情報検索システム特論
Advanced Information Retrieval Systems
第11回 Lecture #11

2023-06-19

湯川 高志 Takashi Yukawa

Alternative Retrieval Models (cont'd)

refer the material for Lecture #09-10 pp.54-80

Page Rank Technology

Page Rank

- Web Information Retrieval
  - Crawling
  - Indexing
  - Retrieving --- Scoring

- Rage Rank is trademark of Google

Motivation

- Want to retrieve well-written Web pages
  → Giving high score for well-written pages

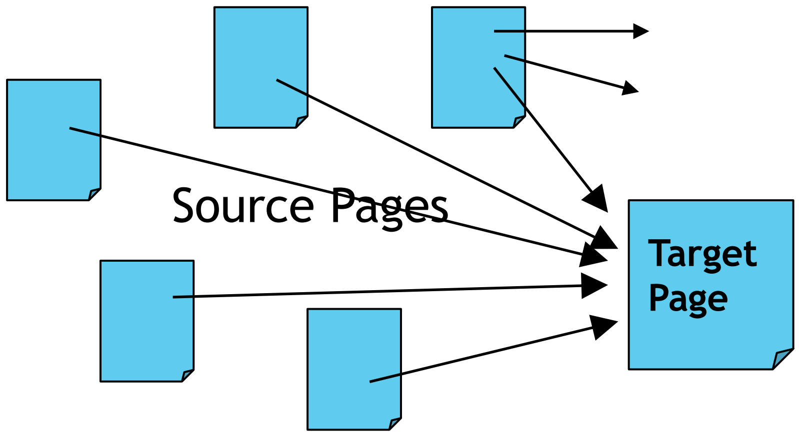
- Recognizing “well-written” – How?
  - Reading the page as a human does
    → natural language processing, knowledge processing

Too complicated

  - Using another information, which is special in WWW

The Idea

- A page which linked by a lot of well-written pages is also a well-written page
- Criteria of “well-written” for the web page
  - If a source page is “well-written”, the target page would be “well-written”
  - If a source page has less outgoing links, the target page would have better quality
  - The target page are linked by more source pages which are “well-written” and have less links, the target page would be very good page



Principle

- A Link form Page i to Page j → Page i votes Page j
- Score of Page j is determined based on
  - The number of votes
  - Score of Page i
  - The number of outgoing links in Page i (less is better)

Computation

- A Link from Page i to Page j

$$a_{i,j} = \begin{cases} 1, \text{there is a link from Page } i \text{ to Page } j \\ 0, \text{there is no link from Page } i \text{ to Page } j \end{cases}$$

- Create a matrix, which $a_{i,j}$ is an element at row i and column j → adjacency matrix
- Transpose the matrix
- Normalize for each column
  - Divide each element by the number of total links

Transition probability matrix

Computation (cont'd)

- Compute the eigenvector for maximum eigenvalue

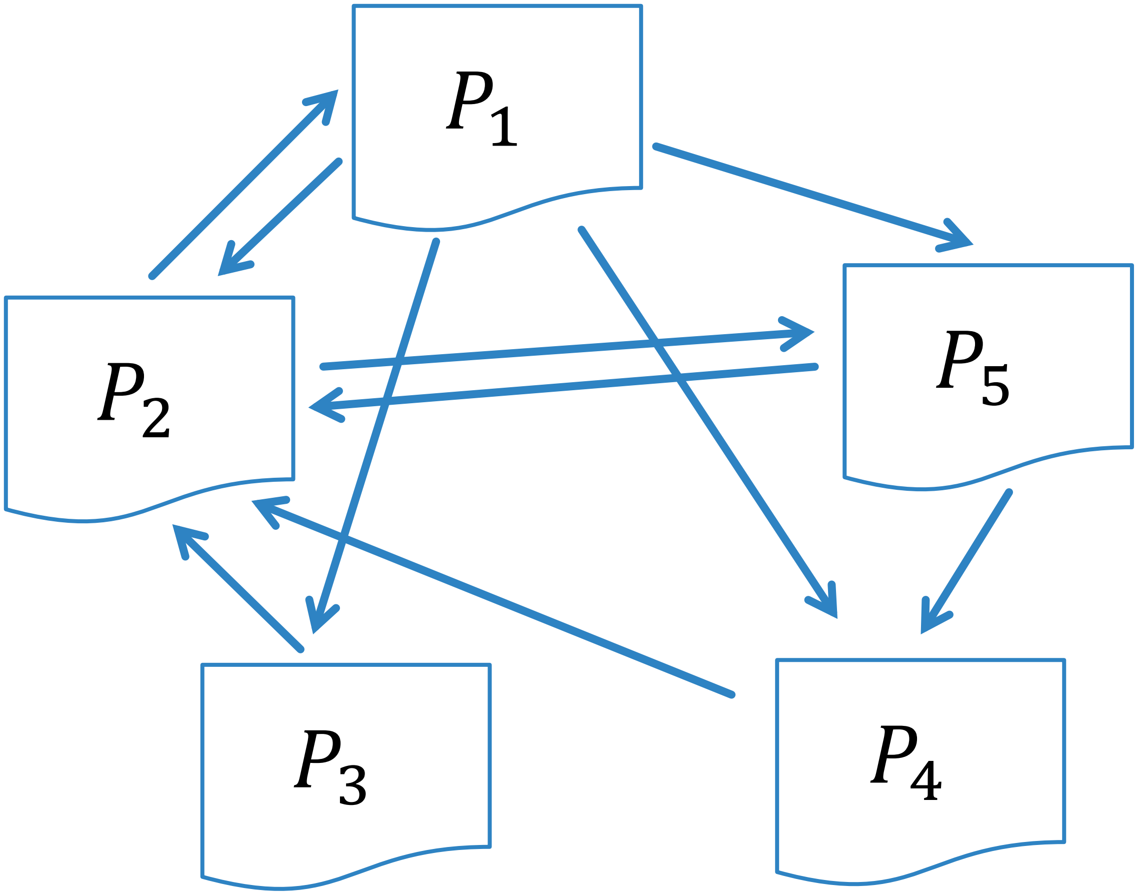
$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

eigenvalue

eigenvector

- Normalize the eigenvector
- i-th element is the score for Page i

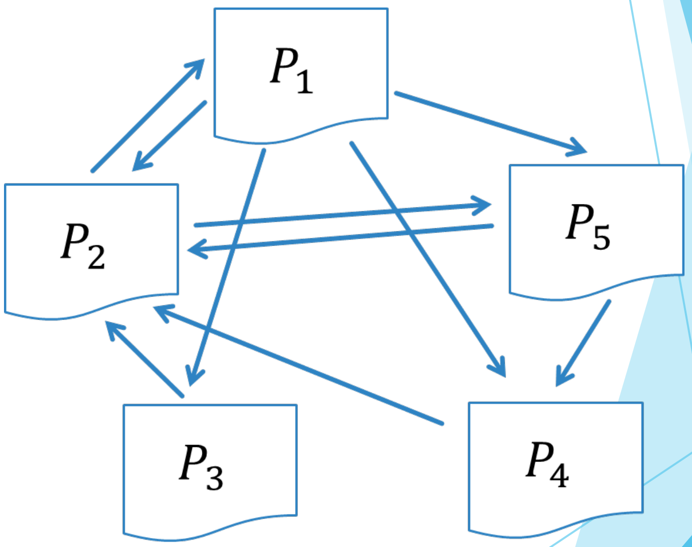
PageRank Example



PageRank Example (cont'd)

- adjacency matrix （隣接行列）

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} P_1 & P_2 & P_3 & P_4 & P_5 \end{array} \\
\begin{array}{c} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{array} &
\begin{pmatrix}
0 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0
\end{pmatrix}
\end{array}
$$



PageRank Example (cont'd)

- transposed adjacency matrix （隣接行列の転置）
  - adjacency matrix

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

  - transposed adjacency matrix

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

PageRank Example (cont'd)

- transition probability matrix
  - transposed adjacency matrix

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

  - transition probability matrix

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

PageRank Example (cont'd)

- Eigenvector for the maximum eigenvalue

$$\begin{pmatrix} 8 & 16 & 2 & 7 & 10 \end{pmatrix}$$

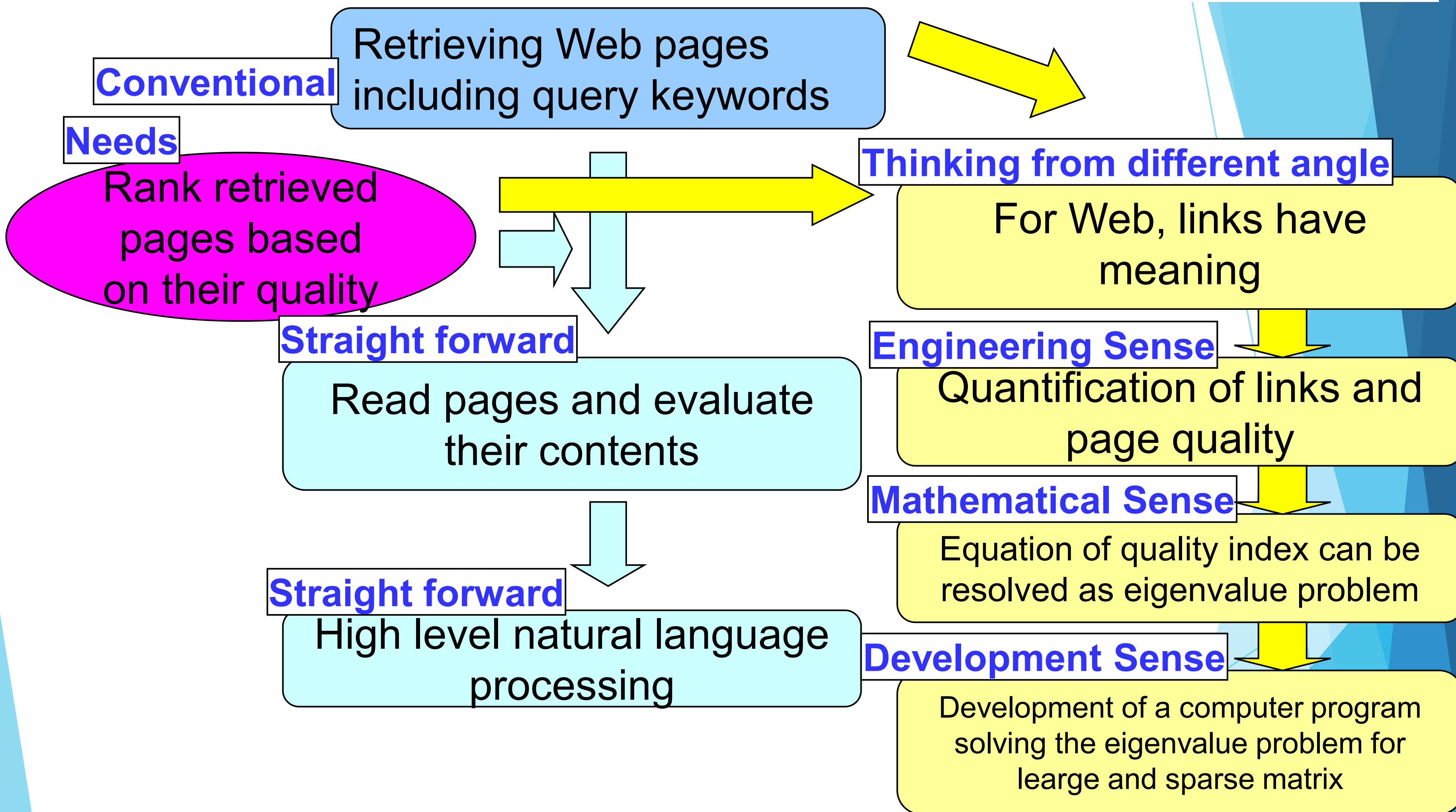
- Normalized eigenvector

$$\begin{pmatrix} \frac{8}{43} & \frac{16}{43} & \frac{2}{43} & \frac{7}{43} & \frac{10}{43} \end{pmatrix}$$

$$\approx \begin{pmatrix} 0.186 & 0.372 & 0.047 & 0.163 & 0.233 \end{pmatrix}$$

$$\boxed{P_1} \quad \boxed{P_2} \quad \boxed{P_3} \quad \boxed{P_4} \quad \boxed{P_5}$$

PageRank Technology : Structure of the Idea



That's it today

Assignment #2 in next class (July 12th)