Web Page Scoring Based on Link Analysis of Web Page Sets

Hitoshi Nakakubo[†], Shinsuke Nakajima[‡]
Kenji Hatano[+], Jun Miyazaki[‡], Shunsuke Uemura*
[†]U-TEC Corporation, Japan
[‡]Nara Institute of Science and Technology, Japan
[+]Doshisha University, Japan
*Nara Sangyo University, Japan

Background

- Web search engine
  - Performance of query processing
  - Retrieval accuracy
    - Link analysis approach
      - cut/information unit [Tajima et al., HT ‘98/Li et al., WWW 2001]
      - PageRank [Page et al., WWW ‘98]
      - HITS [Kleinberg, SODA ‘98]

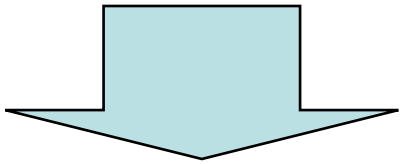
Background

- cut/information unit
  - calculate importance degrees of Web content (=multiple Web pages)

- PageRank/HITS
  - calculate importance degrees of Web page (one Web page) using their hyperlink structure

Problems

- cut/information unit
  - Relativity among Web pages is not considered
    (no guarantee that the Web pages contain one identical topic.)
- PageRank/HITS
  - Relativity of Web contents is not considered
    (no guarantee that Web page unit equal information unit.)



Web pages irrelevant to query keywords are often ranked highly

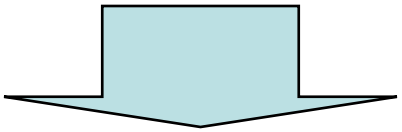
Our Approach

- In order to provide relevant Web pages
  - extracting sets of Web pages containing one identical topic compiled by a unique author
    - by considering relativity among Web pages
  - adopting PageRank algorithm
    - with considering relativity among Web contents

Retrieval accuracy must be improved !!

Web Page Set (WPS)

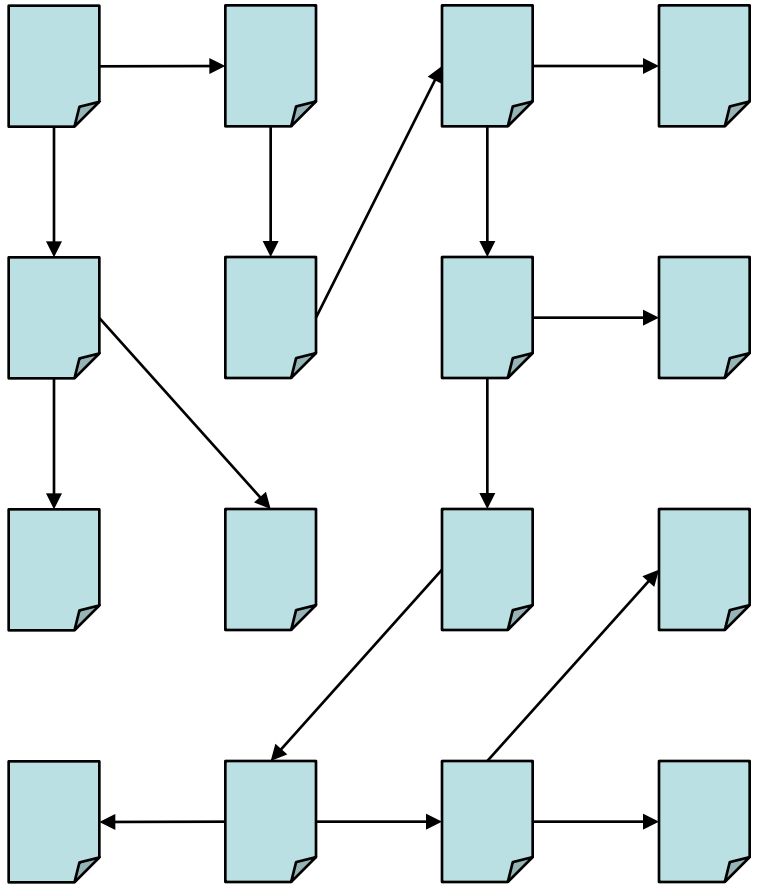
- Web page set (WPS) is
  - compiled by a unique author
    - Quality of a Web page should be homogenized
  - containing one identical topic
    - Importance degree of a Web page should be calculated using one identical topic



We can treat features of Web pages exactly

How to extract WPSs?

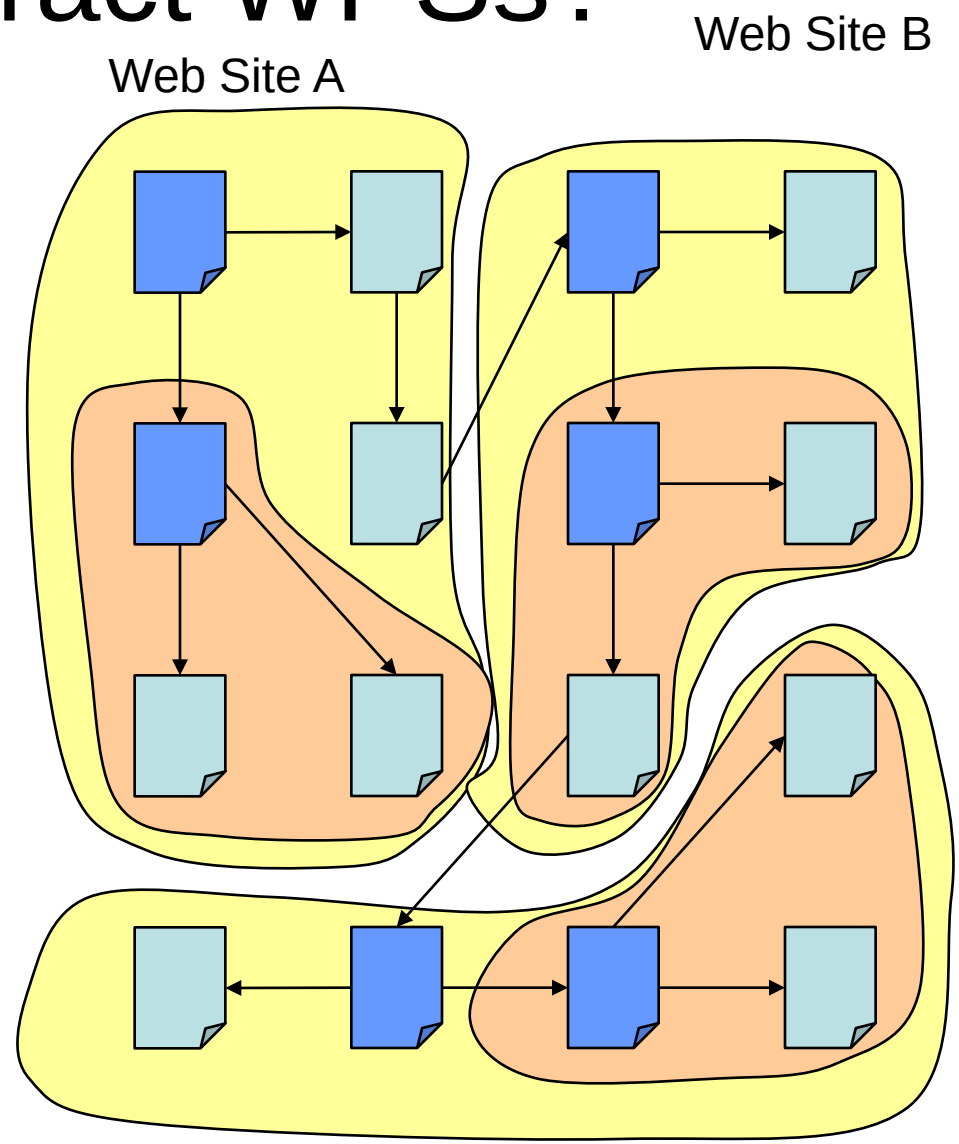
1. extracting Web pages compiled by a unique author
   - Ayan's approach
     1. find entry pages
        - calculate the points of each Web page
          - » URL strings
          - » title of Web page
          - » anchor texts
          - » number of links in the Web page etc.



entry pages

How to extract WPSs?

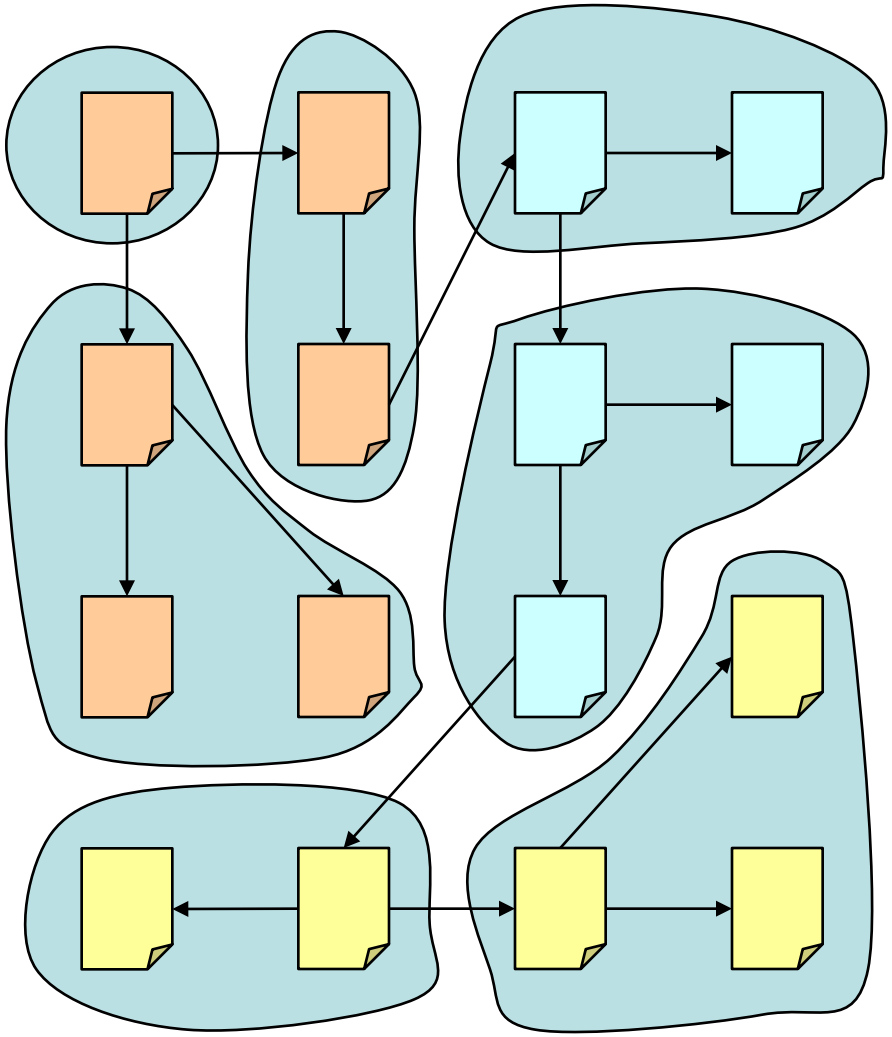
2. determine a boundary of a logical domain
   - an entry page and its descendants are belonging to the same logical domain
   - number of Web pages with in the same logical domain is 10 and above
     - » merged into a parent logical domain



logical domains

How to extract WPSs?

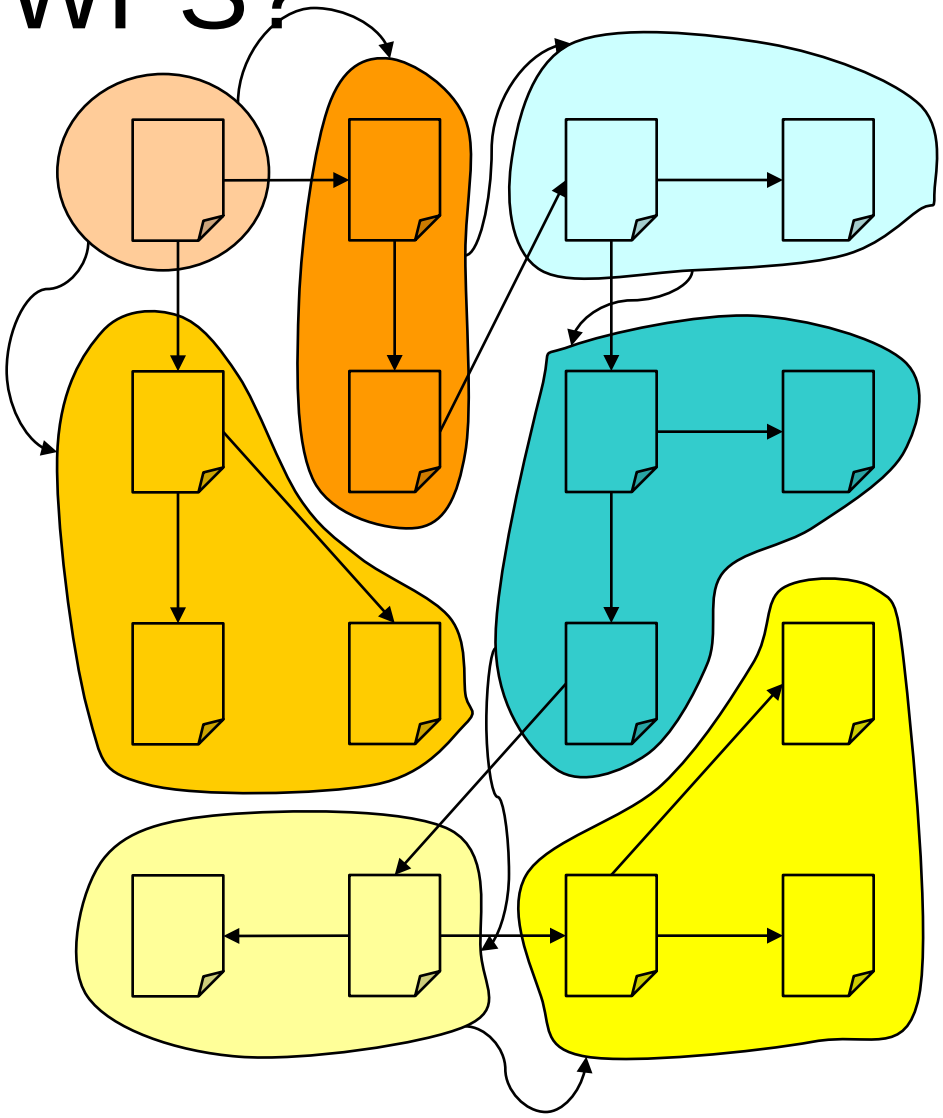
2. determining one identical topic in the same logical domain
   - calculate feature vectors of each Web page
   - apply Ward's method for clustering the Web pages
     - the number of cluster is one tenth of the number of Web pages in each logical domain



WPSs

How to calculate PageRanks of each WPS?

- delete all links among Web pages within the same WPSs
- construct link structures among WPSs
- delete all duplicate links between any two WPSs
- calculate PageRanks of each WPS



Experiments

- Web test collection
  - NW100G-01
    - 100GB (11 million pages)
    - contains mostly English and Japanese pages
    - developed by NTCIR (NII Test Collection for IR) project
- Search topics & relevance judgment
  - NTCIR-4 WEB Info 1
    - categorizes 4 relevance levels (highly relevant, relevant, partially relevant, irrelevant)

Evaluation Measures (1)

- Discounted Cumulated Gain (DCG)

  [Jarvelin, Kekalainen 2000]

  - relevance measure taking account of multiple valued relevance levels

$$dcg(i) = \begin{cases} g(1) & \text{if } i = 1 \\ dcg(i-1) + \dfrac{g(i)}{\log i} & \text{otherwise} \end{cases}$$

$$g(i) = \begin{cases} h \text{ if } d(i) \in H \text{ (highly relevant)} \\ a \text{ if } d(i) \in A \text{ (relevant)} \\ b \text{ if } d(i) \in B \text{ (partially relevant)} \end{cases}$$

Evaluation Measures (2)

- Weighted Reciprocal Rank (WRR)

  [Eguchi et al. 2003]

  – extension of Mean Reciprocal Rank (MRR) [Voorhees 1999] to multiple valued relevance levels

$$mrr = AVG\left(\frac{1}{\text{rank of the first appeared relevant document}}\right)$$

$$wrr(m) = \max(r(i))$$

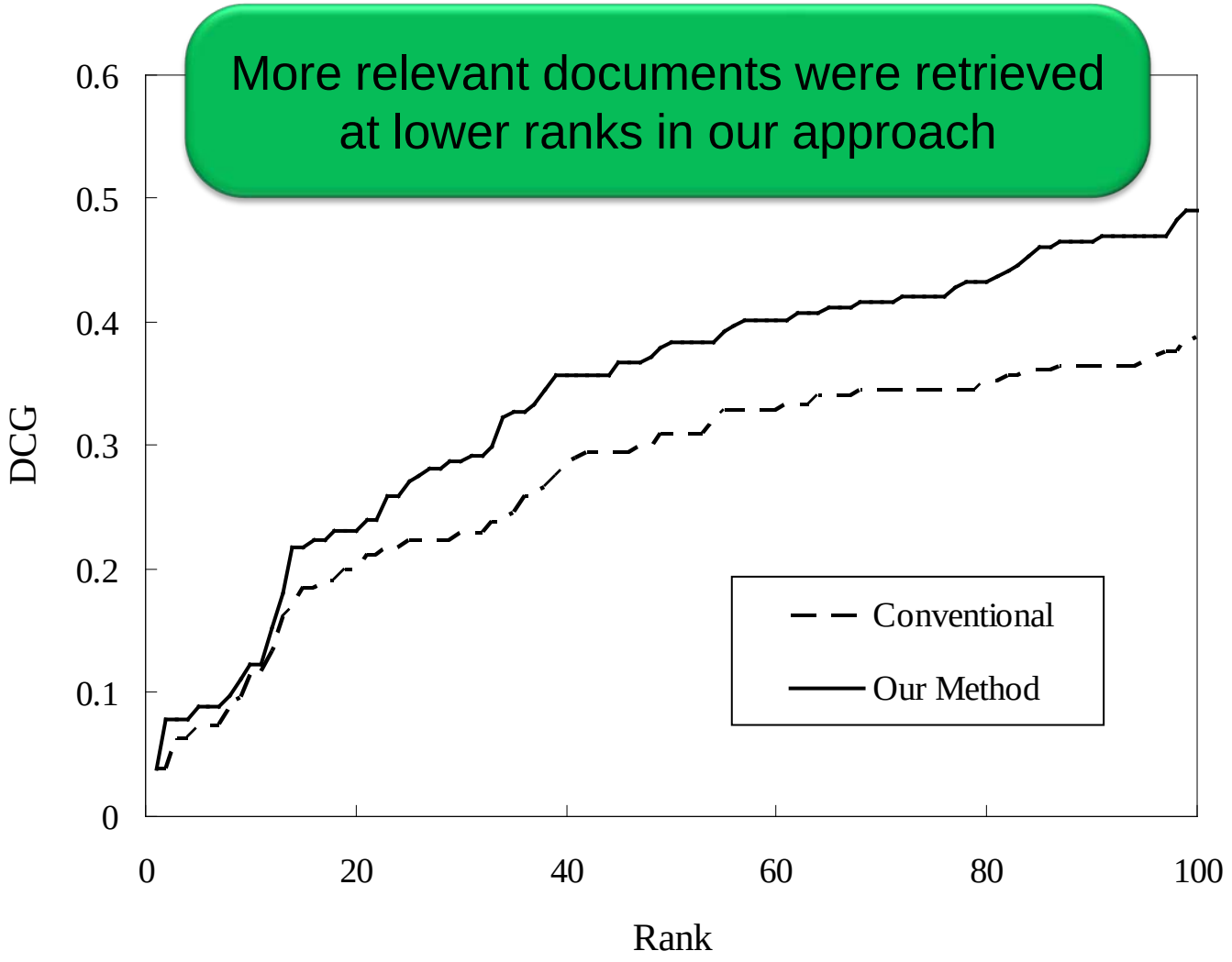
$$r(i) = \begin{cases} \delta_h / (i - 1/\beta_h) & \text{if } d(i) \in H \text{ and } 1 \le i \le m \\ \delta_a / (i - 1/\beta_a) & \text{if } d(i) \in A \text{ and } 1 \le i \le m \\ \delta_b / (i - 1/\beta_a) & \text{if } d(i) \in B \text{ and } 1 \le i \le m \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } \delta \in \{0,1\}, \beta_b \ge \beta_a \ge \beta_h > 1$$

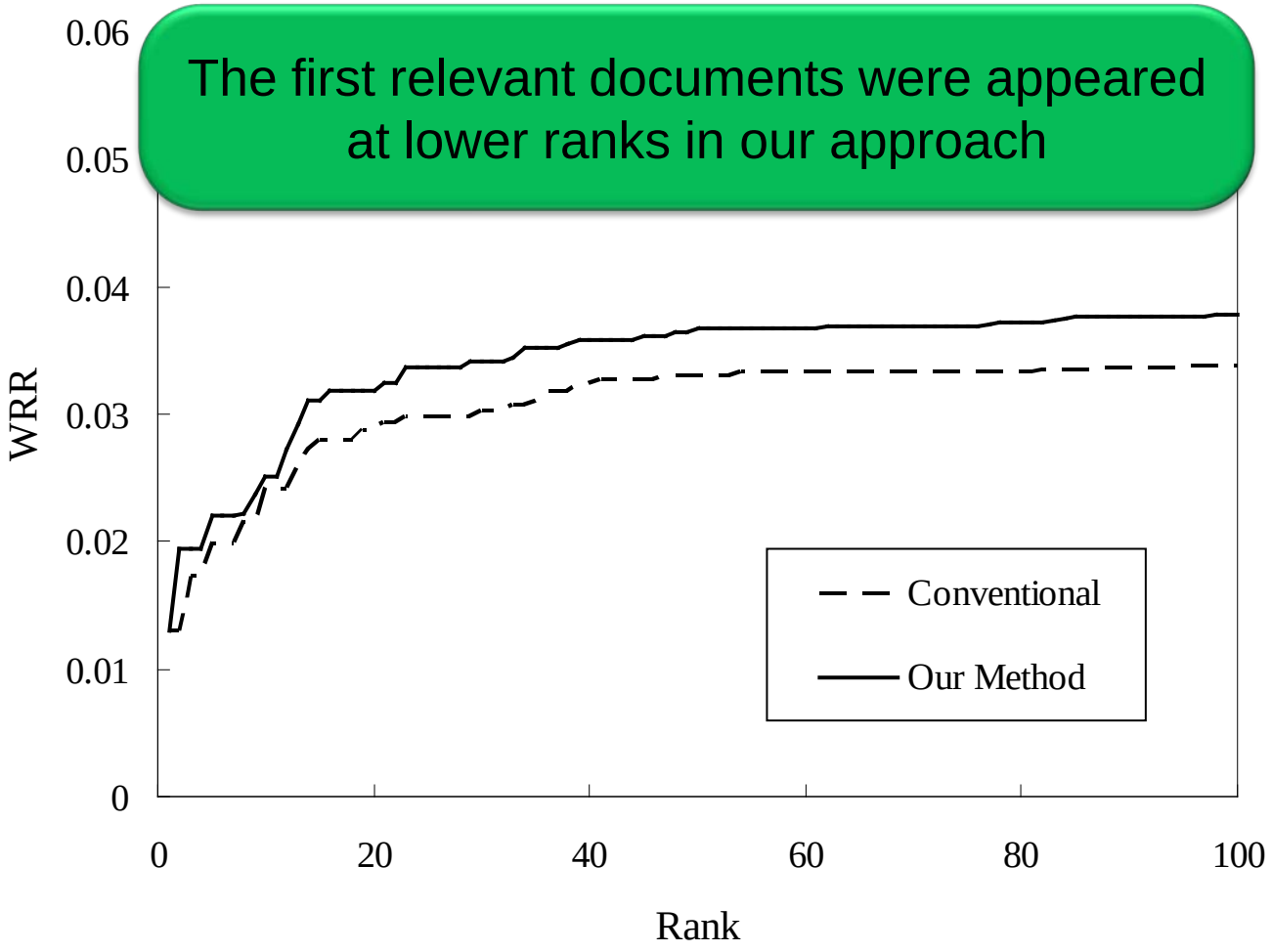
Parameters

- the size of logical domain
  - more than or equal to 10 pages
- the number of WPSs
  - 1/10 of total Web pages
- DCG
  - weight for relevance: $(h, a, b) = (3,2,0)$
- WRR
  - $(\delta_h, \delta_a, \delta_b) = (1,1,0),\ (\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty),\ m = 100$

Evaluation by DCG



Evaluation by WRR



Conclusion

- proposed a new Web page scoring based on the notion of Web Page Set (WPS)
  - better accuracy than conventional ones w.r.t. DCG and WRR evaluation measures

Future Work

- more discussion of the notion of WPS
  - compare possible variations of WPS
- improvement of scoring
  - better (optimal) WPS size and # of clusters
  - better (optimal) distribution of page scores inside WPSs

Danke Schön!