Hierarchical Organization of Web Documents based on Hypertext Classification

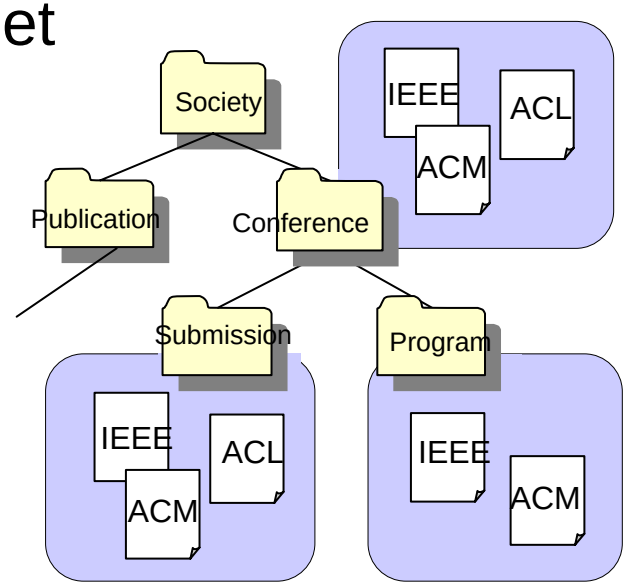
Yusuke Suzuki, Shigeki Matsubara
and Masatoshi Yoshikawa
Nagoya University, Japan

Introduction（1/2）

- Many sites containing common information
  - Academic societies, Universities, Internet service providers …
- When users want to browse particular pages on several sites
  - deadlines for paper submissions to academic conferences
  - service contents provided by ISPs

  they have to bother to seek the target pages from each site

Need for Web directories to organize pages found in several sites based on contents



Introduction（2/2）

- Present state of Web directories
  - Manual design and management

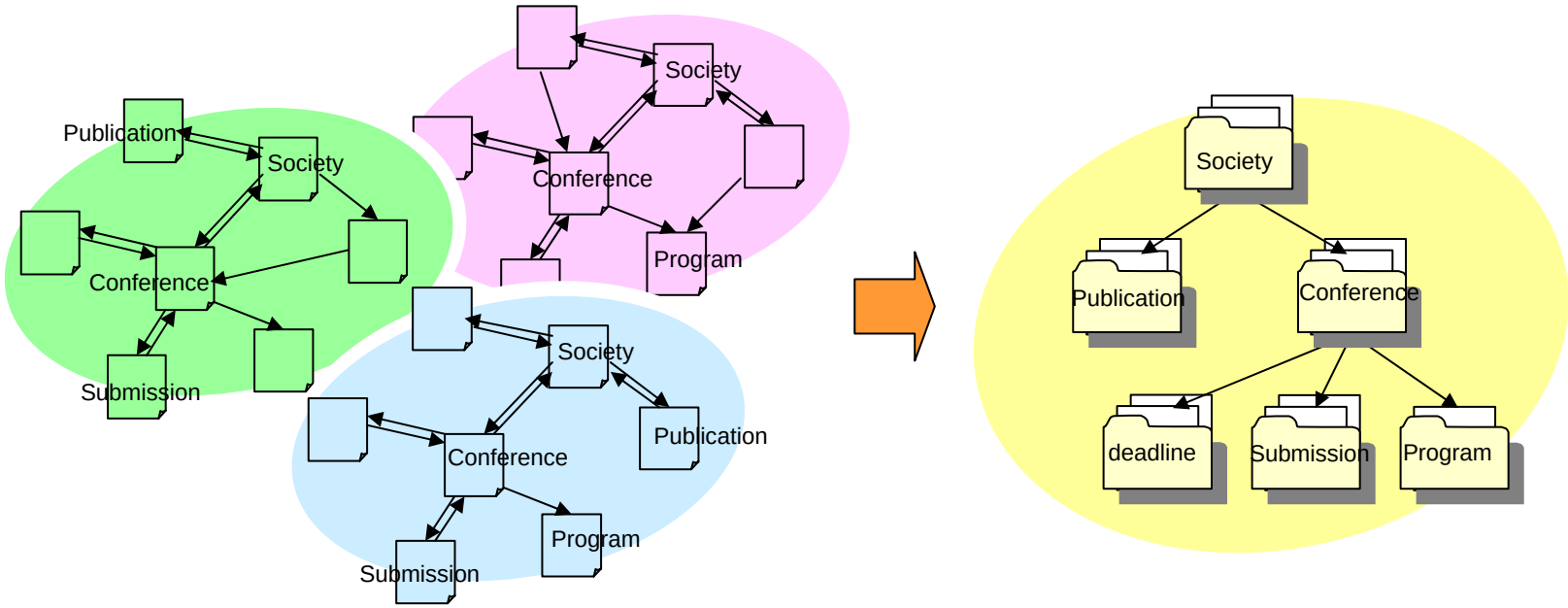
Costly processes
- Design of directory structures for each category
- Categorization of pages into directories
- Response to frequent page updates

- As the size of a directory or data increases, the cost of manually designing and managing the Web directory grows

Purpose

Proposal of a technique for automatically constructing hierarchical Web directories

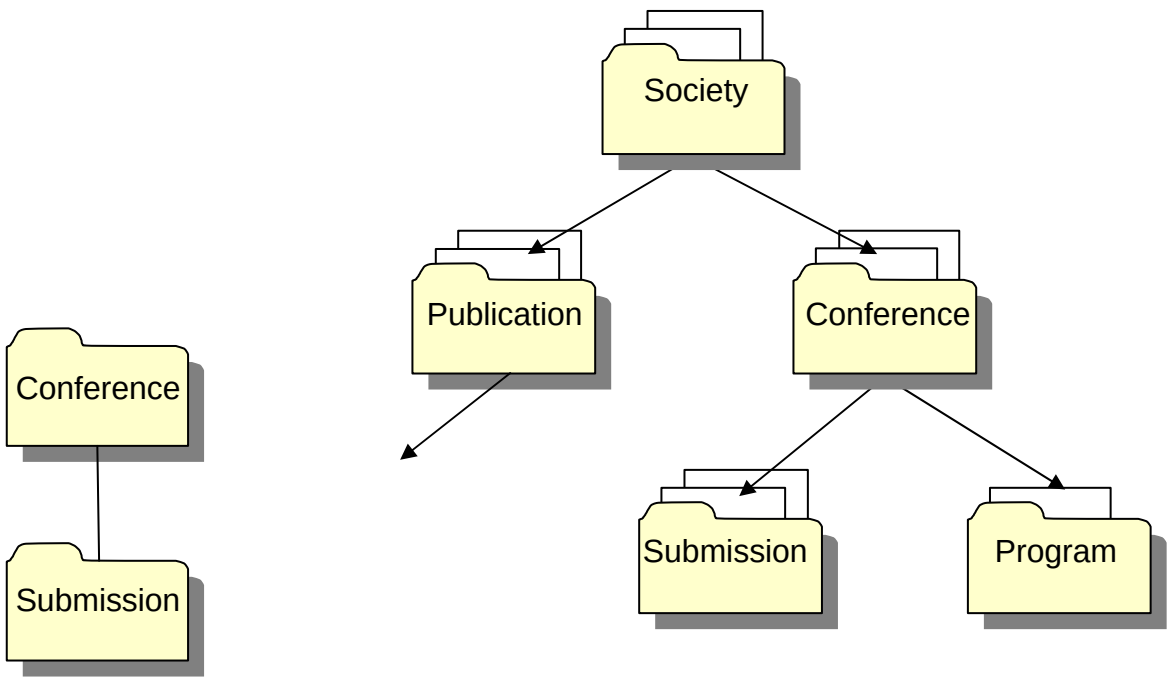
- Putting pages with same contents from several sites into a directory and providing hierarchical Web directories



Basic Ideas（1/3）

To construct hierarchical Web directories

- Find super-sub relations between directories
- Categorize Web pages into directories

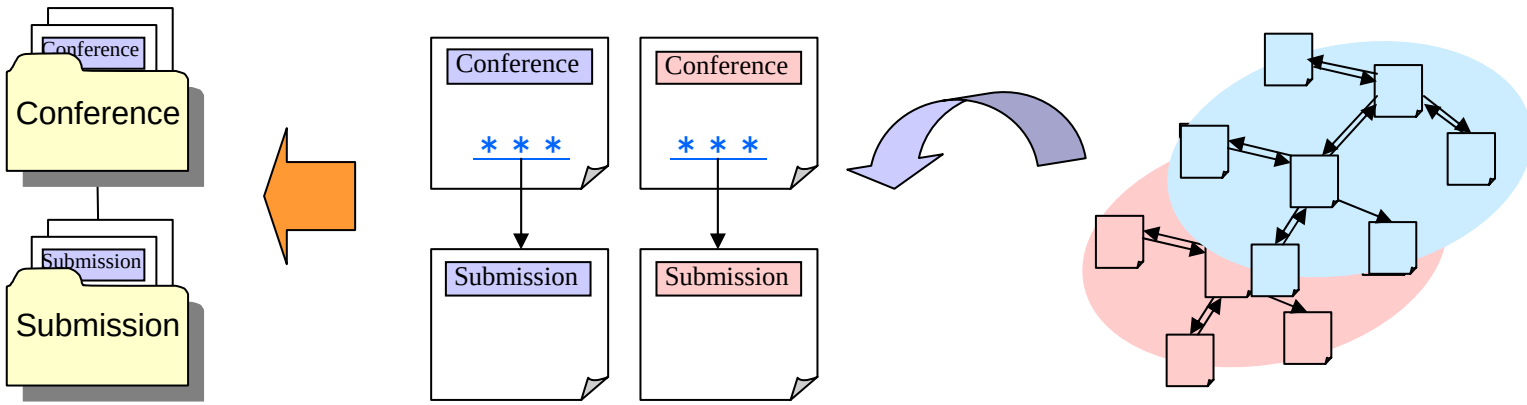


Basic Ideas（1/3）

To construct hierarchical Web directories
- Find super-sub relations between directories
- Categorize Web pages into directories

- A feature of the Web
  - A relation between Web pages is represented by a hyperlink

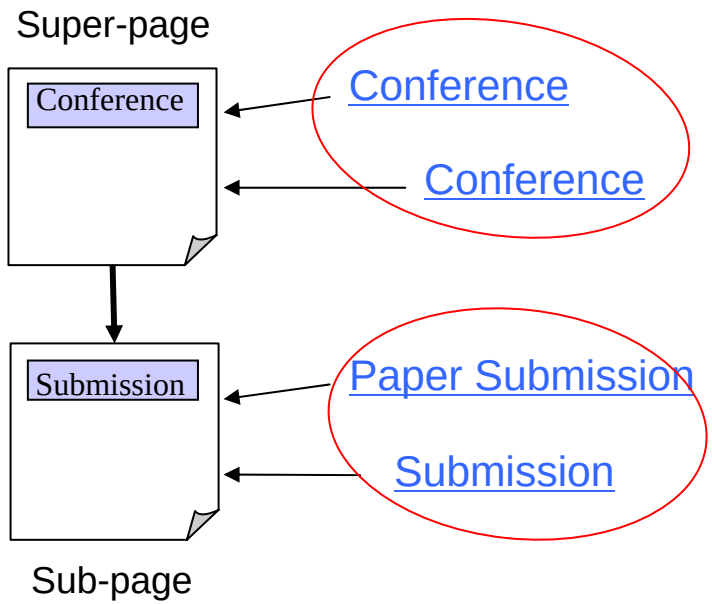
Extract pages with a super-sub relation as a page-pair based on the hyperlinks and replace its relation with a relation between directories



Basic Ideas（2/3）

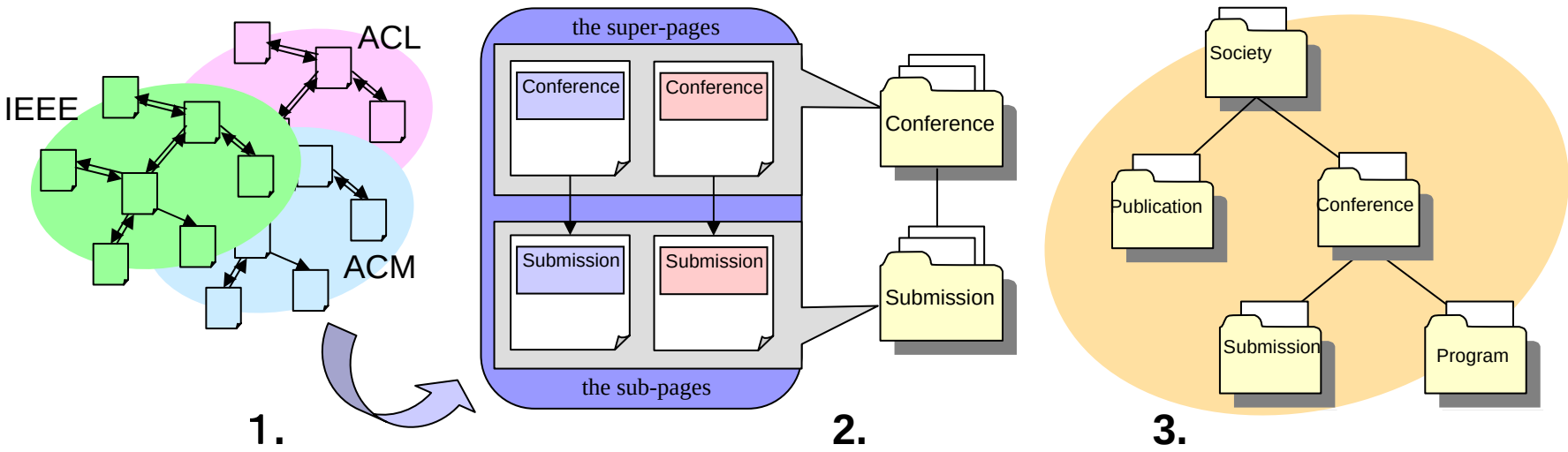
- Representation of pages of a super-sub relation
  - Anchor text
    - Set by creator in order to navigate users to a linked page
    - ⇒ A description representing in brief the whole contents of the linked page

We represent each page of the super-sub relation using the anchor texts



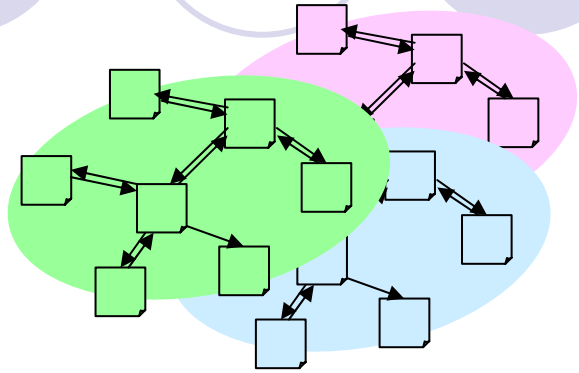
Basic Ideas（3/3）

- Constructing the directory structures
  1. Extract the super-sub relation between the Web pages from each site
  2. Cluster the common super-sub relations and replacing their relation with a super-sub relation between the directories
  3. Construct a Web directory by iterating the integration of the directories



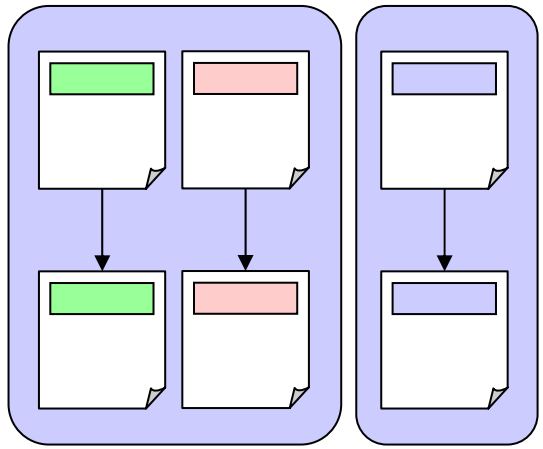
Process of the Proposed Method

Several sites



（1）Extracting super-sub relations

（2）Clustering super-sub relations



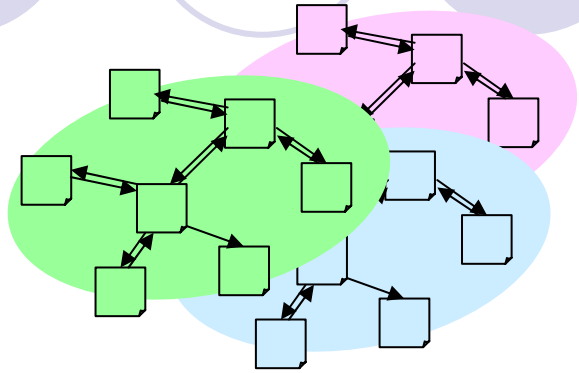
（3）Constructing directory structures

（4）Deciding directory names

Hierarchical Web directory

Process of the Proposed Method

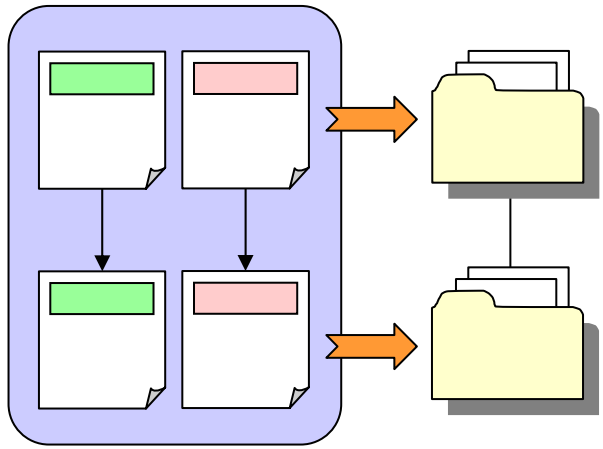
Several sites



（1）Extracting super-sub relations

（2）Clustering super-sub relations

（3）Constructing directory structures

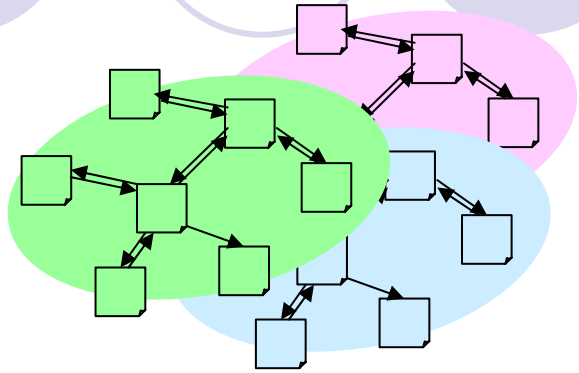


（4）Deciding directory names

Hierarchical Web directory

Process of the Proposed Method

Several sites

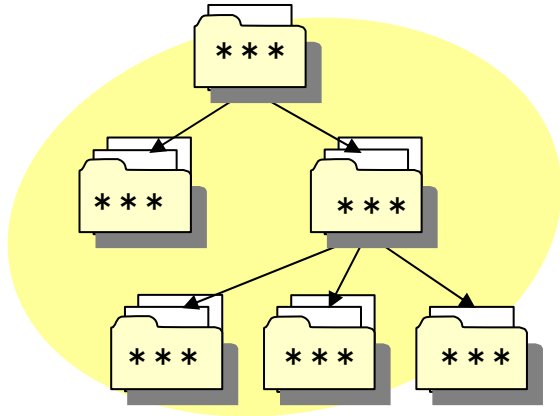


（1）Extracting super-sub relations

（2）Clustering super-sub relations

（3）Constructing directory structures

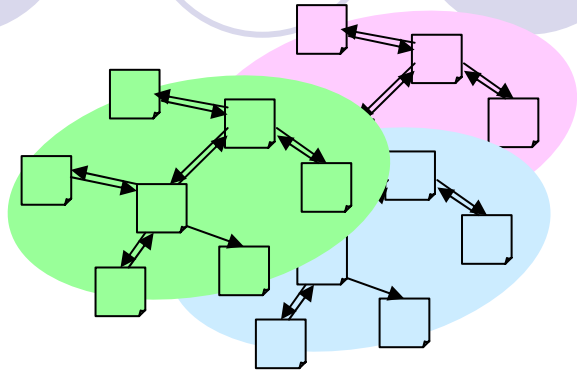
（4）Deciding directory names



Hierarchical Web directory

Process of the Proposed Method

Several sites



（1）Extracting super-sub relations

（2）Clustering super-sub relations

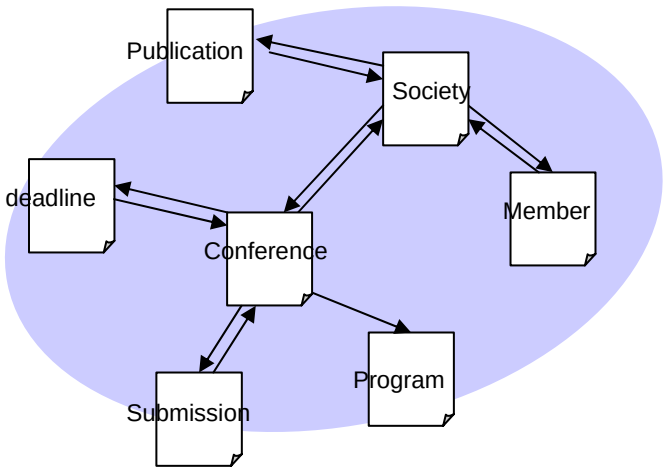
（3）Constructing directory structures

（4）Deciding directory names

Hierarchical Web directory

Extracting Super-sub Relations（1/4）

- Necessary to identify the links connecting a super-page and a sub-page
  - Not all Web pages connected by hyperlinks necessarily have a super-sub relation
- Website creators organize pages into folders and locate them on the server



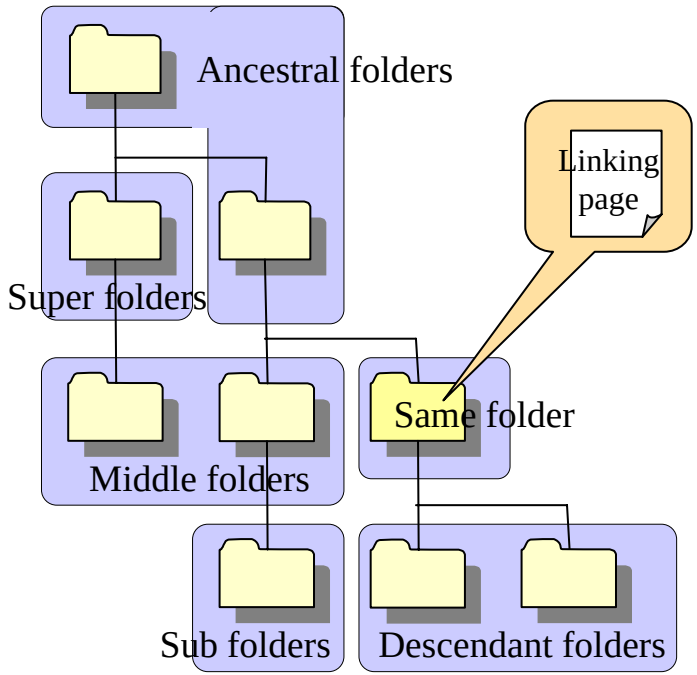
To identify a super-sub relation,
we utilize the creators' knowledge

Extracting Super-sub Relations（2/4）

- Identification of the links connecting a super-page and a sub-page
  - We investigated the relevance between a page's location on a server and the links

Investigation method

1. Extract 200 links from each of four sites
2. Judge whether the linking page and the linked page have a super-sub relation
3. Classify the links into six relative locations and investigate the rate of a super-sub relation in each relative location

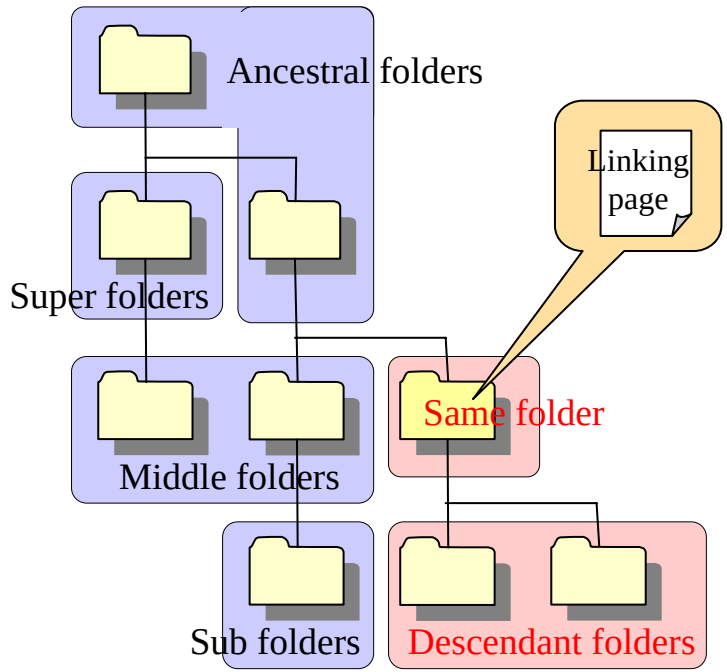


Extracting Super-sub Relations（3/4）

- Result of the investigation

Location of linked pages	Link	Rate of Super-sub relation（%）
Descendant folder	136	91.9
Ancestral folder	151	0.7
Same folder	246	58.1
Sub folder	3	0
Super folder	77	2.6
Middle folder	152	2.7

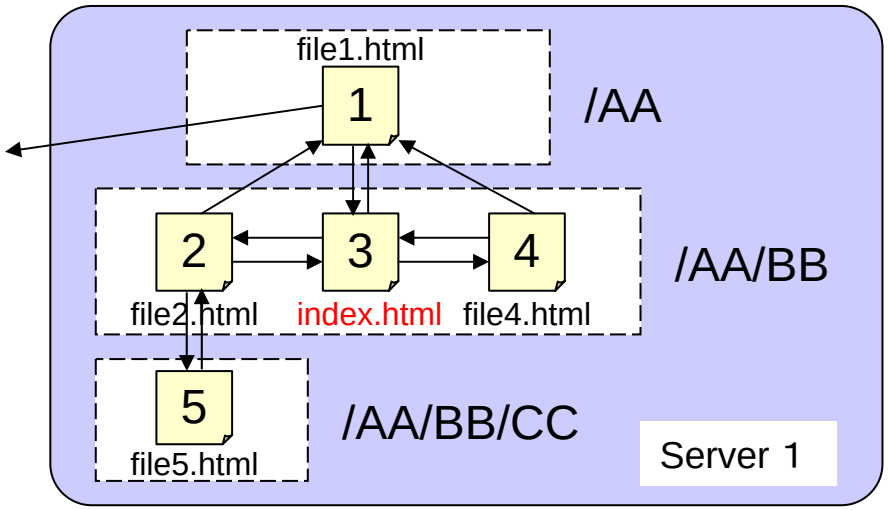
The case that the linking page is “index.html”	41	85.3%

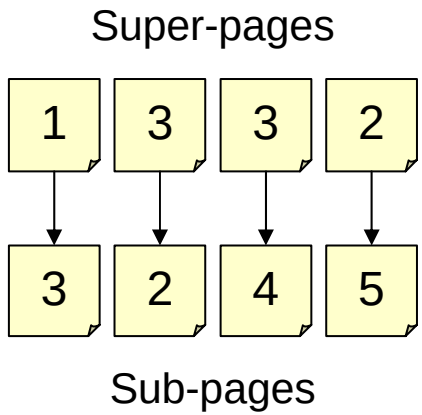


Extracting Super-sub Relations（4/4）

Decision rule for a super-sub relation

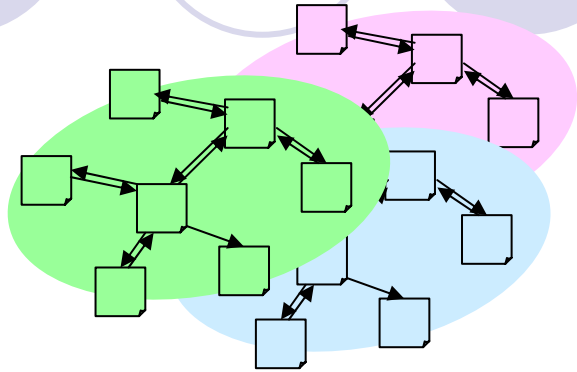
1. Link to the page on the same server
2. Link to the page in the descendant folders or the same folder
3. In the case of the same folder, the linking page is “index.html”
   (Without “index.html,” the linking page is a page which links to the most pages in the same folder）





Process of the Proposed Method

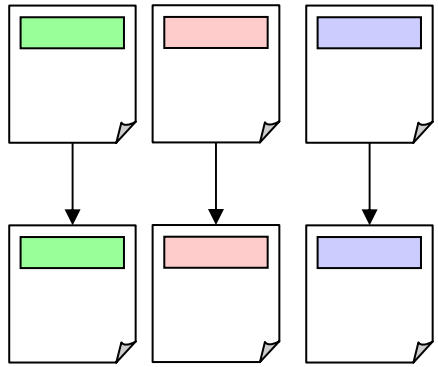
Several sites



（1）Extracting super-sub relations

（2）Clustering super-sub relations

（3）Constructing directory structures



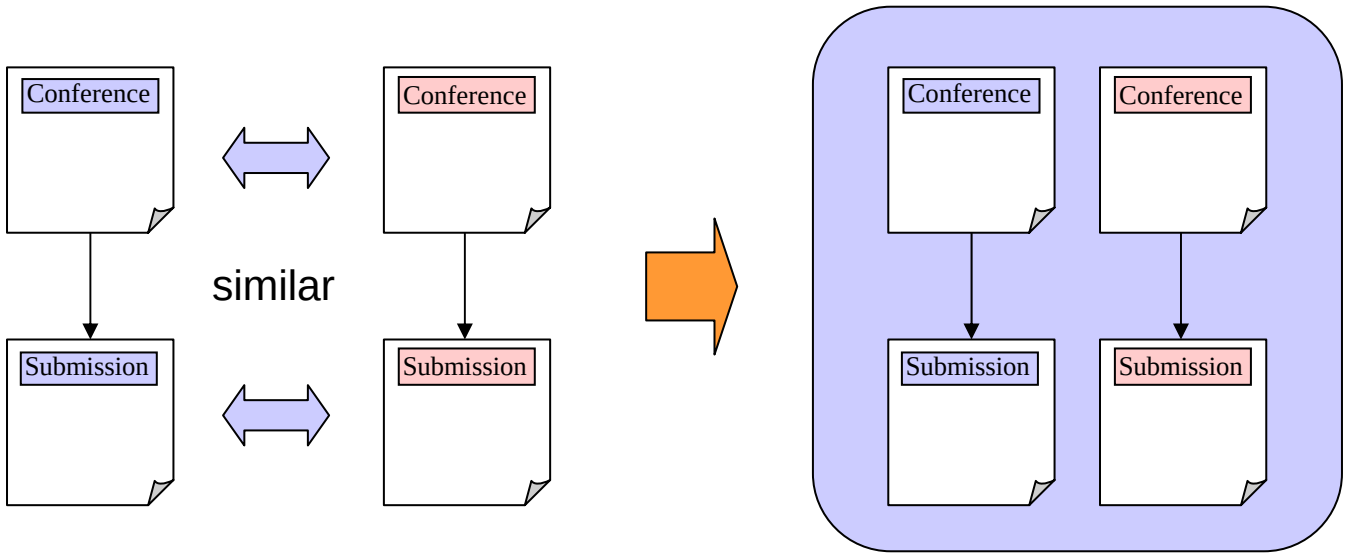
（4）Deciding directory names

Hierarchical Web directory

Clustering Super-sub Relations（1/3）

- The common super-sub relations are clustered
  - Common super-sub relation

    The super-sub relation where both the contents of the super-pages and the contents of the sub-pages are similar
  - In the clustering result, if the number of members in a cluster is below a threshold value, that cluster is excluded



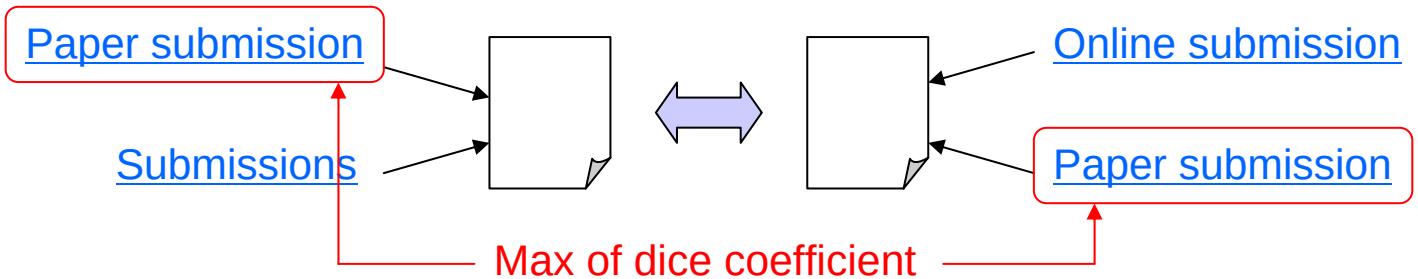
Clustering Super-sub Relations（2/3）

- The similarity between Web pages
  - The maximal Dice coefficient value between the anchor texts linking to each page is adopted as the similarity between the pages

$$sim(d_i, d_j) = \max_{1 \le s \le m, 1 \le t \le n} \left(\frac{2M_{i_s j_t}}{M_{i_s} + M_{j_t}} \right)$$

M_{i_s} : number of nouns in the anchor text $a_{i_s}\ (1 \le s \le m)$ which links to the page d_i

$M_{i_s j_t}$: the number of nouns common to anchor text $a_{i_s}\ (1 \le s \le m)$ and $a_{j_t}\ (1 \le t \le n)$

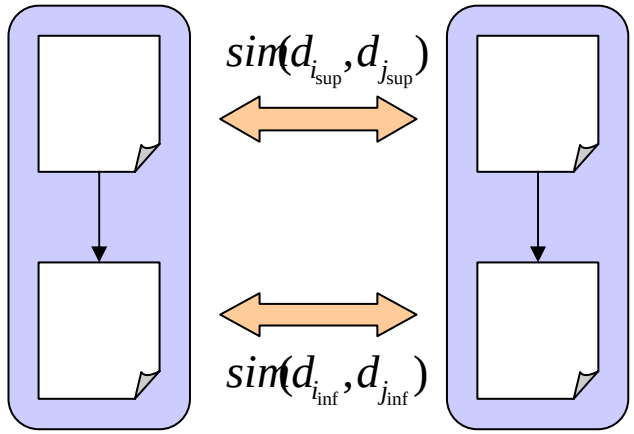


Clustering Super-sub Relations（3/3）

- Clustering method (hierarchical method)
  - Integrating clusters
    1. Both the similarity between the super-pages and between the sub-pages exceeds the threshold value
    2. The average of their similarities is maximal
  - The similarity between the clusters
    - Complete linkage method

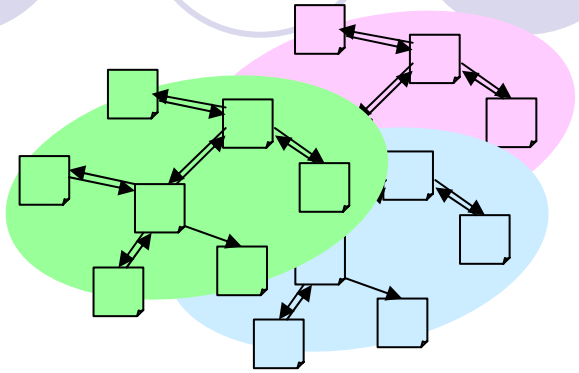
The similarity between the super-sub relations

$$sim(p_i, p_j) = (sim(d_{i_{sup}}, d_{j_{sup}}) , sim(d_{i_{inf}}, d_{j_{inf}}))$$



Process of the Proposed Method

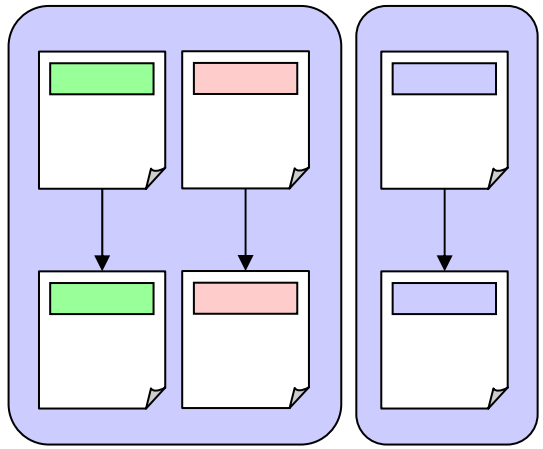
Several sites



（1）Extracting super-sub relations

（2）Clustering super-sub relations

（3）Constructing directory structures

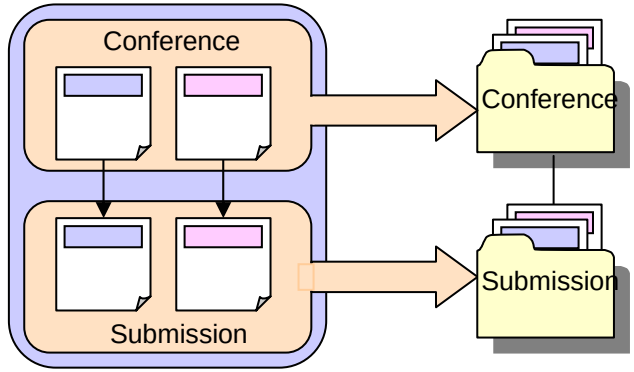


（4）Deciding directory names

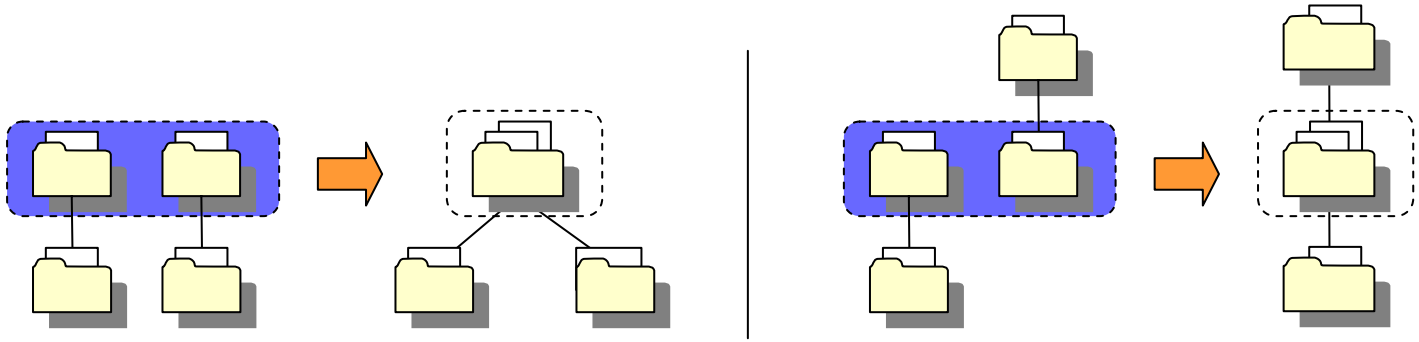
Hierarchical Web directory

Constructing Hierarchical Structures（1/3）

1. Clustered super-sub relations are replaced by the super-sub directory structure



2. Construct the directory structures by iterating the integration of each directory
   - Integration of super-directories
   - Integration of a super-directory and a sub-directory

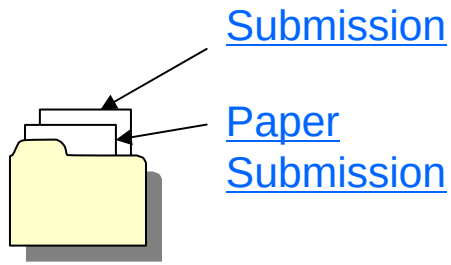


Constructing Hierarchical Structures（2/3）

- Representation of a directory
  - Each directory is represented as a feature vector

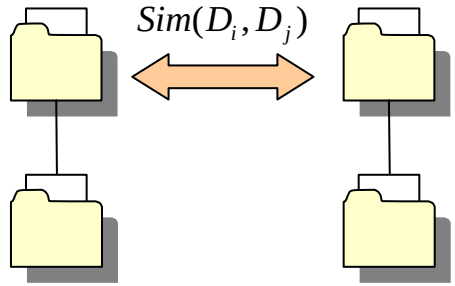
$$\vec{x}_i = (w_{i1}, w_{i2}, \ldots, w_{iN}) \qquad w_{ij} = F_{ij}$$

F_{ij} : the frequency of a noun e_j in a set A_i of the anchor texts which links to the pages in a directory



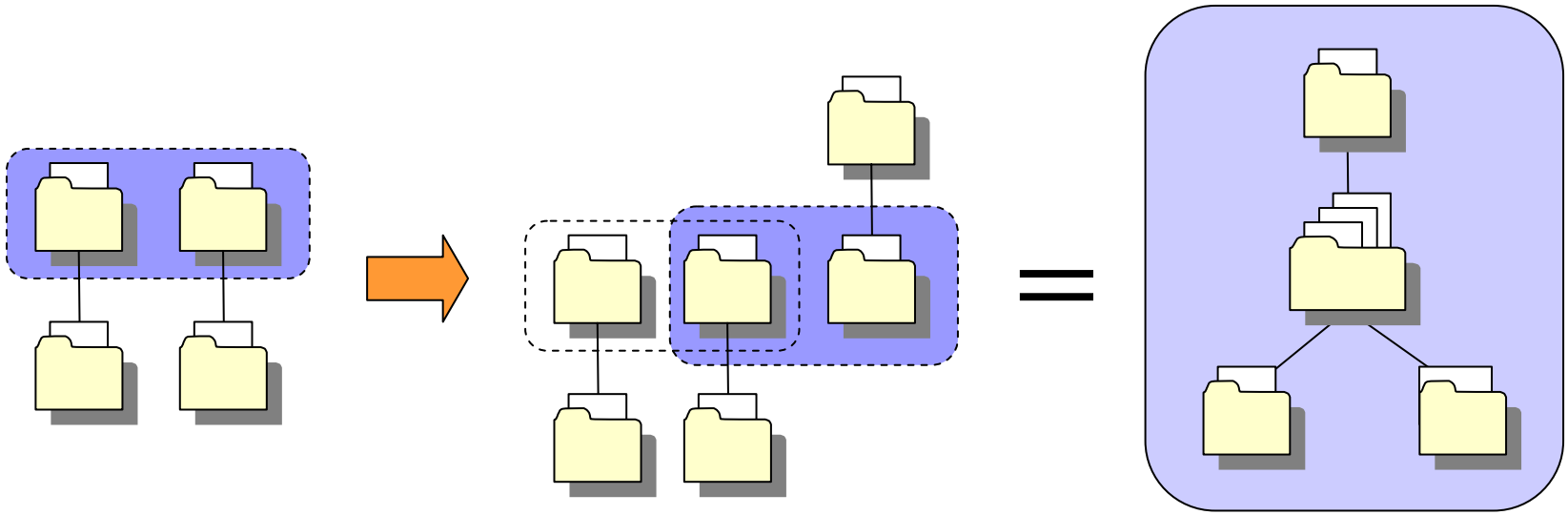
- The similarity between directories
  - Cosign of the feature vectors

$$Sim(D_i, D_j) = \frac{\vec{x}_i \bullet \vec{x}_j}{\left|\vec{x}_i\right|\left|\vec{x}_j\right|} \qquad (i \neq j)$$



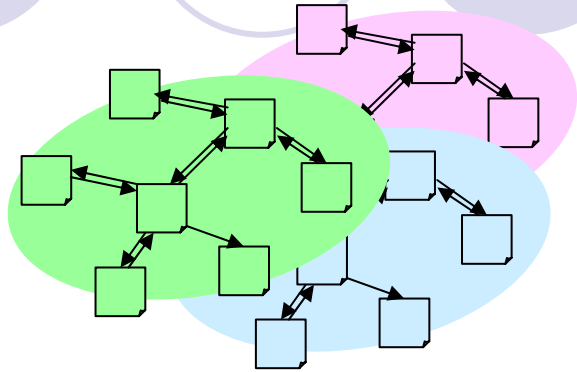
Constructing Hierarchical Structures（3/3）

- Integration of the directories
  - Integrate the directories to satisfy the nature of a tree structure in descending order of similarity
    - When the maximal similarity is less than a threshold value, clustering is stopped



Process of the Proposed Method

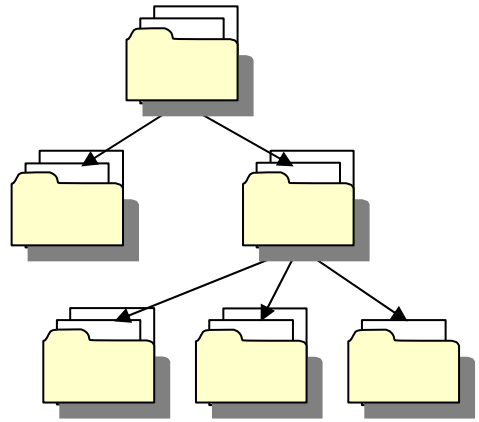
Several sites



（1）Extracting super-sub relations

（2）Clustering super-sub relations

（3）Constructing directory structures

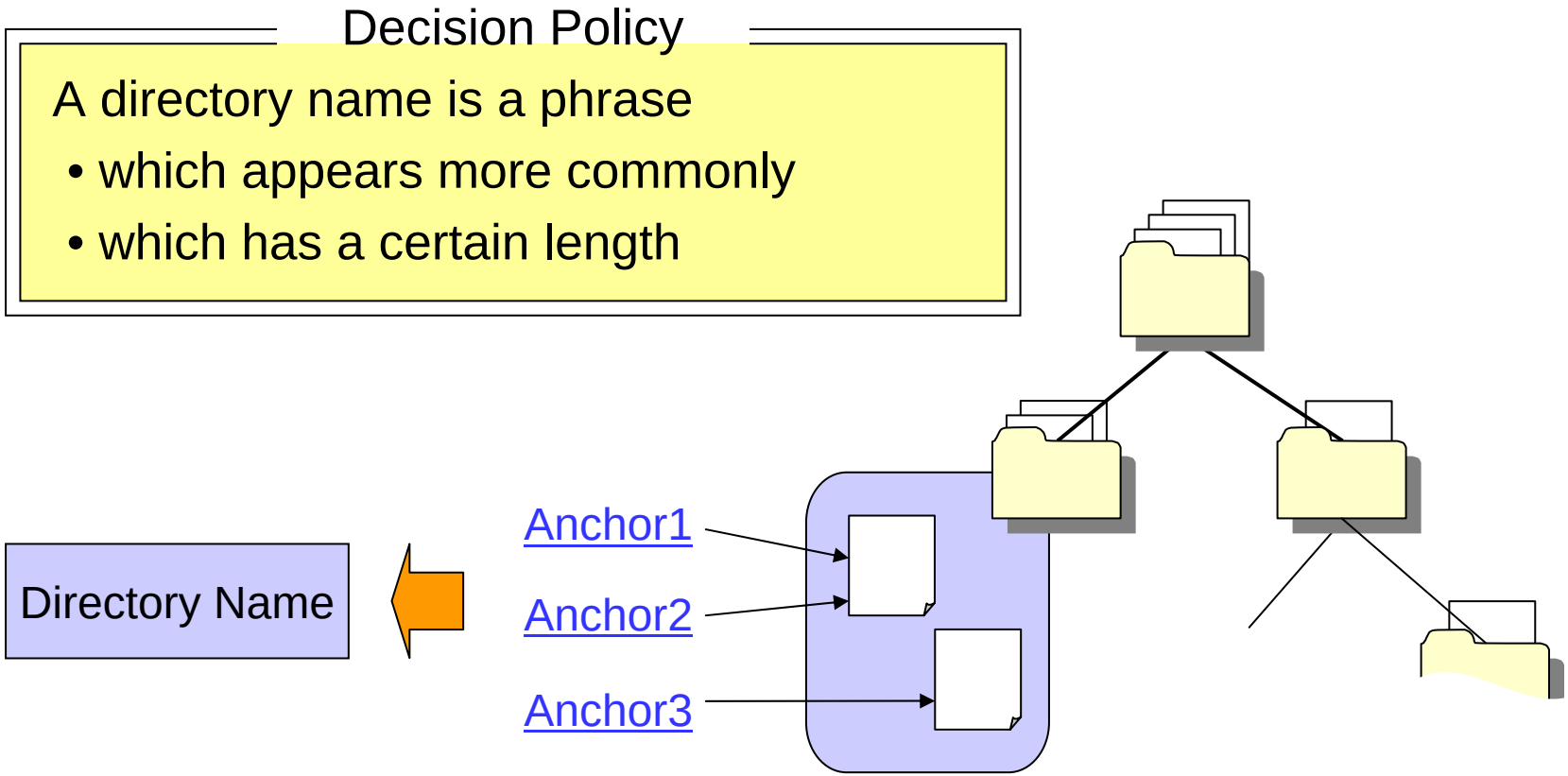


（4）Deciding directory names

Hierarchical Web directory

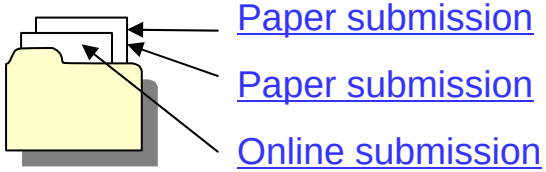
Deciding Directory Names (1/2)

- Each directory name is decided based on a set anchor texts linking to the pages in the directory



Deciding Directory Names (2/2)

- Decision method
  1. Extracting any morphological sequence s_{ij} from a set of anchor texts in each directory
  2. For each s_{ij}, calculate the inclusion rate for each anchor text
  3. Make s_{ij} whose average value is maximal the directory name



any morphological sequences

{ paper, online, submission, paper submission, online submission }

Directory Name
「Paper submission」

The inclusion rate

$$Cover(s_{ij}, a_{ik}) = \frac{F^{i}_{jk}}{|a_{i_k}|}$$

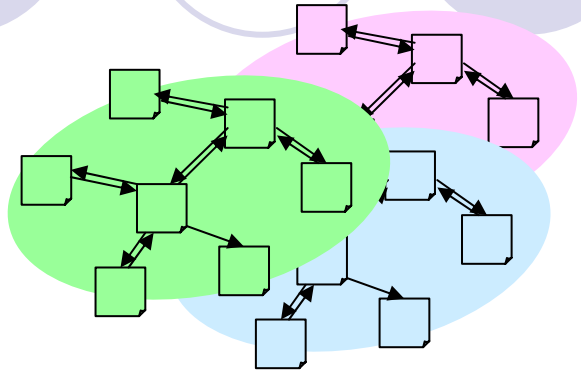
$|a_{i_k}|$: number of morphemes in the anchor text a_{ik} in a directory D_i

F^{i}_{jk} : number of common morphemes in s_{ij} and a_{i_k}

（※ iff a_{i_k} is included s_{ij} . Otherwise 0）

Process of the Proposed Method

Several sites

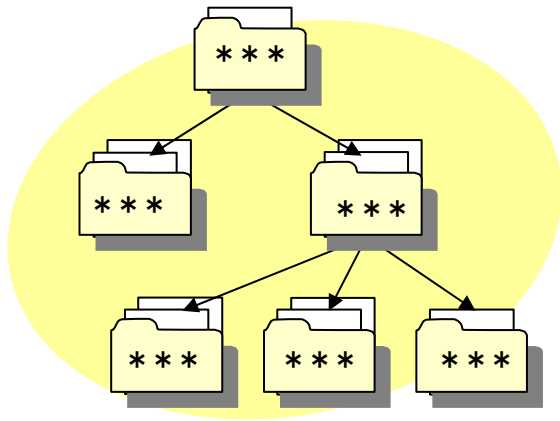


（1）Extracting super-sub relations

（2）Clustering super-sub relations

（3）Constructing directory structures

（4）Deciding directory names



Hierarchical Web directory

Experiments

- Experimental data
  - Sites of graduate schools at Nagoya University

ID	Site	Pages	Links in same server
Ⅰ	Engineering (www.engg.nagoya-u.ac.jp)	126	276
Ⅱ	Environmental Studies (www.env.nagoya-u.ac.jp)	281	1,192
Ⅲ	Information Science (www.is.nagoya-u.ac.jp)	106	267
Ⅳ	Science (www.sci.nagoya-u.ac.jp)	280	887
Ⅴ	Economics (www.soec.nagoya-u.ac.jp)	605	3,288

- Parameter settings
  - Threshold values
    - Similarity in the clustering 0.5
    - Similarity in the construction of directory structure 0.6
    - Minimum number of members in a cluster 2

Experimental Results

- Sample output of the system

A list of the root directory of the generated directory structures

…13 directory structures

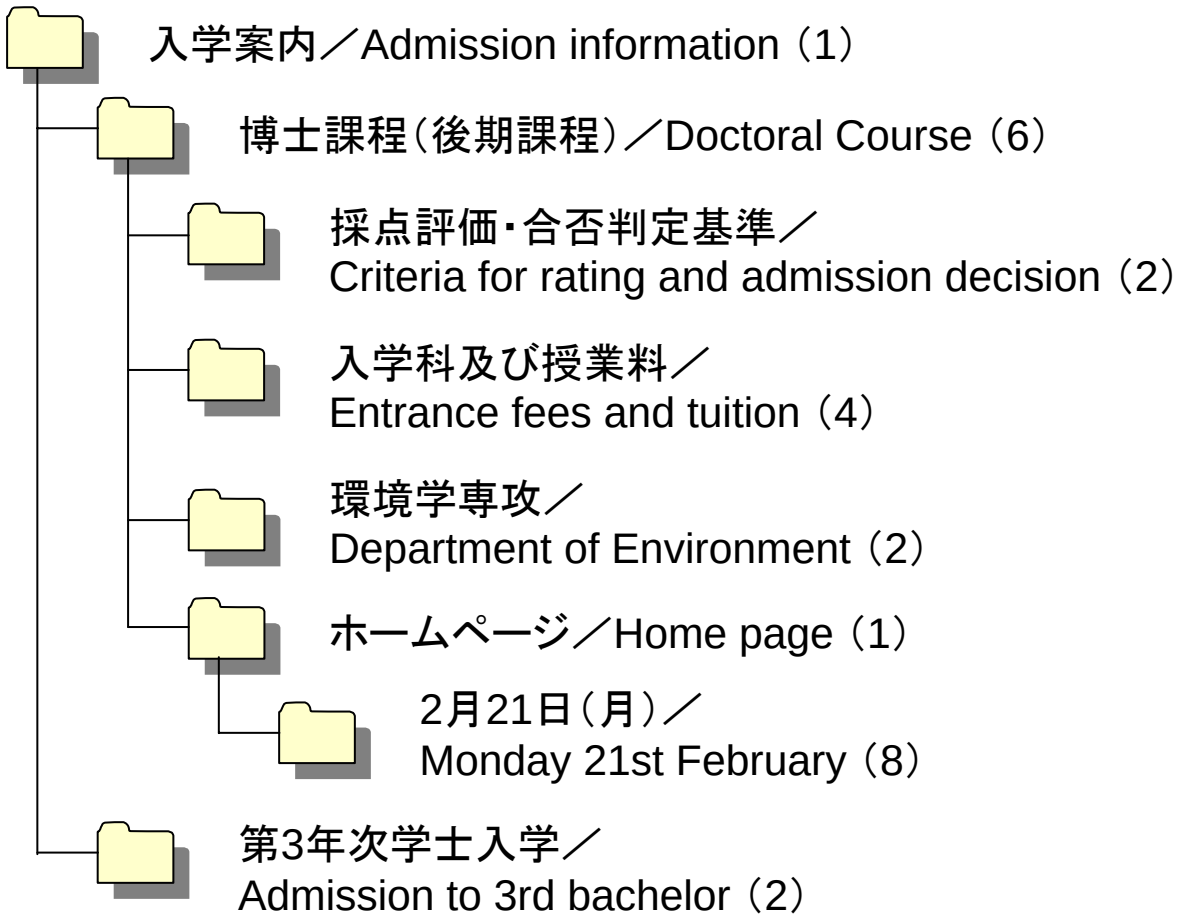
An overview of the directory structure

A collection of links to the pages in the directory

http://plum.itc.nagoya-u.ac.jp/auto_directory/demo/

Experimental Results

- The directory structure : sample 1

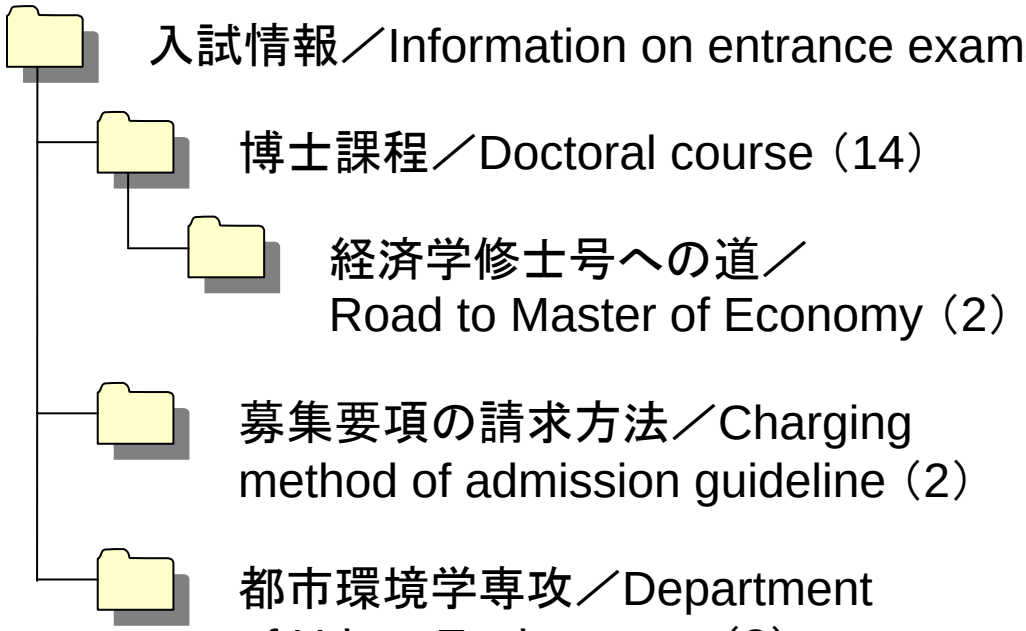


The accuracy of directory categorization
14 / 26

Site ID	Breakdown of pages
Ⅰ	5
Ⅱ	10
Ⅲ	0
Ⅳ	11
Ⅴ	0

Experimental Results

- The directory structure : sample 2



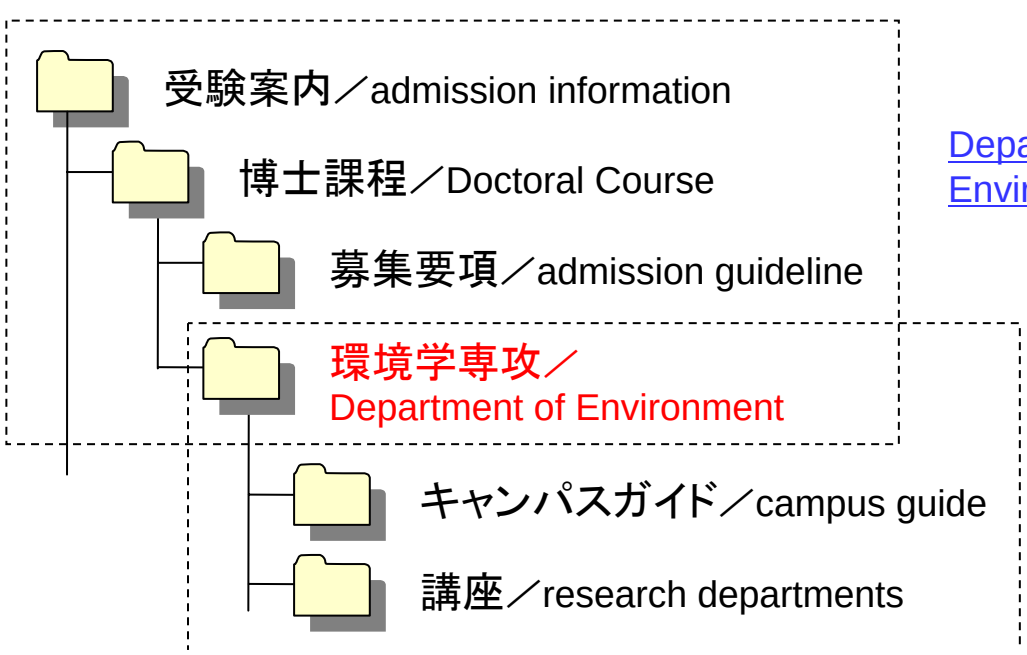
The accuracy of directory categorization
21 / 23

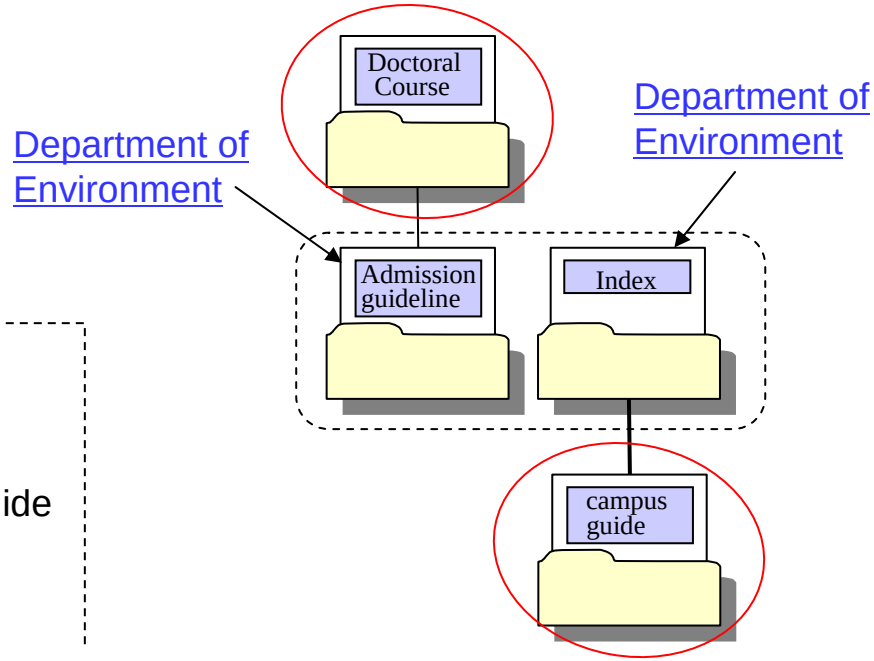
Site ID	Breakdown of pages
Ⅰ	3
Ⅱ	6
Ⅲ	0
Ⅳ	0
Ⅴ	14

Experimental Results

- Failure pattern (1)
  - Though the contents of the pages in two directories are different, their directories were integrated
    - because of the similarity of anchor texts between the integrated directories

⇒ Considering also the relation between another directories

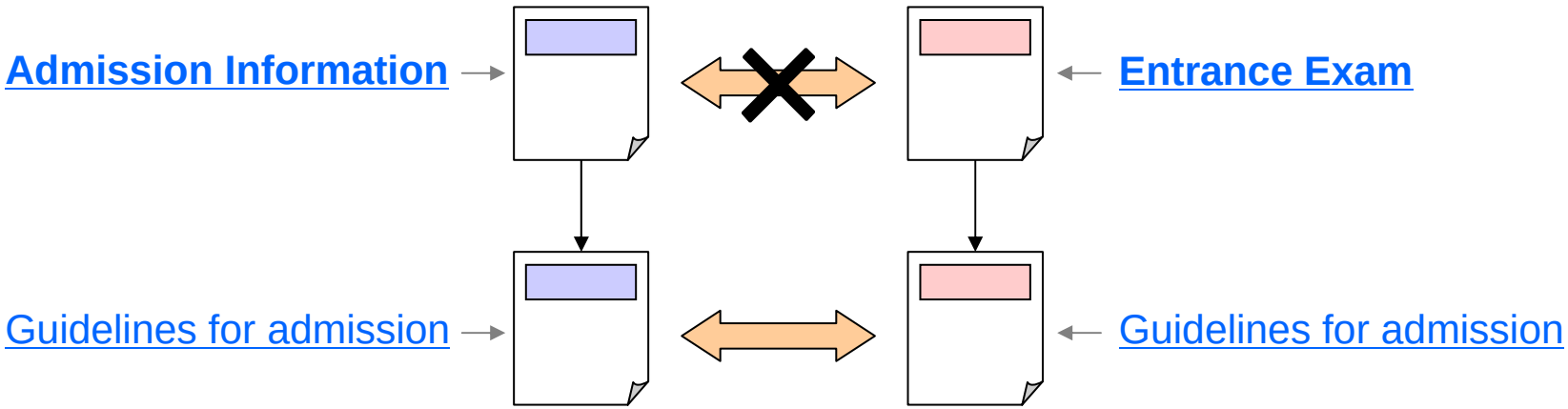




Experimental Results

- Failure pattern (2)
  - The pages of a common super-sub relation were not clustered into the same category
    - because of the mismatch between words in the anchor texts

⇨ Using also information other than anchor texts



Conclusion and Future Work

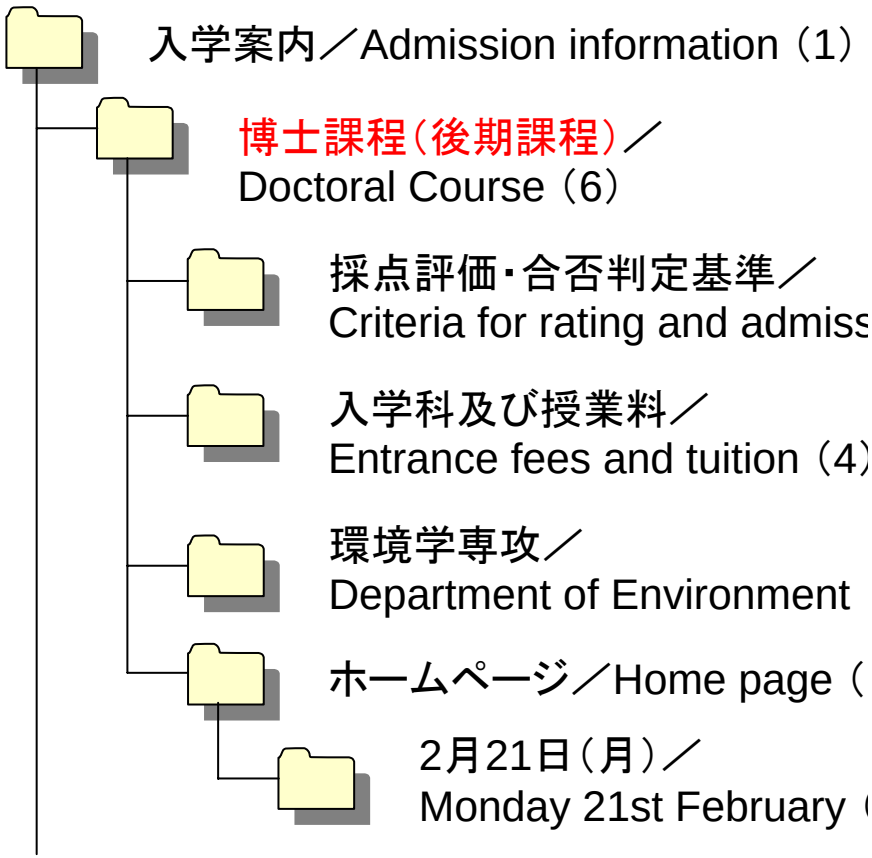
- Conclusion
  - We have proposed a method for constructing a hierarchical Web directory from several sites
  - We experimentally confirmed the feasibility of our method
- Future Work
  - To represent a super-sub relation by using information other than anchor texts
  - To examine the practicality of our method by increasing the amount of the data
    - Apply our method to sites of other categories

Thank you.

- 出力結果（Japanese）
  - http://plum.itc.nagoya-u.ac.jp/auto_directory/main.html

Experimental Results

- The directory structure : sample 1



A set of the anchor texts of “博士課程（後期課程）”

- 博士課程（後期課程）補欠募集
  / Doctoral course (to fill vacancies)
- 博士課程（後期課程）募集要項
  / Guidelines for admission to Doctoral course
- 博士課程（後期課程）
  / Doctoral course
- 博士課程（後期課程）
  / Doctoral course
- 博士課程（前期課程）募集要項
  / Guidelines for admission to Master’s course
- 博士課程（前期課程）
  / Master’s course

Related Works

- Generation of Web directory
  - [Sato et al. 1999]
    - Gather the links to the sites in a category

The generation of the hierarchical structure is not based on contents

- Grouping of Web pages
  - [Harada et al. 1999]
    - Group the folders in which Web pages are contained
  - [Kozima et al. 2002]
    - Extract the strongly connected components as the group by regarding the Web as a directed graph
    - Group the pages in the site hierarchically

The grouping of the pages across the site is not targeted