Hierarchical Organization of Web Documents based on Hypertext Classification

Yusuke Suzuki[1], Shigeki Matsubara[2], and Masatoshi Yoshikawa[2]

[1] Graduate School of Information Science, Nagoya University
[2] Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
Email: suzuki@dl.itc.nagoya-u.ac.jp

ABSTRACT

With the significant increase in accessible resources on World Wide Web, it is important to organize the Web contents effectively in advance. Web directories provide collections of links to lead the users to the sites in the desired category. However, most directories lead the users only to the top pages in the sites. If the users desire to browse the particular pages in the sites, the pages in the sites should be also organized effectively in advance.

This paper proposes a method of automatically constructing the hierarchical Web directories consisting of the pages in several sites. In order to construct a hierarchical directory automatically, it is necessary to obtain the hierarchical super-sub relations between the directories. The method finds the Web pages with the super-sub relations as the hyperlink, and replaces the relation by the hierarchical super-sub relation between directories. The method constructs the hierarchical directories by iterating the integration of directories. As a result of the experiment using five web sites, the hierarchical directories consisting of the Web pages in several sites were constructed. Hence, we confirmed the feasibility of the proposed method.

1 INTRODUCTION

Recently, the accessible resources on World Wide Web are increasing significantly. In order to access the target pages efficiently, it is desired that the Web pages are organized in advance. The Web directories such as Yahoo! and links pages provide collections of links in the same category and help the users to efficiently access the target pages. However, most Web directories and links pages lead the users only to the top pages on the sites. When the users want to browse the particular pages on several sites, for example, the deadlines of the papers in each academic society site or the service contents provided by each Internet service provider site, they have to look for the desired pages by following the links from the top page of each site.

Therefore, it is hopeful that the pages in the related sites are also organized according to the contents. It is one method to put the pages in several sites into the hierarchical directories based on the contents. For example, for several sites of the academic society, the directory structure such as figure 1 is constructed. This advantage is that the users can easily browse the pages with the same contents in several sites and grasp the whole contents in the related sites. However, because the desired directory structures vary by the categories, a large amount of labor is required to design the hierarchical structure and to categorize the pages into the directories manually.

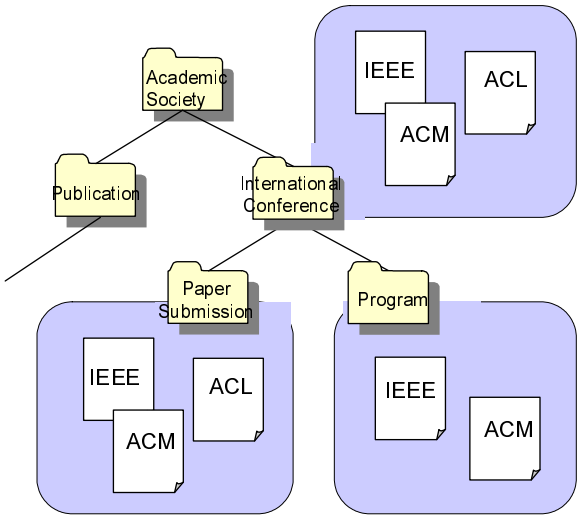


Fig. 1. Hierarchical structure of Web directory

This paper proposes a method of automatically constructing the hierarchical directories consisting of the pages in several sites. The method makes the super-sub directory structure by extracting the semantic super-sub relations between the Web pages and clustering them according to their contents. Then, by integrating the directories with the similar contents, the method constructs hierarchical directory structures.

There are several studies on automatic organization of the Web pages. Harada et al. have propoesed a method for grouping the pages in the site by regarding the pages in the same directory as a group and deciding the core page in the group by the file names and the reference relation by the hyperlinks [3]. Kozima et al. have provided a technique for grouping the pages in the site hierarchically by regarding the Web as a directed graph and decomposing the pages in the site into strongly connected components [4]. However, there studies do not target the grouping of the pages across the site and therefore differ from our study of organizing the pages in several sites.

We have evaluated the feasibility of our proposed method. As a result of an experiment using five web sites in Nagoya University, the hierarchical directory structures containing the pages on several sites were constructed. Therefore, we have confirmed our proposed method to be feasible.

This paper is organized as follows: Section 2 describes the concept of constructing the hierarchical directory structures from Web. Section 3 explains the method of constructing the directory structures. Section 4 evaluates our proposed method by the experiment.

2 OUTLINE OF OUR METHOD

In order to construct a hierarchical directory structure automatically, it is necessary to make the hierarchical super-sub relations between the directories and to categorize the Web pages into the directories.

Assume that, in the figure 1, the Web pages on several academic society sites are categorized into two directories with a super-sub relation: “International Conference” and “Paper Submission”. In this case, it is thought that a semantic super-sub relation exists in advance between the page which was categorized into the super-directory “International Conference” and the page which was categorized into the sub-directory “Paper Submission”. This shows that, in order to construct a hierarchical structure, it is only

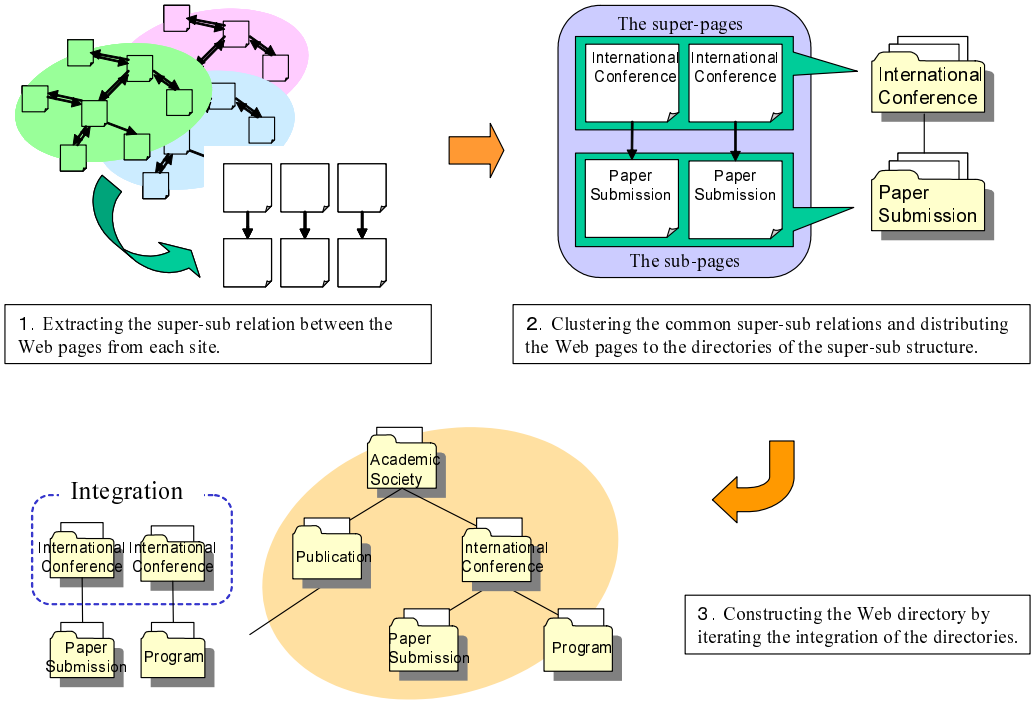


Fig. 2. Overview of the proposed method

necessary to extract the pages with a semantic super-sub relation from web sites. It is a high possibility that the links which connect between the pages with such a semantic relation exist. For example, a page which belongs to the directory "International Conference" links to a page which "Paper Submission" by the hyperlink. If the Web pages with a super-sub relation can be specified by the hyperlink, a super-sub relation between two directories can be produced by replacing a super-sub relation between the pages by a super-sub relation between the directories. In addition, the Web pages can be categorized into the directories at the same time (Figure 2).

2.1 Web Pages Connected by Link and the Super-sub Relation

It is important to identify the Web pages with a semantic super-sub relation because all Web pages connected by links don't have a super-sub relation.

When making a web site, the web site creators put the Web pages in folders and locate them on the server. The operations are done based on the judgment of the creators. For example, the creators trend to put the related Web pages into a folder and to make a folder in which contain the pages with more detailed contents in the folder. Therefore, we think that the Web pages with a super-sub relation are able to be identified by using the location of the Web pages on a server.

In order to utilize such the knowledge of the creators, we investigated the relevance between a super-sub relation between the Web pages connected by links and a location of the Web pages on a server. We extracted 200 links which refer to the pages in the site from each of four sites in Nagoya University, and judged whether the linking page and the linked page have a super-sub relation. Then, we classified the links into six location relations according to a relative location relation of the linked page to the linking page, and researched a rate of a super-sub relation in each location relation. Figure 3 shows the location relations on a server of the linked page to the linking page. "Super folder", "middle folder" and "sub

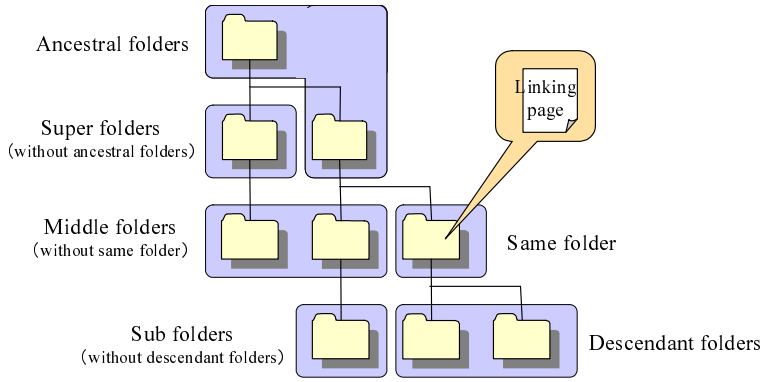


Fig. 3. Location relation of two pages on a server

Table 1. Rate of the super-sub relation

Location of linked pages	Link	Super-sub relation(%)
descendant folder	136	91.9
ancestral folder	151	0.7
same folder	246	58.1
sub folder (without descendant folder)	3	0
super folder (without ancestral folder)	77	2.6
middle folder (without same folder)	152	2.7
total	765	36.0

folder" represent that the depth of the path on server of the linked page is respectively shallower, same and deeper than that of the linking page.

Table 1 displays the results. The total number of links is reduced in order to remove the dead links. "Super-sub relation" represents a rate of the links connected the Web pages with a super-sub relation in each of location relations. 97.5% of the links in a super-sub relation belong to "descendant folder" and "same folder". Moreover, we investigated the relevance in order to identify the links in a super-sub relation in "same folder". When we investigated the case that the linking page in "same folder" is "index.html", the total number of such links is 41 and the rate of the links in a super-sub relation is 85.3%.

2.2 Representation of Super-page and Sub-page

We focus on an anchor text as a method of representing a super-page and a sub-page which are in a super-sub relation. Because an anchor text is set by the creators in order to lead the users into a linked page, it is often a description representing briefly the whole contents of the linked page. We think that characterizing a Web page by the anchor texts enables to represent the super-sub relation more clearly than characterizing by the contents of the page itself because the type of the contents or the text size varies by the pages. Therefore, we represent each page of the super-sub relations using the anchor texts which link to it. If the anchor texts are the reference terms such as "Back", they are excluded as a stopword in order not to represent the contents of the linked page.

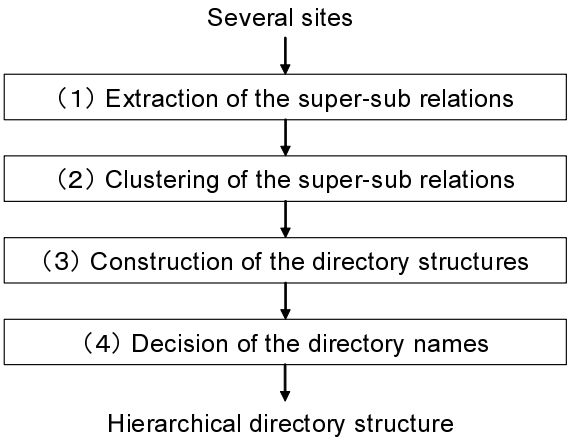


Fig. 4. Flow of constructing the Web directory

3 CONSTRUCTION OF WEB DIRECTORIES

Figure 4 shows the flow of construction processing of Web directories from several sites. The procedure is as follows:

1. Extracting the super-sub relation between the Web pages connected by the links.
2. Clustering the common super-sub relations.
3. Replacing a super-sub relation between the Web pages by a super-sub structure between the directories, and constructing the hierarchical directory structures by integrating the directories.
4. Giving each directory the name.

3.1 Extraction of the Super-sub Relations

For the links connecting between the pages in a site, the Web pages with a super-sub relation are extracted as the pair of the Web pages. Whether the Web pages connected by the link are a super-sub relation is judged by the rules based on the results of the research of Section 2.1. That is, if the Web pages connected by the link fulfill all the following conditions, they are extracted as the page-pair consisting of the linking page and the linked page.

1. Both the linking page and the linked page exist on the same server.
2. The linked page exists in the same folder as the folder in which contains the linking page, or in the folder which is located in a descendant position to the folder in which contains the linking page.
3. In the first case in 2., if the page of "index.html" exists in the folder, the linking page is "index.html". If not so, the linking page is a page which most links to the pages in the same folder.

Hereinafter, when the superior page is defined as d_{sup} and the inferior page is defined as d_{inf}, the page-pair p with a super-sub relation is represented as (d_{sup}, d_{inf}). Also, d_{sup} refers to a super-page and d_{inf} refers to a sub-page.

3.2 Clustering of Super-Sub Relations

For the super-sub relations which were extracted from several sites, the common super-sub relations are clustered. Here, a common super-sub relation is defined that both the contents between the super-pages

and the contents between the sub-pages of two relations are similar. The similarity between the Web pages is calculated by Dice coefficient [1] between the anchor texts which link to each page. That is, for two Web pages, the similarity between each anchor text which links to the one page and each anchor text which links to the other page is all calculated, and its maximal value is adopted as the similarity between the pages. When the anchor text which links to the page d_i is defined as $a_{i_s}(1 \leq s \leq m)$ and the anchor text which links to the page d_j is defined as $a_{i_s}(1 \leq s \leq m)$, the similarity between the pages d_i and d_j is defined as

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

where M_{i_s} is the number of nouns of a_{i_s} and $M_{i_s j_t}$ is the number of common nouns to a_{i_s} and a_{j_t}.

The similarity between the super-sub relations is represented by the similarity between the super-pages and between the sub-pages. The similarity $sim_{sup}(p_i, p_j)$ between the super-pages and the similarity $sim_{inf}(p_i, p_j)$ between the sub-pages to the page-pair p_i and $p_j (i \neq j)$ are calculated as formula (2) and (3), respectively.

$$sim_{sup}(p_i, p_j) = sim(d_{i_{sup}}, d_{j_{sup}}) \tag{2}$$

$$sim_{inf}(p_i, p_j) = sim(d_{i_{inf}}, d_{j_{inf}}) \tag{3}$$

The clustering is done based on the similarities between the super-sub relations. First, a cluster C_i consisting of a page-pair p_i is made as an initial cluster. The integrated clusters are the clusters which fulfill the following condition: 1) both the similarity between the super-pages and between the sub-pages are more than the threshold value α and 2) the average of their similarities is maximal. The calculation of the similarity between the clusters is applied to a complete linkage method [5]. The similarity $sim_{sup}(C_k, C_l)$ between the super-pages and the similarity $sim_{inf}(C_k, C_l)$ between the sub-pages to the cluster C_k and C_l are calculated as formula (4) and (5), respectively.

$$sim_{sup}(C_k, C_l) = \max_{p_i \in C_k, p_j \in C_l} (sim_{sup}(p_i, p_j)) \tag{4}$$

$$sim_{inf}(C_k, C_l) = \max_{p_i \in C_k, p_j \in C_l} (sim_{inf}(p_i, p_j)) \tag{5}$$

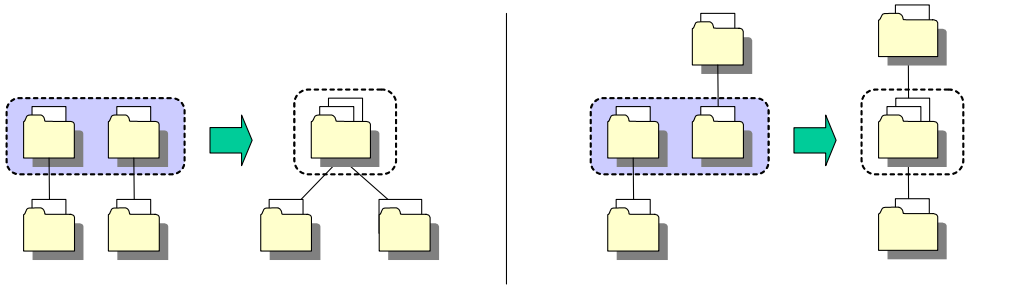
When the cluster which is more than the threshold α disappears, the clustering is stopped. If the number of the Web pages which belong to a cluster is less than m, its cluster is excluded.

3.3 Construction of Hierarchical Structure

The clustered super-sub relations are replaced by the super-sub directory structure. This is done as follows: First, the super-pages $d_{i_{sup}}$ and the sub-pages $d_{i_{inf}}$ of the page-pair p_i in cluster C are each distributed to the directory D_{sup} and D_{inf}. Then, the super-sub structure between the directories is represented as the directory-pair $P = (D_{sup}, D_{inf})$. Hereinafter, D_{sup} refers to a super-directory and D_{inf} refers to a sub-directory.

The hierarchical directory structure is constructed by integrating each directory in sequence. For example, when the super-directories of the directory-pairs are integrated, the directory structure which is a parent-child relation, such as figure 5(a), is produced, and when the super-directory and the sub-directory are integrated, the directory structure which has the third generation relation, such as figure 5(b), is produced.

In integration of the directories, the similarity between the directories is calculated by using a vector space model [2]. When a set of the anchor texts which link to the Web pages in a directory D_i is defined as A_i, the directory D_i is represented as a feature vector which is weighted by the frequency of nouns in



(a) Integration of super-directories (b) Integration of a super-directory and a sub-directory

Fig. 5. Integration of the directories

A_i. When a set of nouns is defined as $\{e_1 \dots e_N\}$ and a weight w_{ij} of a noun e_j is defined as formula (6), a feature vector of a directory D_i is represented as

$$\vec{x_i} = (w_{i1}, w_{i2}, \cdots, w_{iN})$$
$$w_{ij} = F_{ij} \tag{6}$$

where F_{ij} is the frequency of a noun e_j in A_i. By the formula (6), the feature vector $\vec{x}_{i_{sup}}$ of the super-directory $D_{i_{sup}}$ and the feature vector $\vec{x}_{i_{inf}}$ of the sub-directory $D_{i_{inf}}$ in the directory-pair $P_i = (D_{i_{sup}}, D_{i_{inf}})$ are calculated respectively.

The similarity between the directories is calculated by a cosign of the feature vectors. The similarity between a directory D_i and $D_j (i \neq j)$ is defined as formula (7).

$$Sim(D_i, D_j) = \frac{\vec{x}_i \bullet \vec{x}_j}{|\vec{x}_i||\vec{x}_j|} \tag{7}$$

By the formula (7), the similarity $Sim(D_{i_{sup}}, D_{j_{sup}})$ between the super-directories and the similarity $Sim(D_{i_{sup}}, D_{j_{inf}})$ between the super-directory and the sub-directory in the directory-pair P_i and P_j are calculated respectively.

An integration of the directories is done by integrating the directories to satisfy the nature of a tree structure in descending order of the similarity between the directories. Figure 6 shows its process. First, the method calculates all of the similarity $Sim(D_{i_{sup}}, D_{j_{sup|inf}})$ between the directories which are a part of the directory-pair P_i and P_j. Second, the method finds the directory-pairs P_k and P_l in which the similarity between the directories is maximal and more than a threshold value β, and evaluates an integrating validity of them. If they are valid, the directories $D_{k_{sup}}$ and $D_{l_{sup|inf}}$ in the directory-pairs P_k and P_l are integrated.

Here, an integrating validity assures that the constructed directory structure satisfies the nature of a tree structure. The directories are integrated so that the constructed directory structure fulfill the following conditions:

1. Each directory has at most one parent directory.
2. The directory structure is a noncyclic structure.

In case that the directories are integrated or an integrating validity is not satisfied, the method shifts to the directory-pairs which have the highest next similarity between the directories. Repeating this operation until the maximal similarity between the directories is less than a threshold value β, the method constructs the hierarchical directory structures.

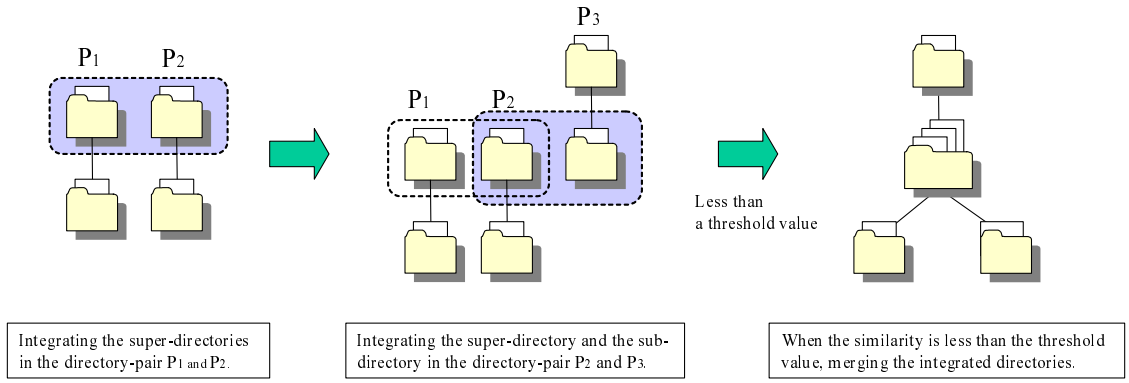


Fig. 6. Construction of the directory structure

When the maximal value of the similarity becomes less than the threshold value, the method merges the integrated directories and makes the new directory. When the integrated directories are defined as $D_1, \cdots, D_n$ and the new directory is defined as D_r and a set of the Web pages in a directory D_i is defined as W_i, a set of the Web pages in the new directory D_r is defined as formula (8).

$$W_r = \sum_{i=1}^{n} \cup W_i \tag{8}$$

3.4 Decision of Directories Names

The directory names are decided based on a set of the anchor texts linking to the Web pages in the directory. The policy of the decision of the directory name is that a directory name is a phrase which appears in common with a set of the anchor texts representing the directory and which has a certain degree of length.

First, the method extracts any morphological sequence s_{ij} from a set of the anchor texts, $A_i = \{a_{i_1}, \cdots, a_{i_M}\}$, which represents a directory D_i and makes them the candidates for its directory name. Second, for each morphological sequence s_{ij}, the inclusion rate $Cover(s_{ij}, a_{i_k})$ to the anchor text a_{i_k} in A_i is calculated as formula (9). Finally, the average inclusion rate $Cover_{ave}(s_{ij}, A_i)$ is calculated as formula (10), and the morphological sequence s_{ij} whose value is maximal makes the directory name.

$$Cover(s_{ij}, a_{ik}) = \begin{cases} \frac{F^i_{jk}}{|a_{i_k}|} & (|s_{ij}| \leq F^i_{jk}) \\ 0 & (otherwise) \end{cases} \tag{9}$$

$$Cover_{ave}(s_{ij}, A_i) = \frac{\sum_{k=1}^{M} Cover(s_{ij}, a_{i_k})}{M} \tag{10}$$

Here, $|a_{i_k}|$ is the number of morphemes in a_{i_k} and F^i_{jk} is the number of common morphemes in s_{ij} and a_{i_k}, $|s_{ij}|$ is the number of morphemes in s_{ij} and M is the number of the anchor texts in A_i.

4 EVALUATION EXPERIMENT

4.1 Outline of Experiment

We evaluated our method to confirm the feasibility of our method for constructing the hierarchical directory structures from several sites. In this experiment, we used five sites of the graduate course in

Table 2. Experimental data and its site

ID	Site	Page	Link
I	www.engg.nagoya-u.ac.jp	126	276
II	www.env.nagoya-u.ac.jp	281	1192
III	www.is.nagoya-u.ac.jp	106	267
IV	www.sci.nagoya-u.ac.jp	280	887
V	www.soec.nagoya-u.ac.jp	605	3288

Nagoya University. Table 2 shows the sites which we used. "Page" represents the number of the Web pages in each site and "Link" represents the number of the links to pages in the site. We gathered the anchor texts which represent the Web pages from each site. In setting of the threshold value, the parameter α, which is used for a clustering of the super-sub relations, was 0.5, the parameter β, which is used for a construction of the directory structure, was 0.6 and the clusters whose the number of members is less than 2 were excluded. We used Chasen [6] for a Japanese morphological analysis.

4.2 Experimental Result

Figure 7 shows the sample output of the constructed hierarchical directory structures. Figure 7 represents a part of thirteen directory structures which were produced by integrating the directories at least one times. Each number in figure 7 represents as follow: 1) the list of the root directory of the constructed directory structure, 2) an overall view of the specific directory structure, 3) the links to the Web pages belonging to the specific directory.

Table 3 and 4 show the example of the directory structures. "Level" represents the hierarchical level of the directory structure and "Page" represents the number of the Web pages in the directory. Also, "Page" is distributed to each site and each "ID" in "Page" corresponds the ID in Table 2. We can see from their table that the pages on several sites are categorized into a directory structure and that the super-sub structure which is valid to some extent is produced. From there results, we confirmed the feasibility of our method.

4.3 Discussion

We discuss the following items based on the experimental results.

4.3.1 The Validity of the Hierarchical Structure

We confirmed the validity to some extent between the super-directory and the sub-directory in the produced directory structure. However, we also observed some directory structures with the invalid super-sub structure. In our proposed method, when constructing the hierarchical structure of the directory, the integrated directories are decided based on the similarity of a set of the anchor texts. Therefore, even if the contents in the pages in two directories are different, their directories are integrated if the anchor texts are similar. For example, in case the directory is represented as "Department of Information", it is difficult to judge the contents of the pages in the directory from its phrase. In an integration of the directories, the current method considers only the relation between the integrating directories. If the method also considers the relation between the another directories which composes the super-sub relation, we think that such a structural contradiction will be reduced.

Fig. 7. Example of the system output

Table 3. Constructed directory structure (1)

Level	Directory name	Page		
		I	II	IV
1	入学案内/Entrance guide	1	0	0
1-1	博士課程 (後期課程)/Doctral course	2	2	2
1-1-2	採点評価・合否判定基準/Criteria for rating and admission decision	0	2	0
1-1-3	入学科及び授業料/Entrance fees and tuition	0	4	0
1-1-4	環境学専攻/Department of Environment	0	2	0
1-1-5	ホームページ/Home page	0	0	1
1-1-5-1	2 月 21 日 (月)/Monday 21st February	0	0	8
1-2	第 3 年次学士入学/Admission to 3rd bachelor	2	0	0

Table 4. Constructed directory structure (2)

Level	Directory name	Page		
		I	II	V
1	入試情報/Information on entrance exams	1	1	1
1-1	博士課程/Doctoral course	2	2	10
1-1-1	経済学修士号への道/Road to Master of Economy	0	0	2
1-2	募集要項の請求方法/Charging method of admission guideline	0	1	1
1-3	都市環境学専攻/Department of Urban Environment	0	2	0

Table 5. Directory names generated by the anchor texts

博士課程（後期課程）/Doctoral course
博士課程 (後期課程) 補欠募集，博士課程 (後期課程) 募集要項，博士課程 (後期課程) × 2， 博士課程 (前期課程) 募集要項，博士課程 (前期課程)
募集要項の請求方法/Charging method of admission guideline
募集要項の請求方法× 2，各種募集要項の請求方法× 3

4.3.2 The Accuracy of Directory Categorization

We judged the accuracy of the categorized page by whether the contents of the pages belonging to each directory match the name of its directory or not. When judging based on this measure, 35 of 49 pages belonging to the directory structures of table 3 and 4 were categorized correctly.

In respect of the reproducibility, we observed that each page of the page-pairs with a common super-sub relation is not clustered into same directory. Also, eight of generated 13 directory structures were composed by only the pages on a single site. There are attributed to the mismatch between words in the anchor texts even if the contents of the pages are similar. In order to cluster as many pages whose contents are similar as possible into same directory, we need to express each page by using also information other than anchor text.

4.3.3 The Validity of Directory Names

Each directory name is decided based on a set of anchor texts in the directory. For example, "博士課程 (後期課程)/Doctoral course" in table 3 and "募集要項の請求方法/Charging method of admission guideline" in table 4 are respectively generated from a set of anchor texts in table 5. Though many directory names were the phrase fitting the decision policy, there were also the directory names which are not always represented correctly. For example, in the sample of "博士課程 (後期課程)/Doctoral course", the pages representing by "前期課程/master's course" are not reflected in the directory name properly.

In addition, the grammatically incorrect directory names such as the directory name beginning with a postposition such as "の/of" were observed . Because this is attributed to the fact that our method considers only until a morphemic level, it is necessary to adopt a grammatical restriction toward a decision of directory names.

5 CONCLUDING REMARKS

In this paper, we have proposed a method for constructing the hierarchical directory structures from several sites and categorizing the Web pages into them based on hyperlinks and anchor texts. We described the evaluation experiment with several sites. In the experiment, because the directory structures into which the pages on several sites were categorized are constructed, we confirmed the feasibility of our proposed method. However, we observed some directory structures with the invalid super-sub structures.

In the future, in order to construct the valid super-sub structures, it is necessary to represent a super-sub relation by additionally using information other than anchor texts. In addition, we will examine the practicality of our method by increasing the amount of the experimental data.

REFERENCE

[1] G. Salton and M. J. McGill: Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[2] G. Salton, A. Singhal, C. Buckley and M. Mitra. "Automatic Text Decomposition Using Text Segments and Text Themes," In *Proceedings of the Hypertext 96*, pp.53-65, 1996.

[3] Masanori Harada, Shinya Sato and Kazuhiro Kazama. "Estimation of Hierarchical Structure of Web for Search Engine," IPSJ SIG Notes, DBS118-14, pp.105-112, 1999.

[4] Shuichi Kozima, Atsuhiro Takasu and Jun Adachi. "Structural analysis and grouping of Web pages," NII Journal, No.4, pp.23-35, 2002.

[5] P. Willett. "Recent trends in hierarchic document clustering: a critical review," *Information Processing and Management*, 24(5), pp.577-597, 1988.

[6] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita and Yoshitaka Hirano. *Japanese Morphological Analysis System Chasen version 2.0 Manual*. NAIST Techinical Report, NAIST-IS-TR99009, 1999.