50} = 0.00515377, \quad (9)$$

$$0.85^{50} = 0.00029576. \quad (10)$$

- リーク率を適切に選べば，定常分布の一意性存在，初期分布への非依存，定常分布への収束，収束のスピード制御まで統制できる．
- その代わりにもとのネットワーク構造が歪められる．

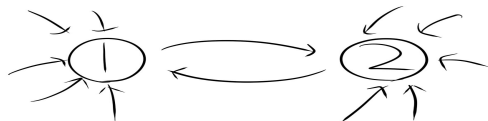
Remark

今やっているの理学ではなく工学．
欲しいのはいろいろな制約条件のもとでの精度のいい近似解．

問題設定と数学

250億元連立一次方程式

- **ウェブページにリンクを張る** というのはどういうことか？
- → **リンク先のページは重要だ** .
- Google のページランク：月に一度の人気投票会 .
- **仮定**：あるページの重要度は，そこにリンクしているページ数と各ページの重要度で決まる .
- 注意：話が堂々めぐりしている .



- 各ウェブページに名前をつける：一般に P とする．
- 各ページ P に対してその重要度を $I(P)$ とする：
ページランクと呼ぶ．
- ランキングを操作したがる人たちがいるので，実際のページランクは公開されていない．

- 全ページ数は $N = 2.5 \times 10^{10}$ で各ページの持ち点は 1 .
- ページ P_j が l_j 個リンクを張っているとする .
- ページ P_j はページ P_i にリンクを張っているとする .
- **仮定** : ページ P_j は自分の持ち点をリンク数で割って , リンク先に均等に配分する .
- **仮定** : 各ページのページランクは , リンクしてくれているページからもらえる点の単純な和 .

- ページ P_i にリンクを張っているページの集合を B_i として決定方程式を書く：

$$I(P_i) = \sum_{P_j \in B_i} \frac{1}{l_j} I(P_j) \quad (11)$$

$$= \sum_{j: P_j \text{ は } P_i \text{ にリンクを張っている}} \frac{1}{l_j} I(P_j). \quad (12)$$

- 問題勃発**：あるページのページランクを知りたいなら，そこにリンクを張っている全てのページのページランクを決定せよ．
- 自縄自縛．

- とりあえず数学的に問題を整理することから始める .
- ハイパーリンク行列 $H = [H_{i,j}]$ を次のように定義する :

$$H_{i,j} = \begin{cases} 1/l_j, & \text{if } P_j \in B_i, \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

- 全要素（エントリー）が非負：

$$H_{i,j} \geq 0, \quad \forall i, j. \quad (14)$$

- 対応するページが1つでもリンクを張っていれば，対応する列の和が1：

$$\sum_{j=1}^N H_{i,j} = 1. \quad (15)$$

Definition

全ての要素が非負で、**全ての**列についてその全要素の和が1になる行列を**確率行列**という。

Remark

一般に、いま定義したハイパーリンク行列は確率行列**ではない**。

- 確率行列は今回の主力兵器 .
- これから扱う行列が確率行列になるようにモデルを修正していく : 最終的に考えるのは H ではない .
- 数学がつかえるように **どうごまかすか** が焦点と言ってもいい .

Remark

今やっているのは理学でなく工学 : すべきは**最良**の成果を求めることではなく , **様々な拘束下**での**最適**な成果を**いい近似**で求めること .

- $I = [I(P_i)]$ は N ($=250$ 億) 次元ベクトル：各要素はページランク．
- ページランク決定方程式を行列形式で書くと

$$I = HI. \quad (16)$$

- ページランクベクトル I は H の固有値 1 の固有ベクトル： H の定常ベクトルという．
- すっきり書けるので，考えやすくなることもある．

Theorem

確率行列は必ず定常ベクトルを持つ .

Remark

一般の文脈では I は H の不動点ともいう .

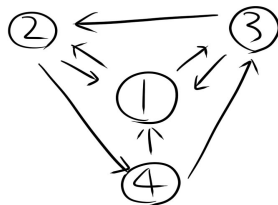
- ここでは概略だけ：

$$I^{k+1} = HI^k. \quad (17)$$

- I^0 を適当に決めて帰納的にベクトル列 (I^k) を定義する．
- $I = \lim_{k \rightarrow \infty} I^k$ が求めるベクトルになる．

- 決定方程式の解の存在 \leftrightarrow 自縄自縛 .
- 定常ベクトルの要素は全て非負か？
- 決定方程式の解の一意性 \leftrightarrow どれを選べばいい？
- 数値計算にのる？
- 250 億次 \leftrightarrow 計算には時間がかかる .
- 高速のアルゴリズムを実装するだけでどうにかなるのか？
- I^k の収束が遅いと現実的な時間内に計算が終わらない .
- 近似列の収束のスピードの保証 .

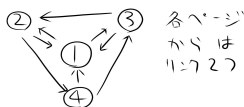
行列とベクトルの積：持ち点分配ゲーム



各ページ
からは
リンク2つ

$$\begin{cases} I(1) = 0I(1) + \frac{1}{2}I(2) + \frac{1}{2}I(3) + \frac{1}{2}I(4) \\ I(2) = \frac{1}{2}I(1) + 0I(2) + \frac{1}{2}I(3) + 0I(4) \\ I(3) = \frac{1}{2}I(1) + 0I(2) + 0I(3) + \frac{1}{2}I(4) \\ I(4) = 0I(1) + \frac{1}{2}I(2) + 0I(3) + 0I(4) \end{cases} \quad (18)$$

$$\Leftrightarrow \begin{bmatrix} I(1) \\ I(2) \\ I(3) \\ I(4) \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} I(1) \\ I(2) \\ I(3) \\ I(4) \end{bmatrix} \quad (19)$$

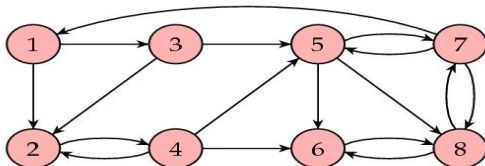


- 全体の得点を 1 として各ページに点を割り振る（初期ベクトル）。
- リンク先に今の自分の持ち点を平等に割り振る（行列をかける）。
- 得点分布の推移を見守る（ I^k の動きを追う）。

$$I^0 := \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow I^1 = \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \\ 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \rightarrow I^2 = \frac{1}{4} \begin{bmatrix} 2 \\ 1 \\ 0 \\ 1 \end{bmatrix} \rightarrow I^3 = \frac{1}{8} \begin{bmatrix} 2 \\ 2 \\ 3 \\ 1 \end{bmatrix} . \quad (20)$$

$$I^4 := \frac{1}{16} \begin{bmatrix} 6 \\ 5 \\ 3 \\ 2 \end{bmatrix} \rightarrow I^5 = \frac{1}{32} \begin{bmatrix} 10 \\ 9 \\ 8 \\ 5 \end{bmatrix} \rightarrow I^6 = \frac{1}{64} \begin{bmatrix} 22 \\ 18 \\ 15 \\ 9 \end{bmatrix} \rightarrow I^7 = \frac{1}{128} \begin{bmatrix} 42 \\ 37 \\ 31 \\ 18 \end{bmatrix}. \quad (21)$$

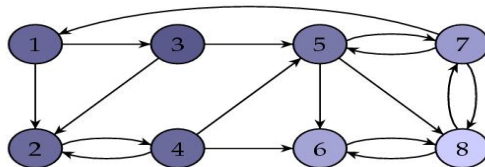
- 順序はだいたい分かる (7:6:5:3) .
- これを無限に続けると点の分配が止まる : 定常ベクトルは定常分布 .
- 本当に分配が止まるのか ?
- 止まる先は初期ベクトルに依存せずに決まるのか ?
- (適当に運動方程式や拡散方程式 .)



$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{bmatrix}, \quad I = \begin{bmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ 0.2950 \end{bmatrix}$$

(22)

グラデーション付きのグラフ



ページランクベクトル I の計算

- 正方行列の固有ベクトルを求める方法はたくさんある．
- H の特性に注目．
- 平均すると，各ウェブページはだいたい10個くらいリンクを張っている．
- H では： H の1つの列を見ると，ある10個以外全ての要素は0．
- Power method (ベキ乗法？) がいい．

- I^0 を適当に選んで，ベクトル列 (I^k) を次のように定義する：

$$I^{k+1} = HI^k. \quad (23)$$

- ベキ乗法の基礎になる定理：

Theorem

適当な確率行列 H に対して適当な初期ベクトルを取れば，ベクトル列 (I^k) は定常ベクトルに収束する．

- ベクトル列 (I^k) はいつでも収束するのか？
- I に収束するとして, I は初期条件 I^0 によらないベクトルになるのか？
- ページランクベクトルが欲しい情報を含んでいるのか？

Remark

今のままでは3つとも **NO** .
モデルの構成を改良していく .



対応する決定方程式と行列：

$$\begin{cases} I(1) &= 0I(1) + 0I(2), \\ I(2) &= 1I(1) + 0I(2). \end{cases} \quad (24)$$

$$\Leftrightarrow \begin{bmatrix} I(1) \\ I(2) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} I(1) \\ I(2) \end{bmatrix}. \quad (25)$$

$$I^0 := \begin{bmatrix} 1 \\ 0 \end{bmatrix}, I^1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, I^2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, I^3 = I = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (26)$$

- 両方のページランクが0：情報が得られない．
- 問題： P_2 がリンクを張っていない $\rightarrow P_1$ からの情報の連鎖が消える．
- 情報が外に出ていかないページを**ぶらさがり結節** (**dangling node**) という．
- 実際にはこういうページはたくさんある．

ハイパーリンク行列 H の 確率論的解釈

- 馴染みのないトピックについて調べたい .
- 引っかかったページから適当にリンクをたどっていく .
- しばらくやっていると , よくリンクされているページがいくつか見つかる .
- "All roads lead to Rome" : これらのページは重要そうだ .
- これを確率論でモデル化する .

- ネットサーフィン：リンクをたどって適当に動き回る．
- 面白そうなページにはブックマークを付けておいて，あとでじっくり見る．
- 各ページには1秒だけいることにして，1秒たったら適当にリンクをたどっていく．
- ページ P_j には l_j 個のリンクがあって，とくにページ P_i へのリンクがある．
- 次に P_i に行く確率は $1/l_j$ ．

- ページ P_j にいる “平均（見込み）時間” を T_j とする：
- ページ P_j からページ P_i に行くとき，ページ P_i で過ごす “時間” は T_j/l_j ．
- P_i で過ごす時間は P_i にリンクを張っているページ全体からの和になるから，

$$T_i = \sum_{j: P_j \text{ は } P_i \text{ にリンクを張っている}} \frac{T_j}{l_j}. \quad (27)$$

- これはページランク決定方程式と同じ．

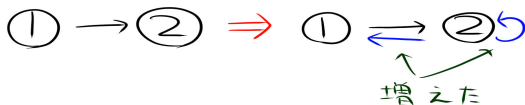
- **dangling node** (リンクのないページ) にぶつかる .
- ここでは次のようにモデル化する .
- 次のページを適当に選ぶ : dangling node はすべてのページにリンクを張る .
- 決定方程式を「修正」: dangling node の列の各要素が 0 から $1/N$ になる .
- 修正行列を A : **確率**行列 S が登場 :

$$S = H + A. \quad (28)$$

- S に対して決定方程式を考えていく .

- リンクが張られていない .
- ページランクに全く寄与しない = どのページにも点を分配しない .
- これをどのページにも均等に持ち点を分配すると解釈しなおす .

- さっき出したまずい例は次のように変わる．
- P_2 ランクは P_1 の 2 倍：何となく直感と合致．



$$H = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/2 \end{bmatrix}, \quad S = \begin{bmatrix} 0 & 1/2 \\ 1 & 1/2 \end{bmatrix}, \quad (29)$$

$$I = \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}. \quad (30)$$

ベキ乗法の基礎

Definition

A : 線型空間 V 上の線型作用素, $\lambda \in \mathbb{C}$, $v \in V \setminus \{0\}$. λ を A の固有値といい, v を固有値 λ に属する固有ベクトルという:

$$Av = \lambda v. \quad (31)$$

Example (かなりいいかげん)

$$\left(-i \frac{d}{dx}\right) e^{ikx} = k e^{ikx}.$$

実数 k は, 微分作用素 (運動量演算子) の “固有値” で, e^{ikx} はその “固有ベクトル”.

Remark

V が有限次元ならば、必ず固有値は存在する。
 $\dim V = N$ とすると、固有値は重複を込めて N 個存在する。
(代数学の基本定理。)

- ベキ乗法は大きさ最大の固有値に対する固有ベクトルを求めるための方法．
- (簡単にするための) 仮定: S の固有値は

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \quad (32)$$

- 固有ベクトルは自然基底 $\{v_i\}$:

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, v_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \quad (33)$$

- I^0 を初期ベクトルとしてこれを展開：

$$I^0 = c_1 v_1 + \sum_{i=2}^n c_i v_i, \quad (34)$$

$$I^{\mathbf{1}} = SI^0 = c_1 v_1 + \sum_{i=2}^n c_i \lambda_i^{\mathbf{1}} v_i, \quad (35)$$

$$\vdots \quad (36)$$

$$I^{\mathbf{k}} = SI^{k-1} = c_1 v_1 + \sum_{i=2}^n c_i \lambda_i^{\mathbf{k}} v_i. \quad (37)$$

- $j \geq 2$ のとき . $|\lambda_j| < 1$ なので ,

$$\lambda_j^k \rightarrow 0 \quad \text{if } j \geq 2 \quad (38)$$

$$I^k \rightarrow I = c_1 v_1. \quad (39)$$

- 極限で得られる I は固有値 1 に属する固有ベクトル (定常ベクトル) .

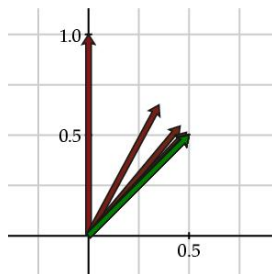
Remark

収束のスピードを決めるのは $|\lambda_2|$.

$|\lambda_2|$ が小さいほど , 収束のスピードが増す .

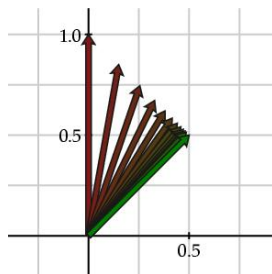
- 行列と固有値：

$$S = \begin{bmatrix} 0.65 & 0.35 \\ 0.35 & 0.65 \end{bmatrix}, \quad \lambda_1 = 1, \quad \lambda_2 = 0.3. \quad (40)$$



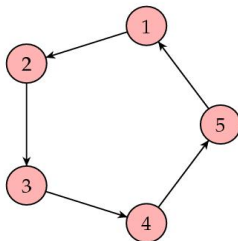
- 行列と固有値：

$$S = \begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix}, \quad \lambda_1 = 1, \quad \lambda_2 = 0.7. \quad (41)$$

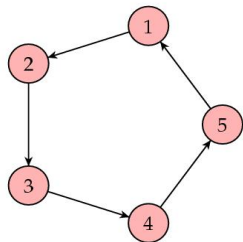


さらなる問題点

- 第2固有値が $|\lambda_2| < 1$ となるものばかりではない。



$$S = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \quad (42)$$



- ベクトル列が収束しない（ことがある）:

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} .$$

- 定常ベクトルは

$$I = \begin{bmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{bmatrix}. \quad (43)$$

- ページランクを一種の得点とみなす：得点の等分配．

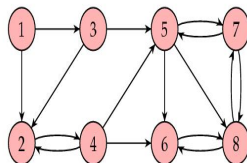
- $|\lambda_2| < 1$ となるためには S が原始的であればいい．

Definition

ある自然数 m が存在して S^m の行列要素が全て正になるとき，正方行列 S は原始的であるという．強連結性を持つともいう．

- m 個のリンクをたどっていけば，どこへでも行ける．
- 原始的（強連結）な確率行列をひねり出すためにモデルを修正したい．

もう一つ困る例：しかも本質的に困る



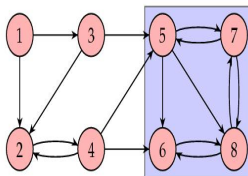
$$S = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/2 & 0 \end{bmatrix}, \quad I = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.12 \\ 0.24 \\ 0.24 \\ 0.4 \end{bmatrix}$$

(44)

- はじめの4つのページランクが0.
- 他のページからのリンクがあるにもかかわらず.
- “得点”が吸い上げられてしまう.
- “importance sink” : dangling nodes の一般化.
- dangling nodes は見つけやすいが, importance sink は見つけにくい.

Request

全てのページのページランクは厳密に正としたい.



- 青い箱が小さいネットワークを作ってしまったている .
- 一旦入ったら出ていけない .
- 行列の言葉でいえば , S が **可約** のとき :

$$S = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix} . \quad (45)$$

- ネットワークが連結：任意の2つのページ P_1, P_2 に対して，必ずリンクの列があって，それをたどって P_1 から P_2 に行ける．
- 注意：同じリンクを逆にたどれなくてもいい（相互リンク）
- 可約でない行列を既約という（群の表現論由来の言葉）．
- 連結性と既約性は単に言葉づかいを変えただけ．

Remark

行列が強連結ならば連結．

Summary

行列 S は確率行列であれ：定常ベクトルを持て．
既約（連結）であれ：定常ベクトルの成分が全て正となれ．
原始的（強連結）であれ：第2固有値 $|\lambda_2| < 1$ となれ．

Remark

原始的ならば既約．

最後の修正

- 無茶なネットサーフィンを要求する .
- パラメータ $\alpha \in (0, 1)$ を導入 .
- S の復習 : 今いるページにリンクがあればその中のどれかを使って飛ぶ . リンクがなかったら (今いるページを含めて) 適当にどこかに飛ぶ (確率論) .
- 二重に確率を使う : 確率 α で S にしたがってリンクをたどる . 確率 $(1 - \alpha)$ で適当にどこかに飛ぶ .
- S の時点で現実的にほぼ不可能なネットサーフィンだが , さらに不可能性が増した .

- Google 行列 G の定義 :

$$G := \alpha S + (1 - \alpha) \mathbf{N}, \quad (46)$$

$$\mathbf{N} := \begin{bmatrix} 1/N & \cdots & 1/N \\ 1/N & \cdots & 1/N \\ \vdots & \cdots & \vdots \\ 1/N & \cdots & 1/N \end{bmatrix} \quad (47)$$

- これは全ての条件を満たす .
- G について決定方程式 $I = GI$ を考えれば , 定常ベクトルは一意的に存在し , ベキ乗法もうまく使える .

Theorem (Perron-Frobenius)

正方行列 G が確率行列で全ての要素が正 (> 0) ならば, G は固有値 1 をもち, 定常ベクトルの成分を全て正に取れる.

Theorem (離散時間マルコフ連鎖の収束定理)

正方行列 G が強連結性を持つ確率行列ならば, 任意の初期分布 I^0 にたいして $I^k := G^k I^0$ は必ず定常ベクトルに収束する.

$$G = \alpha S + (1 - \alpha)N. \quad (48)$$

- $\alpha = 1$ のとき $G = S$: もとのネットワーク構造に沿って考える .
- $\alpha = 0$ のとき $G = N$: 全てのページが対等 , もとのリンク構造の情報が潰れる .
- α が 1 に近ければ , もとのネットワークの情報が強く出る (はず) .
- 数学としては , $\alpha \neq 1$ ならいい : $\alpha = (1 - 10^{-34})$.

- 収束の評価の問題：Google 行列の第2固有値は $|\lambda_2| = \alpha$ となる．
- α が1に近いと，ベキ乗法の収束が遅くなる．
- Brin と Page は経験的に $\alpha = 0.85$ とした．

$$(1 - 10^{-10})^{50} = 0.9999999995, \quad (49)$$

$$(1 - 10^{-10})^{10^{10}} = 0.36787941, \quad (50)$$

$$0.99^{50} = 0.60500606, \quad (51)$$

$$0.9^{50} = 0.00515377, \quad (52)$$

$$0.85^{50} = 3 \times 10^{-4}, \quad (53)$$

$$0.85^{100} = 9 \times 10^{-8}. \quad (54)$$

I の計算

- 今扱っているのは $n = 250$ 億というそこそこ大きな数．
- まともにやったら計算大変そう．

$$GI^k = \alpha HI^k + \alpha AI^k + (1 - \alpha)NI^k \quad (55)$$

- 復習： H の各列はだいたい10個くらいがノンゼロ
→ HI^k は各要素に対して10個くらいの要素を計算すればいい．
- A と N の列は同じ： AI^k と NI^k は dangling node
か全てのページの I^k の値を加えればいい．

- 次の評価が出る：

$$\|G^k I^0 - I\|_1 \leq \alpha^k \|I^0 - I\|_1, \quad (56)$$

$$\|I\|_1 := \sum_{i=1}^N |I(P_i)|. \quad (57)$$

- 数学的に**最悪**の評価が出る．
- $\alpha = 0.85$ のとき，50 回から 100 回くらいアルゴリズムを回せばそこそこの精度が出せることがアприオリに分かる．
- 計算は数日で終わるらしい．

$$(1 - 10^{-10})^{50} = 0.999999995, \quad (58)$$

$$(1 - 10^{-10})^{10^{10}} = 0.36787941, \quad (59)$$

$$0.99^{50} = 0.60500606, \quad (60)$$

$$0.9^{50} = 0.00515377, \quad (61)$$

$$0.85^{50} = 3 \times 10^{-4}, \quad (62)$$

$$0.85^{100} = 9 \times 10^{-8}. \quad (63)$$

- ウェブは常に変わり続ける．
- ニュースサイトの内容はよく変わる．
- ページの追加やリンクが外されたり追加されたりして，ハイパーリンク構造も変わり続ける．
- Google はページランクを毎月再計算しているという噂．

まとめ

- 1998 年：ウェブはその時点でのサーチエンジンの限界を超えるペースで発展しつつあった．
- その当時，ビジネスとしてサーチエンジンを作っている所はほぼすべて技術的詳細を公開していなかった．
- Brin と Page は学術的振興を図った．
- 他のランク付けアルゴリズムもある．
- Teoma サーチエンジンの基礎，Jon Kleinberg による HITS アルゴリズム．

- スパム，もしくはアフィリエイト．
- 確率行列の非対称性と定常ベクトルの非自明性．

- Michael Berry, Murray Browne, Understanding Search Engines: Mathematical Modeling and Text Retrieval. Second Edition, SIAM, Philadelphia. 2005.
- Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, 33: 107-17, 1998. Also available online at <http://infolab.stanford.edu/pub/papers/google.pdf>
- Kurt Bryan, Tanya Leise, The \$25,000,000,000 eigenvector. The linear algebra behind Google. SIAM Review, 48 (3), 569-81. 2006. Also available at <http://www.rose-hulman.edu/~bryan/google.html>
- Google Corporate Information: Technology.
- Taher Haveliwala, Sepandar Kamvar, The second eigenvalue of the Google matrix.
- Amy Langville, Carl Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2006.
- This is an informative, accessible book, written in an engaging style. Besides providing the relevant mathematical background and details of PageRank and its implementation (as well as Kleinberg's HITS algorithm), this book contains many interesting "Asides" that give trivia illuminating the context of search engine design.