

情報アクセスシステム

第8回

リンク解析その2

Topic-sensitive PageRank, HITS

兵庫県立大学

情報科学研究科/社会情報科学部

山本 岳洋

t.yamamoto@sis.u-hyogo.ac.jp

2021年度前期・水曜2限 社会情報科学部

- 各種リンク解析アルゴリズム

- Topic-sensitive PageRank
(Biased PageRank, Personalized PageRank)
- HITSアルゴリズム

Topic-sensitive PageRank


(Biased PageRank,
Personalized PageRank)

- アイデア: **トピック**に依存したPageRank値計算
 - トピック:
 - ビジネス, 健康, コンピュータ, スポーツなど
ODP (Open Directory Project) の上位16カテゴリ
 - 例: スポーツ分野におけるPageRank値を求めたい


Open Directory Project













5

人手により収集・カテゴリ分類されたウェブディレクトリ

Curlie [About](#) [Forum](#) [Donate](#) 

Collect the best websites for any topic!
Search or browse by category

Search Curlie in English 

 Arts Movies, Television, Music...	 Business Jobs, Real Estate, Investing...	 Computers Internet, Software, Hardware...
 Games Video Games, RPGs, Gambling...	 Health Fitness, Medicine, Alternative...	 Home Family, Consumers, Cooking...
 News Media, Newspapers, Weather...	 Recreation Travel, Food, Outdoors, Humor...	 Reference Maps, Education, Libraries...
 Regional US, Canada, UK, Europe...	 Science Biology, Psychology, Physics...	 Shopping Clothing, Food, Gifts...

<https://curlie.org/en>

2021年6月現在, curlie.orgが運営

Topic-Sensitive PageRank

(Biased PageRank, Personalized PageRank)

- アイデア：テレポーテーションを偏らせる

ビジネス 健康

$\mathcal{C} = \{c_1, c_2, \dots, c_m\}$ ODPのトピック集合

$\mathcal{T} = \{T_1, T_2, \dots, T_m\}$ T_j はトピック c_j に所属するページ集合

トピック c_j に対するPageRank値ベクトルを

$\mathbf{p}_j = (p_{j1}, p_{j2}, \dots, p_{jn})^T$ とすると

Topic-Sensitive PageRankアルゴリズム

$$\mathbf{p}_j = d\mathbf{A}^T \mathbf{p}_j + (1 - d)\mathbf{v}_j \quad \mathbf{v}_{ji} = \begin{cases} \frac{1}{|T_j|} & i \in T_j \\ 0 & i \notin T_j \end{cases}$$

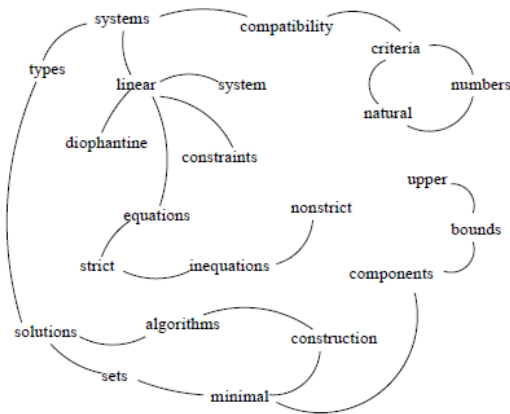
偏ったテレポーテーションに基づくPageRankアルゴリズムを
Biased PageRank, Personalized PageRankと呼ぶこともある

Topic-Sensitive PageRank (Biased PageRank)の直感的な理解

- シードページ集合 T_j から出発したランダムサーファ어가確率 d で閲覧中のページのリンクを辿り, **確率 $(1-d)$ でシードページ集合のいずれかに戻る**
 - 参考: 戻るページが1つだけだと,
Random walk with Restartとも呼ばれる
- ノードに対する「事前に分かっている重要度」をリンクを通じて伝播しているとも解釈できる
 - たとえば, v_j を個人のブックマークから用意すれば,
パーソナライズされたページの重要性が計算できる

● TextRank, LexRank

- 多くの重要な文（単語）と類似する文（単語）は重要
- 文書要約，文書からの重要キーワード発見に利用



Keywords assigned by TextRank:

linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Keywords assigned by human annotators:

linear constraints; linear diophantine equations; minimal generating sets; nonstrict inequations; set of natural numbers; strict inequations; upper bounds

$$\mathbf{p} = d\mathbf{S}^T\mathbf{p} + (1 - d)\frac{1}{n}\mathbf{e}$$

$$S_{ij} = \frac{1}{Z} \text{sim}(i, j) \quad \text{sim}(i, j): \text{文や単語間の類似度や共起度}$$

Z : 確率行列にするための定数項

図の出典: R. Mihalcea, P. Tarau: TextRank: Bringing Order into Text, EMNLP2004, pp. 404-411.

G. Erkan, D. R. Radev: LexRank: Graph-based Lexical Centrality as Saliency in Text Summarization, Journal of artificial intelligence research, 22, pp. 457-479, 2004.

● VisualRank

- 多くの重要な画像と類似する画像は重要
- 画像の典型性評価に利用



$$\mathbf{p} = d\mathbf{S}^T\mathbf{p} + (1 - d)\frac{1}{n}\mathbf{e}$$

$$S_{ij} = \frac{1}{Z} \text{sim}(i, j) \quad \text{sim}(i, j): \text{画像 } i, j \text{ 間の類似度}$$

Z : 確率行列にするための定数項

HITSアルゴリズム

HITSアルゴリズム

(Hypertext Induced Topic Search)

- 目的: Webコミュニティ発見

- あるトピックに関して, Web上の (隠された) コミュニティをリンク構造から発見したい
- 具体的には, 優れたオーソリティとハブを発見したい

- **オーソリティ (Authority)**

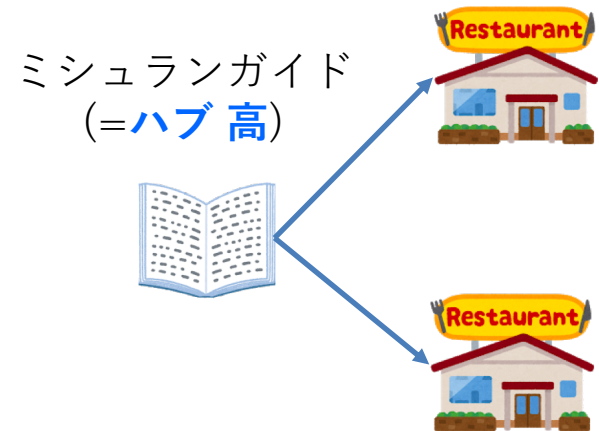
- 多くの (重要な) **ハブ** から
リンク されている 重要なページ

- **ハブ (Hub)**

- 多くの (重要な) **オーソリティ** に
リンク している 重要なページ

- **ハブとオーソリティは相互再帰的な関係**

ミシュランガイドで
紹介される店
(=**オーソリティ** 高)



有向グラフ: $G = (V, E)$ ※ 元論文ではここでの V はクエリに依存した集合だが, この講義では扱わない

ページ i のオーソリティ値を a_i , ハブ値を h_i とすると,

$$\left\{ \begin{array}{l} a_i = \sum_{(j,i) \in E} h_j \\ h_i = \sum_{(i,j) \in E} a_j \end{array} \right. \quad \text{式(4)}$$

※HITSはマルコフ連鎖ではないので注意

式(4)の別表現

隣接行列 (Adjacency matrix) : L : $n \times n$ 行列

$$L_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$

オーソリティ値ベクトル $\mathbf{a} = (a_1, a_2, \dots, a_n)^T$,
ハブ値ベクトル $\mathbf{h} = (h_1, h_2, \dots, h_n)^T$ とおくと,

$$\begin{cases} \mathbf{a} = L^T \mathbf{h} \\ \mathbf{h} = L \mathbf{a} \end{cases}$$

それぞれの式の
 \mathbf{h} と \mathbf{a} を置換すると



$$\begin{cases} \mathbf{a} = L^T L \mathbf{a} \\ \mathbf{h} = L L^T \mathbf{h} \end{cases}$$

べき乗法に基づくHITSアルゴリズム

input: 隣接行列 L

output: オーソリティ値ベクトル \mathbf{a}
ハブ値ベクトル \mathbf{h}

$$\mathbf{a}^{(0)} \leftarrow \mathbf{h}^{(0)} \leftarrow (1, 1, \dots, 1)^T$$

$$k \leftarrow 1$$

repeat

$$\mathbf{a}^{(k)} = L^T L \mathbf{a}^{(k-1)}$$

$$\mathbf{h}^{(k)} = L L^T \mathbf{h}^{(k-1)}$$

$$\mathbf{a}^{(k)} \leftarrow \mathbf{a}^{(k)} / \|\mathbf{a}^{(k)}\|_1 \quad // \text{正規化}$$

$$\mathbf{h}^{(k)} \leftarrow \mathbf{h}^{(k)} / \|\mathbf{h}^{(k)}\|_1 \quad // \text{正規化}$$

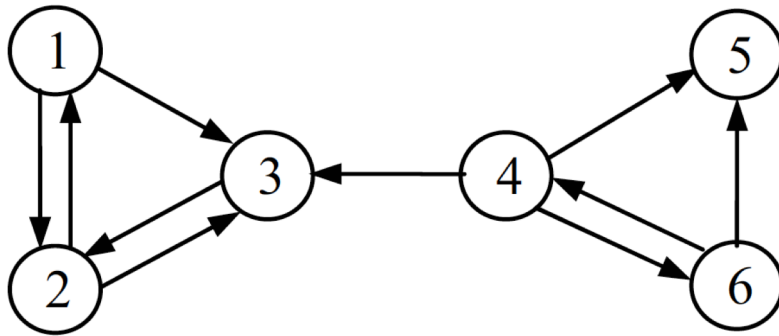
$$k \leftarrow k + 1$$

until $\|\mathbf{a}^{(k)} - \mathbf{a}^{(k-1)}\|_1 < \varepsilon_a$
and $\|\mathbf{h}^{(k)} - \mathbf{h}^{(k-1)}\|_1 < \varepsilon_a$

return \mathbf{a}_k and \mathbf{h}_k

常に収束するが、初期値の
与え方によっては異なる
ベクトルに収束する場合もある
(LL^T の性質による)

参考: 前回の講義資料の例に対して HITSを適用してみる



networkxのhits()で
簡単に求められる

```
hub, authority = nx.hits(G)
```

Authority

```
[(3, 0.35), (5, 0.21), (6, 0.15), (2, 0.13), (1, 0.10), (4, 0.06)]
```

Hub

```
[(4, 0.35), (1, 0.24), (2, 0.22), (6, 0.13), (3, 0.06), (5, 0.00)]
```

参考

- **HITSがコミュニティ発見と言われる理由**
 - HITSはグラフ中の最も密に繋がっている部分グラフにおけるオーソリティ, ハブを求めている

- **参考: 関連する技術**
 - SALSA
 - HITSのマルコフ連鎖版
 - Co-HITS
 - 2部グラフ上での一般的なランダムウォークアルゴリズム

● 耐スパム性

- 出リンクを操作するのは容易なため、ハブ値を簡単に操作できる
- 結果的にオーソリティ値も操作できてしまう

この資料のまとめ

● 各種リンク解析アルゴリズム

- Topic-sensitive PageRank
(Biased PageRank, Personalized PageRank)
 - 偏ったテレポーションをかける
= 一定確率で特定のノードに戻る
- HITS
 - ハブとオーソリティの考え方
 - 隣接行列を用いてハブとオーソリティをどのように定式化できるか？