Web Information System Design
No.8 データのWeb

萩野　達也　(hagino@sfc.keio.ac.jp)

Web文書とWebアプリケーション

- Web文書
  -
  -
  -
  -
  -
- Webアプリケーション
  -
  -
  -
  -
  -

検索エンジンは万能アプリか？

- 少ないキーワードからWeb文書を探し出す
  - 個のキーワードを与えることが多い

- 検索エンジンの仕組み
  1. Webサイトを　　してWeb文書を集めてくる
     - ハイパーリンクをたどる
  2. すべての単語をキーワードだと思い索引を作る
     -
  3. 検索で与えられたキーワードから作品を使ってWebページを探す
     - AND, OR, NOT
  4. 　　を使って検索結果の表示順を決める
     -

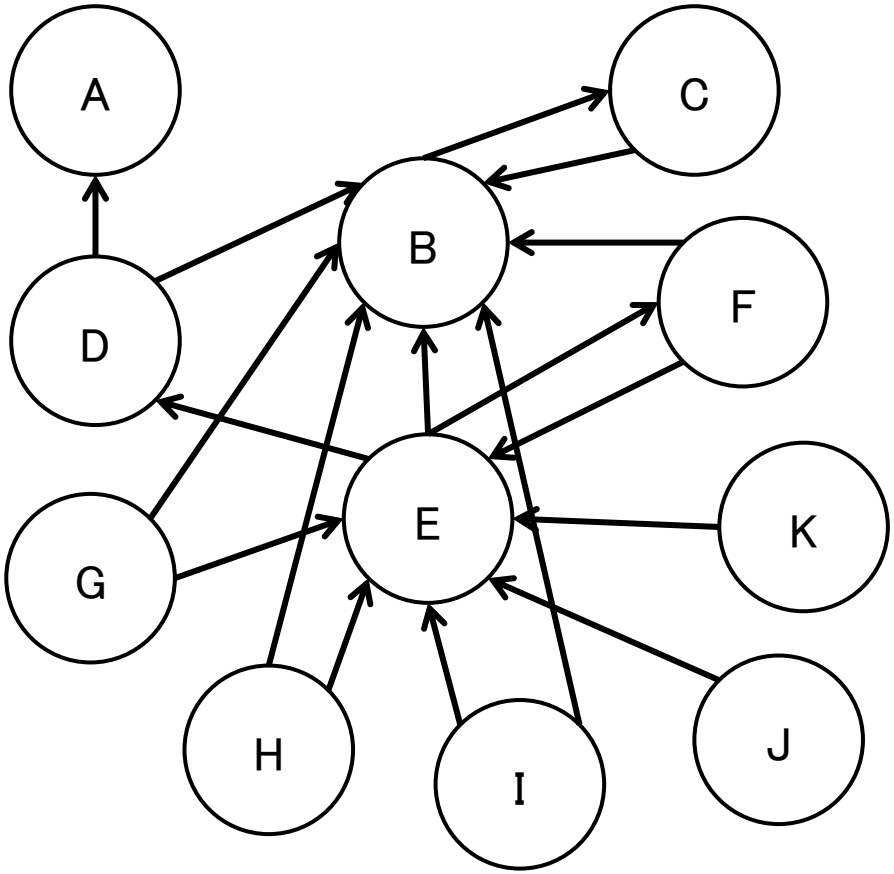
ページランクのアルゴリズム

- Webページ: $p_1, p_2, \ldots, p_N$
  - $M(p_i)$ = p_iにリンクしているページの集合
  - $L(p_j)$ = p_jから出ているリンクの数
  - N = 前ページ数
  - d = ダンピング・ファクター　0.85 (85%　　, 15%　　)

- ページランク: $PR(p_i)$
  - $PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$

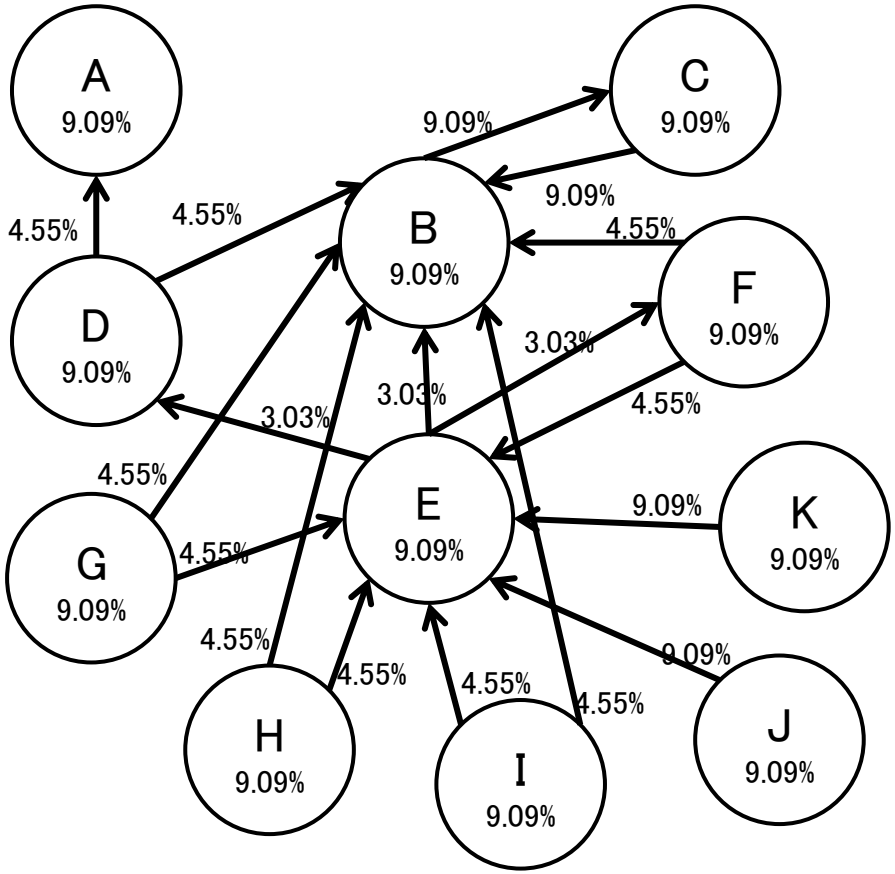
- 計算
  - ページランクの初期値: $PR(p_i) = \frac{1}{N}$
  - 上記の式を使ってページランクを更新していく

例



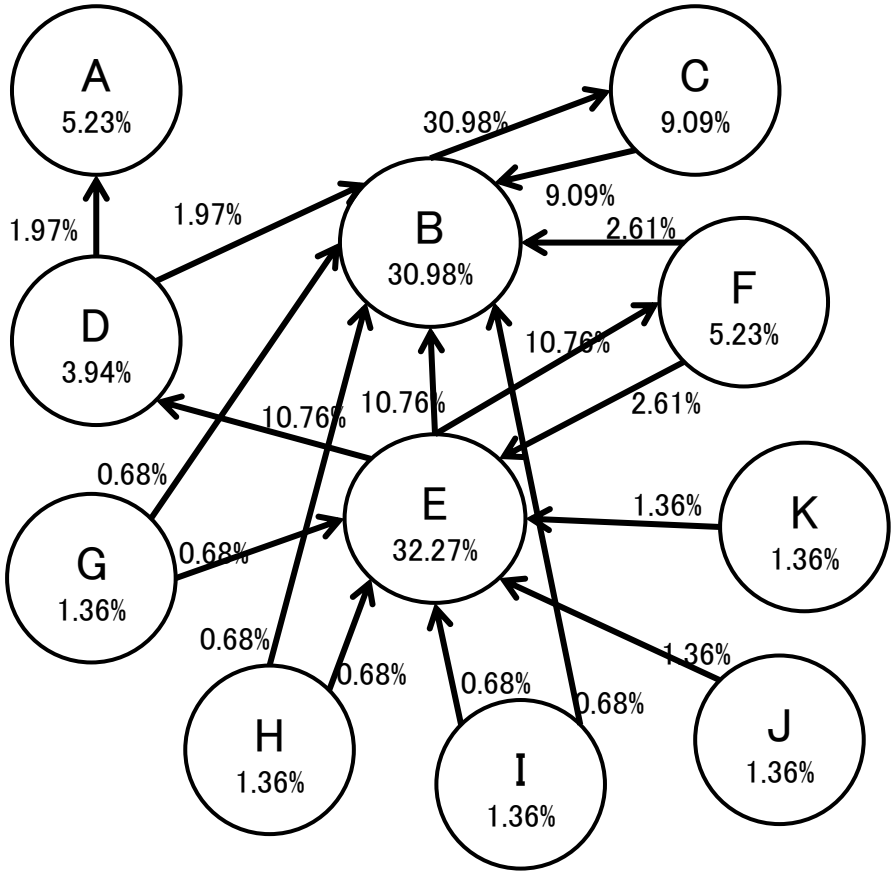
page	$M(p_i)$	$L(p_j)$
A		0
B	C, D, E, F, G, H, I	1
C	B	1
D	E	2
E	F, G, H, I, J, K	3
F	E	2
G		2
H		2
I		2
J		1
K		1

例（t=0）



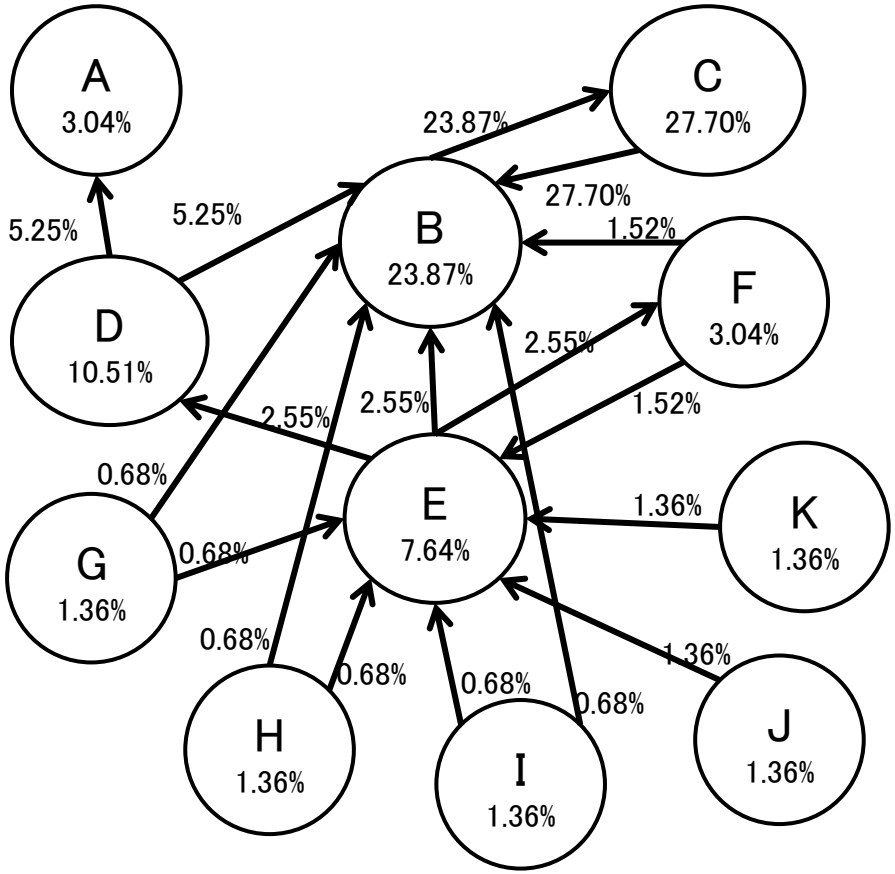
page	$PR(p_i:\mathbf{0})$	$PR(p_i:\mathbf{1})$
A	9.09%	5.23%
B	9.09%	30.98%
C	9.09%	9.09%
D	9.09%	3.94%
E	9.09%	32.27%
F	9.09%	5.23%
G	9.09%	1.36%
H	9.09%	1.36%
I	9.09%	1.36%
J	9.09%	1.36%
K	9.09%	1.36%

例（t=1）



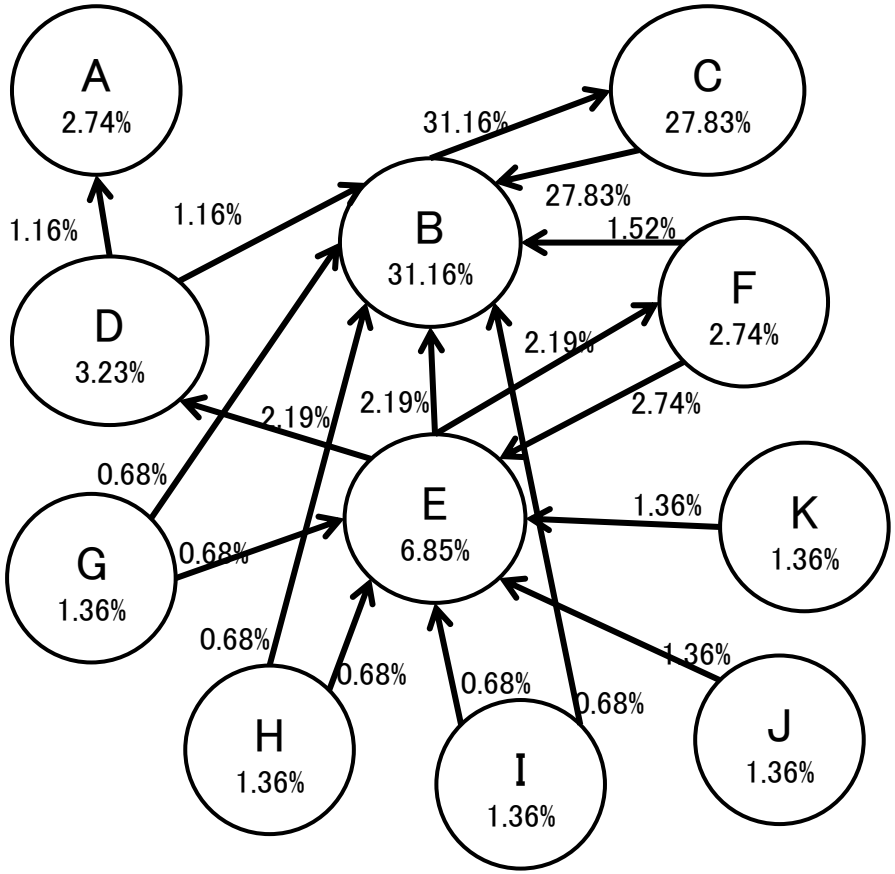
page	$PR(p_i:\mathbf{1})$	$PR(p_i:\mathbf{2})$
A	5.23%	3.04%
B	30.98%	23.87%
C	9.09%	27.70%
D	3.94%	10.51%
E	32.27%	7.64%
F	5.23%	3.04%
G	1.36%	1.36%
H	1.36%	1.36%
I	1.36%	1.36%
J	1.36%	1.36%
K	1.36%	1.36%

例（t=2）



page	$PR(p_i:\mathbf{2})$	$PR(p_i:\mathbf{3})$
A	3.04%	5.83%
B	23.87%	34.57%
C	27.70%	21.65%
D	10.51%	3.53%
E	7.64%	6.71%
F	3.04%	5.83%
G	1.36%	1.36%
H	1.36%	1.36%
I	1.36%	1.36%
J	1.36%	1.36%
K	1.36%	1.36%

例（t=30）



page	$PR(p_i:\mathbf{30})$
A	2.74%
B	31.16%
C	27.83%
D	3.23%
E	6.58%
F	2.74%
G	1.36%
H	1.36%
I	1.36%
J	1.36%
K	1.36%

検索エンジンで十分か？

- 検索エンジンができること:
  -
  -
  -
- 検索エンジンができないこと:
  -
  -
  -
  -

Web上のデータ

- Webには有益なデータがたくさんある:

Webにあるデータの形式

- HTMLの表
  - 良く使われる，見やすい
  - データを取り出すのが困難（データ間の関係が分からない）

辻堂駅　東海道本線 横浜・東京方面（上り）　▶土曜・休日の時刻表を表示

時	平日
5	00 27 50
6	快高 02 11 19 快籠 27 34 40 44 47 湘ラ 52 55 58
7	01 快高 04 07 10 13 16 19 22 快籠 25 29 31 34 37 快高 40 46 49 51 54 58
8	01 04 09 12 15 19 快籠 22 27 湘ラ 30 35 43 湘ラ 51 54
9	01 10 23 33 45 快籠 50 55
10	05 15 26 38 快籠 49 59
11	14 32 快籠 50 53
12	02 14 35 44 快籠 50
13	03 15 26 32 39 快籠 50 56
14	07 15 25 38 快籠 50 54
15	04 17 32 40 快籠 50 58
16	02 16 26 29 42 快籠 50 55
17	07 14 快高 19 24 34 46 快籠 51 54
18	品 05 11 快前 22 26 品 30 35 45 快籠 50 57
19	08 20 快高 23 32 42 48 快籠 54 58
20	09 快籠 21 25 37 快高 42 50
21	00 14 快高 21 25 39 51
22	04 19 25 33 快籠 39 45 58
23	品 12 品 27 品 39

列車種別・列車名：　無印=普通 快=快速 湘ラ=湘南ライナー
行き先・経由：　無印=東京 籠=籠原 高=高崎 品=品川 前=前橋

```
<table>
  <tr>
    <th>時</th>
    <th colspan="19">平日</th>
  </tr>
  ...
  <tr>
    <th>6</th>
    <td>快高<br />02</td>
    <td>11</td>
    <td>19</td>
    <td>快籠<br />27</td>
    <td>34</td>
    ....
  </tr>
  ....
</table>
```

CSV

- Comma-Separated Values
  - テキスト形式
  - 列はコンマで区切られる
  - データ交換によく用いられる
  - ExcelからCSVを出力可能

World Country

```
Country,Capital,Population,Area(km2),Official Languages
Japan,Tokyo,126659683,377944,Japanese
United States,"Washington, D.C.",318133000,9826675,English
United Kingdom,London,63705000,243610,English
France,Paris,66616416,640679,French
China,Beijing,1350695000,9596961,Standard Chinese
India,New Delhi,1210193444,3287590,"Hindi,English"
...
...
```

XMLで表現

- 時刻表

```
<?xml version="1.0" encoding="Shift_JIS"?>
<timetable>
  <station name="辻堂">
    <line name="東海道" dir="上り" week="平日">
      <train at="6:02" dest="高崎" kind="計測" />
      <train at="6:11" />
      <train at="6:19" />
      <train at="6:27" dest="籠原" kind="快速" />
      ...
      <train at="6:62" kind="湘南ライナー" />
      ...
    </line>
    ...
  </station>
  ...
</timetable>
```

XMLで表現

- World Country

```
<?xml version="1.0" encoding="Shift_JIS"?>
<world>
  <country name="Japan">
    <capital>Tokyo</capital>
    <population>126659683</polulation>
    <area unit="km2">377944</area>
    <language>Japanese</language>
  </country>
  ...
  <country name="India">
    <capital>New Delhi</capital>
    <population>1210193444</polulation>
    <area unit="km2">3287590</area>
    <language>Hindi</language>
    <language>English</language>
  </country>
  ...
</world>
```

HTMLの表 vs CSV vs XML

	HTML の表	CSV	XML
利点			
欠点			

データベース

- 関係データベース
  - データを表として表す
  - 依存関係から正規化する
    - 主キー
  - 関係代数
    - 制限，射影，結合
  - 検索言語
    - SQL

Country

id	name	capital	population	area
1	Japan	Tokyo	126659683	377944
2	United States	Washington, D.C.	318133000	9826675
3	India	New Delhi	1210193444	3287590
…	…	…	…	…

↑
主キー

Language

country	language
1	Japanese
2	English
3	Hindi
3	English
…	…

複合キー

Webデータの特徴

- 開世界である
  -
  -
    -
  -
- 複数のデータの結合
  -
  -
  -
- 矛盾するデータ
  -
  -
  -

データの結合

- 個人のCD管理
  -
  -
  -
- くすりの管理
  -
  -
  -
- スケジュール管理
  -
  -
  -
  -
  -
- 食事管理
  -
  -
  -

どのようなデータが欲しいか？

- どのようなデータが欲しいか？あればうれしいか？
  -
  -
- 現在あるデータおよびその形式は？
  -
  -
  -

まとめ

- 文書とアプリケーション
  - 静的情報
  - オンラインショッピング
- 検索エンジン
  - 全文検索
  - ページランク
- 文書とデータ
  - データ形式
  - データベース
  - データの組み合わせ
  - AAAの原則：Anybody can say anything about any topic.
  - 開世界