Web Information System Design
No.6 Search Engine

Tatsuya Hagino (hagino@sfc.keio.ac.jp)

Web Documents vs Web Application

- Web Documents
  - Home pages of universities, companies and organizations
  - News
  - Information about products, …
  - Blogs and twitters
  - Multimedia: photos, movies

- Web Application
  - Online shopping: goods, books, …
  - Online banking
  - Online reservation: hotel, train, airplane, movies, events, …
  - Online games
  - Search engine: google, yahoo, bing, …

Search Engine as Generic Application

- Find Web documents with few keywords
  - Only two or three keywords are given.

- Mechanism of search engine
  1. Collect Web documents by crawling Web sites
     - Follow hyperlinks
  2. Store all the words as keywords and create index
     - Full text search
  3. When keywords are given, use index to find Web pages related to the keywords
     - AND, OR, NOT
  4. Order Web pages using page rank algorithm
     - Pages which are referenced more are more valuable

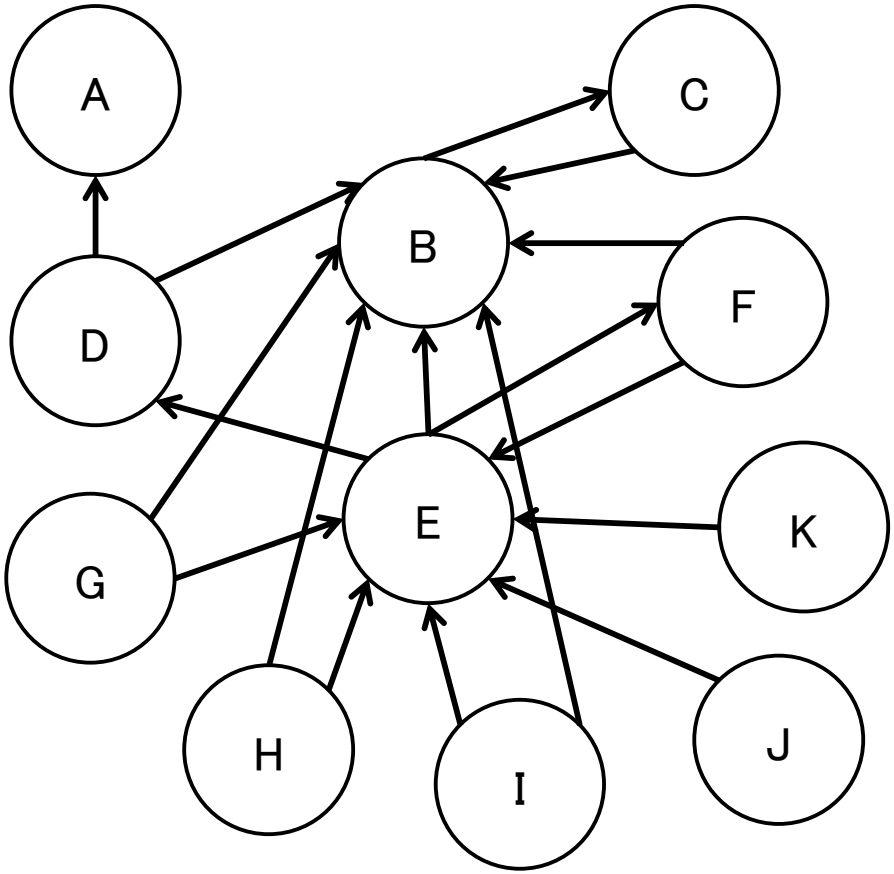
Page Rank Algorithm

- Pages: p_1, p_2, ..., p_N
  - $M(p_i)$ = the set of pages that link to p_i
  - $L(p_j)$ = the number of outbound links on page p_j
  - N = the total number of pages
  - d = dumping factor 0.85 (85% follow links, 15% start over)

- Page rank: $PR(p_i)$
  - $PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$

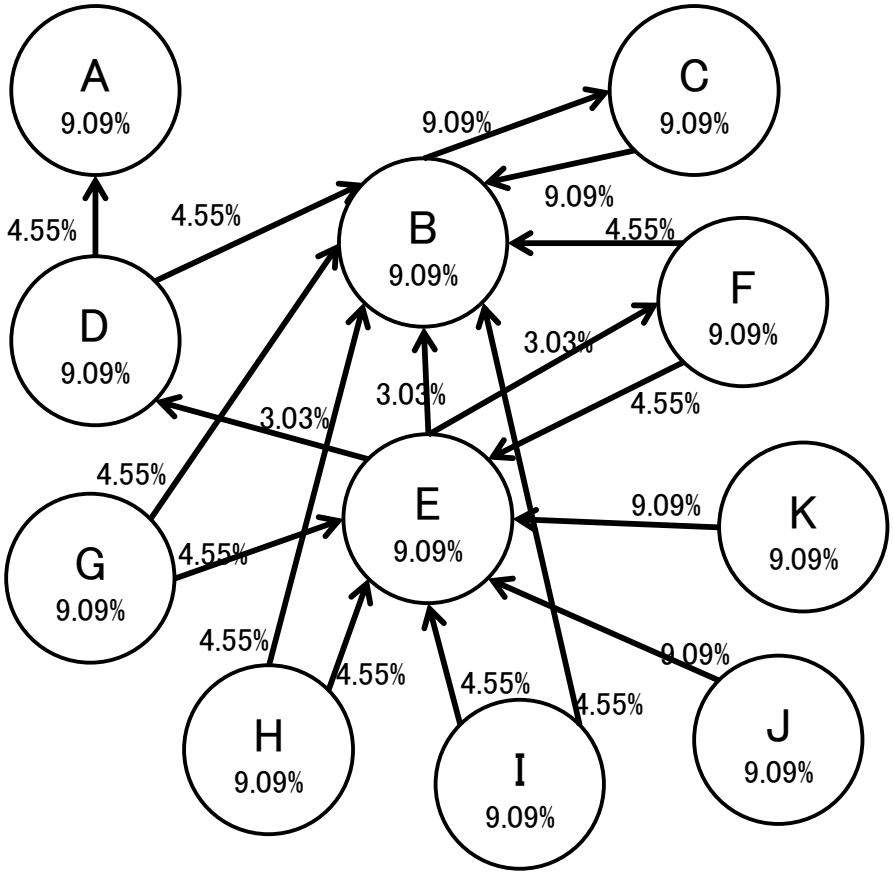
- Computation
  - Initial page rank: $PR(p_i) = \frac{1}{N}$
  - Iteratively update page rank using the above formula

Example



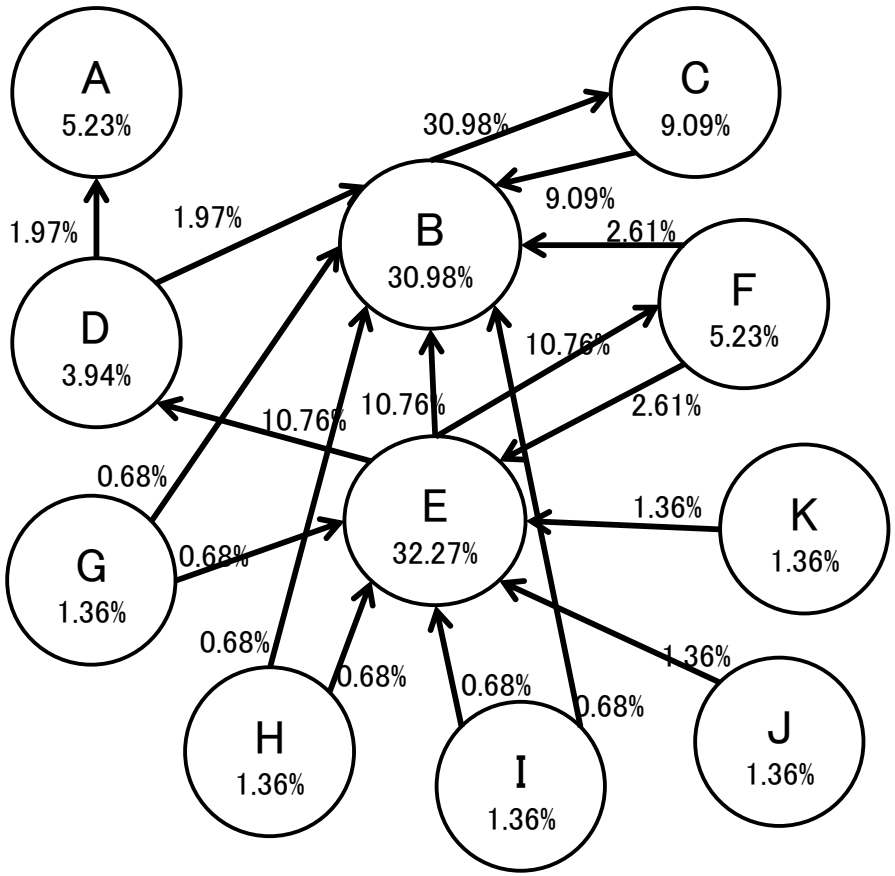
page	$M(p_i)$	$L(p_j)$
A		0
B	C, D, E, F, G, H, I	1
C	B	1
D	E	2
E	F, G, H, I, J, K	3
F	E	2
G		2
H		2
I		2
J		1
K		1

Example (t=0)



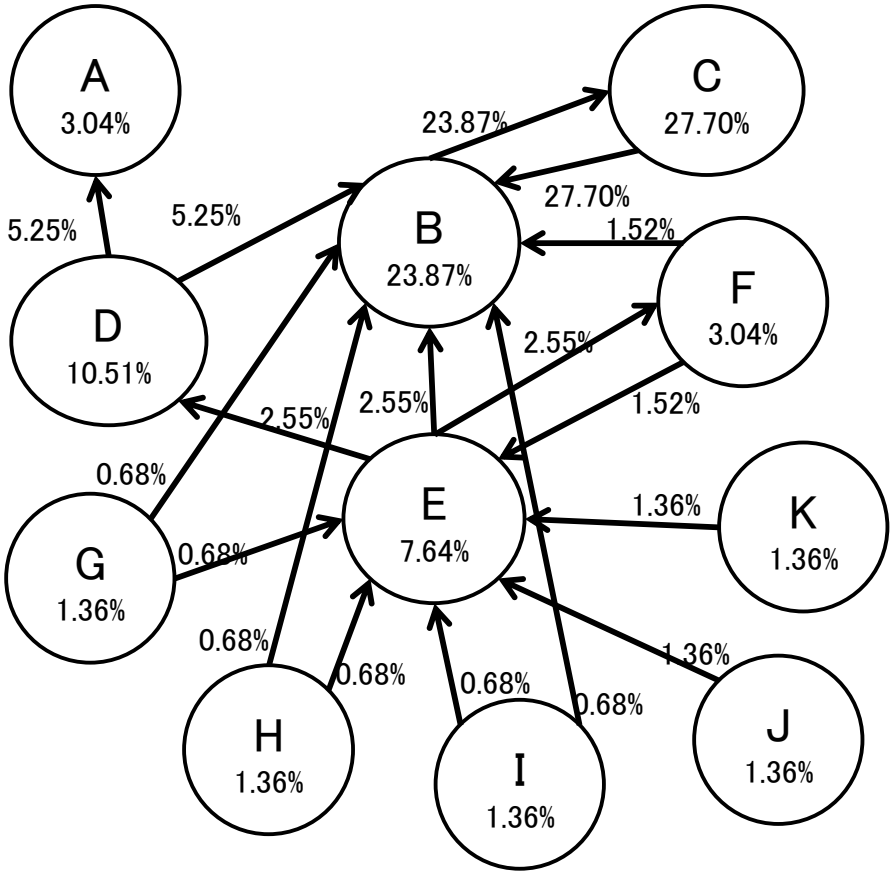
page	$PR(p_i: \mathbf{0})$	$PR(p_i: \mathbf{1})$
A	9.09%	5.23%
B	9.09%	30.98%
C	9.09%	9.09%
D	9.09%	3.94%
E	9.09%	32.27%
F	9.09%	5.23%
G	9.09%	1.36%
H	9.09%	1.36%
I	9.09%	1.36%
J	9.09%	1.36%
K	9.09%	1.36%

Example (t=1)



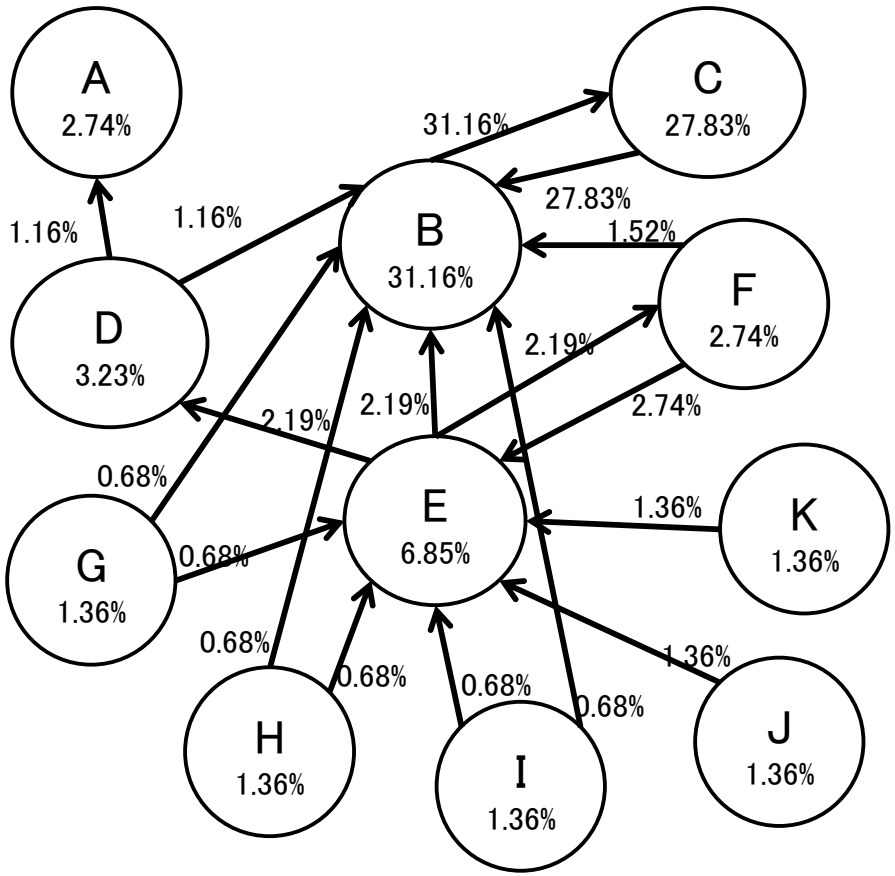
page	$PR(p_i: 1)$	$PR(p_i: 2)$
A	5.23%	3.04%
B	30.98%	23.87%
C	9.09%	27.70%
D	3.94%	10.51%
E	32.27%	7.64%
F	5.23%	3.04%
G	1.36%	1.36%
H	1.36%	1.36%
I	1.36%	1.36%
J	1.36%	1.36%
K	1.36%	1.36%

Example (t=2)



page	$PR(p_i: 2)$	$PR(p_i: 3)$
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%

Example (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%

Is Search Engine Perfect?

- Search Engine can do:
  - Find Web pages relevant to given keywords.
  - Do not need to categorize pages.
  - Do not need any feedback from users.

- Search Engine cannot do:
  - Find hidden pages or dynamic pages.
  - Combine information of different Web sites.
  - Create Summary page.
  - Online banking, shopping, ...