情報学科CSコース

情報システム（3年後期）
第11回
（田島担当分第3回）

田島 敬史

2012 年 12 月 19 日

Web 分析

- コンテンツ
  - 要約，トピック検出
- リンク構造
  - ページの人気度，Webコミュニティの発見
- 利用履歴，アクセスログ
  - そのサイトのアクセスログ，検索エンジンでのアクセスログ
  - ページの人気度，サイトデザインの改善
- 更新履歴
  - クローラのスケジューリング

Web 検索エンジンにおけるランキング手法

- 初期の検索エンジンのランキング手法
  - 「検索式に強くマッチするページ $\fallingdotseq$ ユーザが求めるページ」
  - 例: tf-idf 法，特定のタグ内の単語
- 90年代半ばより Web では
  - 「検索式に強くマッチするページ $\neq$ 有用なページ」
  - 個人の日記ページ
  - SPAM: ページをランキング上位に押し上げるための小細工
    例: よく検索されるキーワードを小さなフォントや見えない色のフォントでたくさん埋め込む
- 解決策: 「他ページからリンクされるページ $\fallingdotseq$ 有用なページ」

  $\Rightarrow$ リンクの情報を用いたランキング手法

PageRank^{TM} (google.com)

概要

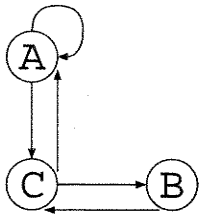
- 多くのページからリンクされるページは有用なページだろう
- 多くの有用なページからリンクされていればなお有用なページ

$\Downarrow$

- 各ページに「有用度」を割り当てる
- 有用度をリンクに沿ってページからページへと波及させる
- ページから複数のリンクが出ている場合，そのページの有用度をそれらへ均等に分割

PageRank^{TM} (google.com)

計算方法



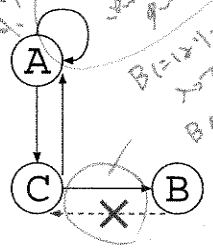
$$\begin{pmatrix} A_0 \\ B_0 \\ C_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} A_{i+1} \\ B_{i+1} \\ C_{i+1} \end{pmatrix} = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{pmatrix} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix}$$

$$\lim_{i\to\infty} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} = \begin{pmatrix} 6/5 \\ 3/5 \\ 6/5 \end{pmatrix} \leftarrow \cdots \leftarrow \begin{pmatrix} 9/8 \\ 1/2 \\ 11/8 \end{pmatrix} \leftarrow \begin{pmatrix} 5/4 \\ 3/4 \\ 1 \end{pmatrix} \leftarrow \begin{pmatrix} 1 \\ 1/2 \\ 3/2 \end{pmatrix} \leftarrow \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

PageRank^{TM} (google.com)

dead end problem

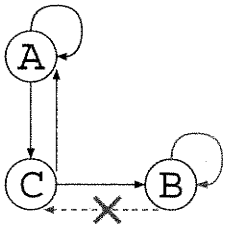


$$\begin{pmatrix} A_{i+1} \\ B_{i+1} \\ C_{i+1} \end{pmatrix} = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix}$$

$$\lim_{i\to\infty} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \leftarrow \cdots \leftarrow \begin{pmatrix} 5/8 \\ 1/4 \\ 3/8 \end{pmatrix} \leftarrow \begin{pmatrix} 3/4 \\ 1/4 \\ 1/2 \end{pmatrix} \leftarrow \begin{pmatrix} 1 \\ 1/2 \\ 1/2 \end{pmatrix} \leftarrow \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

PageRank^{TM} (google.com)

spider trap problem



$$\begin{pmatrix} A_{i+1} \\ B_{i+1} \\ C_{i+1} \end{pmatrix} = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix}$$

$$\lim_{i\to\infty} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix} \leftarrow \cdots \leftarrow \begin{pmatrix} 5/8 \\ 2 \\ 3/8 \end{pmatrix} \leftarrow \begin{pmatrix} 3/4 \\ 7/4 \\ 1/2 \end{pmatrix} \leftarrow \begin{pmatrix} 1 \\ 3/2 \\ 1/2 \end{pmatrix} \leftarrow \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

PageRank^{TM} (google.com)

dead end problem と spider trap problem の回避

$$\begin{pmatrix} A_{i+1} \\ B_{i+1} \\ C_{i+1} \end{pmatrix} = (1-tax) \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} + tax \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

↑ spider trap を避けるために税を徴収

↑ dead end problem を避けるために税を再分配

$$\lim_{i\to\infty} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} = \begin{pmatrix} 7/11 \\ 21/11 \\ 5/11 \end{pmatrix} \leftarrow \cdots \leftarrow \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad (tax = 0.2)$$

PageRank^{TM} (google.com)

ランダムウォークとしての解釈

$$\begin{pmatrix} A_{i+1} \\ B_{i+1} \\ C_{i+1} \end{pmatrix} = (1-\epsilon) \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix} \begin{pmatrix} A_i \\ B_i \\ C_i \end{pmatrix} + \epsilon \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

↑ 時間$i+1$に各ページにいる確率

↑ ジャンプしない場合

↑ 時間iに各ページにいる確率

↑ ランダムにジャンプする場合

ϵ = 「リンクをたどらずにランダムにジャンプする確率」

HITS (Hypertext Induced Topic Search)

"Authoritative Sources in a Hyperlinked Environment"
by Jon M. Kleinberg, Journal of ACM, 46(5), 1999

- 与えられたキーワードを含むページ集合を取得 (Focused Subgraph)
- このページ集合に対してハブ—オーソリティ解析
  そのキーワードに対応するトピックに関する「オーソリティ」と「ハブ」を求める.
- YST (Yahoo! Search Technology) が採用したと言われている.
- Jon Kleinberg が 2006 年, IMU Nevanlinna 賞を受賞

ハブ—オーソリティ解析

概要

- ページには二種類の情報がある
  - 情報を提供するページ
  - 情報を持つページへのリンクを提供するページ (例: ポータル, リンク集ページ)
- オーソリティ = 有用な情報を持つページ
- ハブ = 有用なリンクを持つページ

ハブ—オーソリティ解析

概要

各ページに「オーソリティ度」と「ハブ度」を割り当てる

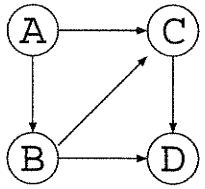
- ページのオーソリティ度とは？
  多くの良いハブからリンクされているページは良いオーソリティ
- ページのハブ度とは？
  多くの良いオーソリティをリンクしているページは良いハブ

相互再帰的に定義される

ハブ—オーソリティ解析

計算方法

$$\overrightarrow{h}_{i+1} = A\,\overrightarrow{a_i}$$



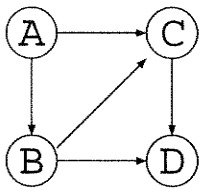
$$\begin{pmatrix}2\\2\\1\\0\end{pmatrix} = \begin{pmatrix}0&1&1&0\\0&0&1&1\\0&0&0&1\\0&0&0&0\end{pmatrix}\begin{pmatrix}1\\1\\1\\1\end{pmatrix}$$

↑ ハブ度　↑ 隣接行列 A　↑ オーソリティ度

ハブ—オーソリティ解析

計算方法

$$\overrightarrow{a}_{i+1} = A^t\,\overrightarrow{h_i}$$



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

↑ オーソリティ度　↑ A^t　↑ ハブ度

ハブ—オーソリティ解析

計算方法

$$A^tA = \begin{pmatrix}0&0&0&0\\1&0&0&0\\1&1&0&0\\0&1&1&0\end{pmatrix}\begin{pmatrix}0&1&1&0\\0&0&1&1\\0&0&0&1\\0&0&0&0\end{pmatrix} = \begin{pmatrix}0&0&0&0\\0&1&1&0\\0&1&2&1\\0&0&1&2\end{pmatrix}$$

オーソリティ度: $\overrightarrow{a_{2i}} = (A^tA)^i\,\overrightarrow{a_0}$

$$(A^tA)^i\begin{pmatrix}1\\1\\1\\1\end{pmatrix} \to \begin{pmatrix}0\\2\\4\\3\end{pmatrix} \to \begin{pmatrix}0\\6\\13\\10\end{pmatrix} \to \begin{pmatrix}0\\19\\42\\33\end{pmatrix} \to \begin{pmatrix}0\\61\\136\\108\end{pmatrix} \to \cdots$$

ハブ—オーソリティ解析

計算方法

A^tA は実数対称行列

→ 以下を満たす対角行列 D が存在

$$A^tA = U^{-1}DU$$

よって

$$(A^tA)^i\overrightarrow{a_0} = U^{-1}D^iU\overrightarrow{a_0}$$

通常は，さらに正規化を行う

$$\overrightarrow{a_{i+2}} = \frac{A^tA\overrightarrow{a_i}}{|A^tA\overrightarrow{a_i}|}$$

ハブ—オーソリティ解析

計算方法

ハブ度も同様:

$$\overrightarrow{h_{2i}} = (AA^t)^i \overrightarrow{h_0}$$

AA^t は実数対称行列

→ 以下を満たす対角行列 D_h が存在

$$AA^t = U_h^{-1} D_h U_h$$

よって

$$(AA^t)^i \overrightarrow{h_0} = U_h^{-1} D_h^i U_h \overrightarrow{h_0}$$

正規化:

$$\overrightarrow{a_{i+2}} = \frac{AA^t \overrightarrow{a_i}}{|AA^t \overrightarrow{a_i}|}$$

ハブ—オーソリティ解析

リンク関係の主成分分析としての解釈

$$\begin{pmatrix} h = \lim_{i\to\infty} h_i \\ a = \lim_{i\to\infty} a_i \end{pmatrix} \text{ とおくと } \begin{pmatrix} h = \mu A a \\ a = \eta A^t h \end{pmatrix} \text{ よって } \begin{pmatrix} h = \mu\eta A A^t h \\ a = \mu\eta A^t A a \end{pmatrix}$$

$\mu\eta = \lambda^{-1}$ とおくと,

$$\begin{pmatrix} \lambda h = AA^t h \\ \lambda a = A^t A a \end{pmatrix}$$

h, a は AA^t, A^tA の固有ベクトル
(最大の固有値に対する)

ハブ—オーソリティ解析

リンク関係の主成分分析としての解釈

- $AA^t(i,j)$ = ノードi, j が共にリンクしているノード数
- $A^tA(i,j)$ = ノードi, j を共にリンクしているノード数

⇓

- h は「どんなノードをリンクしているか」に関する主成分分析の第一主成分の係数
- a は「どんなノードからリンクされているか」に関する主成分分析の第一主成分の係数