Automatic Construction of Web Directory using Hyperlink and Anchor Text

Yusuke Suzuki [†], Shigeki Matsubara [‡] and Masatoshi Yoshikawa [‡]
[†] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya
[‡] Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya
Email: suzuki@dl.itc.nagoya-u.ac.jp

***Abstract*- This paper proposes a technique for automatically constructing Web directories from several sites. To construct the hierarchical structure of the directories, the technique finds Web pages with a super-sub relation, which are connected by hyperlinks, and replaces the relation with a super-sub hierarchical relation between directories. The technique constructs hierarchical directories by iterating the integration of directories. As a result of an experiment using five Web sites, it was possible to construct hierarchical directories containing Web pages from several sites.**

I. INTRODUCTION

To find target information on the WWW efficiently, it is ideal that the Web pages should be organized in advance. Link collections help users to efficiently access target sites. Web directories such as Yahoo! [1] and Google [2] are also considered as hierarchical link collections. However, many link collections collect only the top pages of Web sites. In such traditional link collections, when users want to browse particular pages on several sites, for example, deadlines for paper submissions to academic conferences or service contents provided by Internet service providers, they have to bother to seek their target pages by following several links from the top pages on each site.

Therefore, in order to access to target pages more effectively, it is hopeful that not only top pages on sites but also individual pages on sites are also linked. One method is to put the pages existing in several sites into contents-based hierarchical directories. For example, for several sites belonging to academic societies, a directory structure such as Fig. 1 is constructed. The advantage here is that users can easily browse pages featuring similar contents in several sites and grasp the entire contents in related sites. However, because the desired directory structures vary according to target category, a large amount of labor is required to design a hierarchical structure and to categorize pages into directories manually.

This paper proposes a technique for automatically constructing a Web directory consisting of pages in several related sites. The technique produces the directory structure by extracting the semantic relations between Web pages and clustering them according to their contents. Then, it constructs hierarchical directory structures by integrating directories with similar contents.

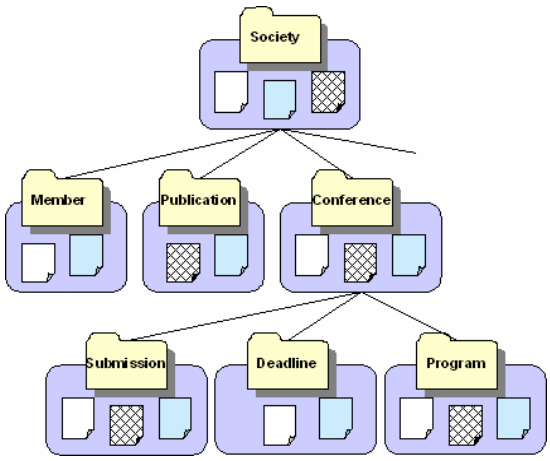

Fig.1. Hierarchical structure of Web directory

There are several studies on automatic generation of link collections [3], [4]. These studies generate automatically link collections by giving a category word to each category. However, such link collections link to only top pages on sites. Our study constructs a Web directory from related sites selected by such method or manual method. Kojima et al. have formulated a technique for grouping pages in a site hierarchically by regarding the Web as a directed graph and decomposing the pages in the site into strongly connected components [5]. However, this study does not target the grouping of pages across sites and therefore differ from our study, in which we organize pages from several sites.

We have evaluated the feasibility of our technique. As a result of an experiment using five Web sites, we constructed hierarchical directory structures containing pages in these sites. Therefore, we have confirmed the technique to be feasible for automatic construction of Web directories.

This paper is organized as follows: Section 2 describes the concept of constructing the hierarchical directory structures

[1] http://dir.yahoo.com/

[2] http://www.google.com/dirhp/

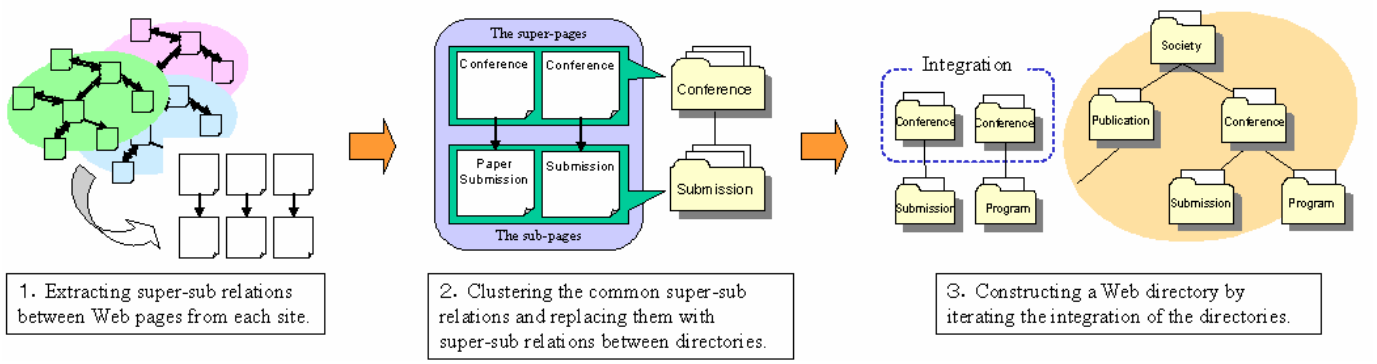

Fig.2. Overview of the proposed method

from the Web. Section 3 explains the method of constructing the directory structures. In Section 4, we evaluate our proposed method experimentally, and after concluding remarks in Section 5.

II. OUTLINE OF OUR METHOD

To construct a hierarchical Web directory automatically, it is necessary to create super-sub relations between directories and to categorize Web pages into those directories.

Assume that, in Fig. 1, the Web pages on several sites belonging to academic societies are categorized into two directories with the super-sub relations: "conference" and "submission." In this case, it is thought that a semantic super-sub relation exists in advance between pages in the super-directory "conference" and pages in the "submission" sub-directory. This shows that in order to construct a hierarchical structure, it is only necessary to extract pages with a semantic super-sub relation from Web sites. Links might exist between pages in such a semantic relation. For example, a page belonging to the directory "conference" links to a "submission" page via a hyperlink. If Web pages with a super-sub relation can be specified by a hyperlink, the super-sub relation between two directories can also be produced by replacing a super-sub relation between the pages with a super-sub relation between the directories. In addition, the Web pages can be categorized into the directories at the same time. Figure 2 shows an overview of the proposed method.

A. Super-sub Relation between Web Pages

It is important to identify Web pages with a semantic super-sub relation because Web pages connected by links do not necessarily have a super-sub relation.

When building a Web site, creators put Web pages into folders and locate them on the server. The operations are done based on the desires of the sites' creators. For example, creators tend to put related Web pages into the same folder and put pages containing more detailed contents into the lower folder. Therefore, we think that Web pages with a super-sub relation can be identified by using their individual location on a server.

To utilize such knowledge of Website creators, we investigated the relevance between a super-sub relation of Web pages connected by links and the location of Web pages on a server. We extracted 200 links from each of four sites at Nagoya University[3], and judged whether linking pages and linked pages have a super-sub relation. Then, we classified the links into four groups according to the relative location relations of linked pages to linking pages, and investigated the rate of super-sub relations in each location relation. Figure 3 shows the location relations on a server of the linked pages to the linking pages.

Table I shows the results. The total number of links is reduced in order to remove the dead links. "Rate" represents the rate of links connected the Web pages with a super-sub relation in each of the location relations; 97.5% of the links belong to "descendant folder" or "same folder." In addition, we investigated the relevance to identify the links in a super-sub relation in "same folder." The total number of such links for which the linking page in "same folder" is "index.html" is 41 and the percentage of these links in a super-sub relation is 85.3%.

B. Representing Super-pages and Sub-pages

We use an anchor text to represent a super-page and a sub-page in a super-sub relation. Because an anchor text is set by the creators to lead the users into a linked page, it is often a brief description representing the whole contents of the linked page. We think that characterizing a Web page by the anchor texts enables clearer representation of the super-sub relation than characterizing by the contents of the page itself because the type of contents or the text size may vary from page to page. Therefore, we represent each page of the super-sub

[3] http://www.nagoya-u.ac.jp/en/

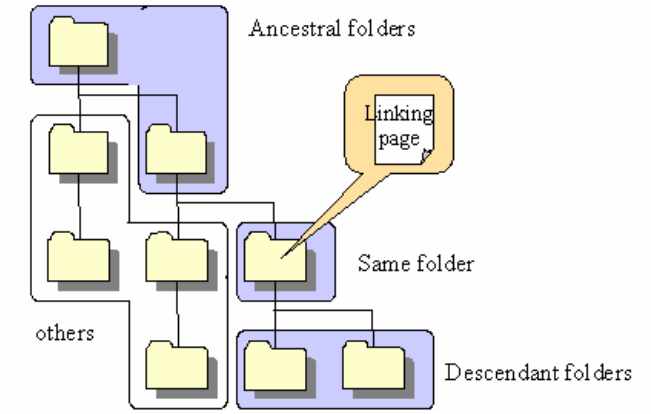

Fig.3. Location relation of two pages on a server

TABLE I

RATE OF SUPER-SUB RELATIONS

Location of linked pages	Link	Rate (%)
Descendant folder	136	91.9
Ancestral folder	151	0.7
Same folder	246	58.1
0thers	232	2.6
Total	765	36.0

relations using anchor texts. If the anchor texts are reference terms such as "Back," they are excluded as a stopword so as not to represent the contents of the linked page.

III. CONSTRUCTING WEB DIRECTORIES

Below we explain the process of constructing Web directories from several sites.

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

A. Extracting the Super-sub Relations

For links connecting pages in a site, Web pages with a super-sub relation are extracted as pairs of Web pages. Whether the Web pages are in a super-sub relation or not is determined based on the investigation in Section II-*A*. That is, if Web pages connected by a link fulfill all of the following conditions, they are extracted as page-pairs.

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 containing the linking page, or in the folder located in a descendant position to the folder containing the linking page.
3. In the first case in 2., if the page "index.html" exists in the folder, the linking page is "index.html." If not, the linking page is a page that links to the most pages in the same folder.

Here after, when the superior page is defined as $d_{\sup}$ and the inferior page is defined as $d_{\inf}$, any 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.

B. Clustering Super-Sub Relations

For the super-sub relations we extracted from several sites, the common super-sub relations are clustered. Here, a common super-sub relation means that both the contents of the super-pages and the contents of the sub-pages are similar. The similarity between the Web pages is calculated by Dice coefficient [1] between the anchor texts used for the link to each page. That is, when the anchor text in the link to page d_i is $a_{i_s}(1 \le s \le m)$ and the anchor text in the link to page d_j is $a_{j_t}(1 \le t \le n)$, the similarity between pages d_i and d_j is defined as

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

where M_{i_s} is the number of nouns of a_{i_s} and $M_{i_s j_t}$ is the number of nouns cooccurring 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 $p_j (i \ne j)$ are calculated as Eqs. (2) and (3), respectively.

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

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

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 required to fulfill the following conditions: 1) that both the similarity between the super-pages and between the sub-pages exceed the threshold value α, and 2) that the average of their similarities is maximal. The calculation of the similarity between the clusters is applied to a complete linkage method [6]. The similarity $sim_{\sup}(C_k, C_l)$

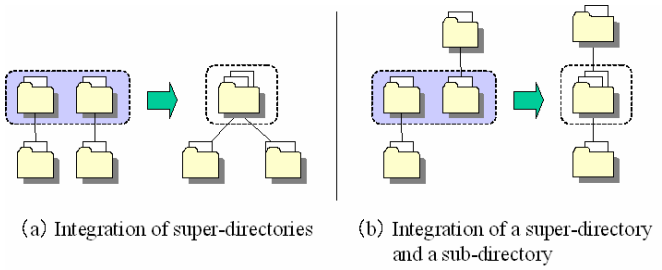

Fig.4. Integration of the directories

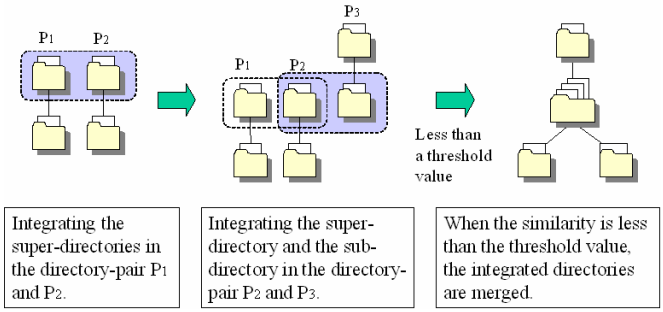

Fig.5. Construction of the directory structure

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 Eqs. (4) and (5), respectively.

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

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

When the cluster whose similarity exceeds threshold α disappears, the clustering is stopped. If the number of Web pages belonging to a cluster is less than m, its cluster is excluded.

C. *Constructing a Hierarchical Structure*

The clustered super-sub relations are replaced by a super-sub directory structure. This is achieved in the following way: 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 $P = (D_{\sup}, D_{\inf})$. Here after, $D_{\sup}$ 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 in a parent-child relation is produced, as Fig. 4(a) shows. On the other hand, when the super-directory and the sub-directory are integrated, the directory structure which has the third-generation relation is produced as show in Fig. 4(b).

In integrating the directories, the similarity between the directories is calculated by using a vector space model [2]. When a set of anchor texts linking to the Web pages in a directory D_{i} is defined as A_{i}, the directory D_{i} is represented as a feature vector weighted by the frequency of nouns in A_{i}. Let a set of nouns be $\{e_1 \cdots e_N\}$ and a weight w_{ij} of a noun e_j be given by Eq. (6).

$$w_{ij} = F_{ij} \quad (6)$$

Here, F_{ij} is the frequency of a noun e_j in A_{i}. Then, a feature vector of a directory D_{i} is represented as $\overrightarrow{x_i} = (w_{i1}, w_{i2}, \cdots, w_{iN})$. By the Eq. (6), the feature vector $\overrightarrow{x_{i_{\sup}}}$ of the super-directory $D_{i_{\sup}}$ and the feature vector $\overrightarrow{x_{i_{\inf}}}$ of the sub-directory $D_{i_{\inf}}$ in the directory-pair $P_i = (D_{i_{\sup}}, D_{i_{\inf}})$ are calculated in turn.

The similarity between directories D_{i} and $D_j (i \neq j)$ is defined by Eq. (7).

$$Sim(D_i, D_j) = \frac{\overrightarrow{x_i} \bullet \overrightarrow{x_j}}{\left|\overrightarrow{x_i}\right| \left|\overrightarrow{x_j}\right|} \quad (7)$$

Equation (7) is used to calculate 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.

The directories are integrated by integrating the directories satisfying the nature of a tree structure in descending order of similarity between the directories. Figure 5 illustrates this process. First, the method calculates all the similarity $Sim(D_{i_{\sup}}, D_{j_{\sup\inf}})$ between the directories that are a part of the directory-pair P_i and P_j. Second, it 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 whether an integration of them will be valid. 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, validity of the integration assures that the constructed directory structure satisfies the nature of a tree structure. The directories are integrated so that the constructed directory structure fulfills the following conditions:

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

If the directories are integrated or the validity of integration is not satisfied, the method shifts to the directory-pairs that have the highest next similarity between directories. Repeating this operation until the maximal similarity between

the directories is less than a threshold value β, the method constructs hierarchical directory structures.

When the maximal value of the similarity becomes less than the threshold value, the method merges the integrated directories and creates a 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 Web pages in a directory D_i is defined as W_i, then a set of Web pages in the new directory D_r is defined as Eq. (8).

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

D. Deciding Directory 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 deciding a directory name is that a directory name be a phrase appearing in common with a set of anchor texts representing the directory and having a certain length.

First, the method extracts any morphological sequence s_{ij} from a set of 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 using Eq. (9). Finally, the average inclusion rate $Cover_{ave}(s_{ij}, A_i)$ is calculated with Eq. (10), and the morphological sequence s_{ij} whose value is maximal constitutes directory name.

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

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

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

[4] http://www.engg.nagoya-u.ac.jp/

[5] http://www.env.nagoya-u.ac.jp/

[6] http://www.is.nagoya-u.ac.jp/

[7] http://www.sci.nagoya-u.ac.jp/

[8] http://www.soec.nagoya-u.ac.jp/

TABLE II

SITES EXAMINED AND THEIR DATA

ID	Site	Pages	Links
I	Graduate School of Engineering[4]	126	276
II	Graduate School of Environmental Studies[5]	281	1192
III	Graduate School of Information Science[6]	106	267
IV	Graduate School of Science[7]	280	887
V	Graduate School of Economics[8]	605	3288

IV. EVALUATION

A. Outline of the Experiment

We evaluated the feasibility of our method for constructing hierarchical directory structures from several sites. In this experiment, we used five sites operated by graduate schools at Nagoya University, which were made independently. Table II displays an outline of the sites. "Pages" represents the number of Web pages in each site and "Links" represents the number of links to pages in each respective site. We gathered anchor texts that represent the Web pages from each site. In setting the threshold value, parameter α (used for clustering the super-sub relations) was 0.5, and parameter β (used for constructing the directory structure) was 0.6. Clusters whose the number of members is less than 2 were excluded. We used Chasen [7] to conduct a morphological analysis in Japanese.

B. Experimental Results

Figure 6 shows a sample output of the constructed hierarchical directory structures. Figure 6 represents a part of thirteen directory structures that were produced by integrating the directories at least once. Each number in Fig. 6 represents the following: 1) a list of the root directories in the constructed directory structure, 2) an overall view of the specific directory structure, and 3) links to the Web pages belonging to the specific directory.

Tables III and IV show examples of the directory structures. "Level" represents the hierarchical level of the directory structure and "Page" represents the number of Web pages in the directory. Also, "Page" is distributed to each site and each "ID" in "Page" corresponds to the ID in Table II. We can see from these tables that the pages on several sites are categorized into a directory structure and that a super-sub structure, which is valid to some extent, is produced. From these results, we were able to confirm the feasibility of our method.

Fig.6. Example of system output

TABLE III

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) / Doctoral course	2	2	2
1-1-1	採点評価･合否判定基準 / Criteria for rating and admission decisions	0	2	0
1-1-2	入学料及び授業料 / Entrance fees and tuition	0	4	0
1-1-3	環境学専攻 / Department of Environment	0	2	0
1-1-4	ホームページ / Homepage	0	0	1
1-1-4-1	2月21日(月) / Monday 21st February	0	0	8
1-2	第3年次学士入学 / Admission to 3rd year of bachelor course	2	0	0

TABLE IV

CONSTRUCTED DIRECTORY STRUCTURE (2)

Level	Directory name	Page		
		I	II	V
1	入学情報 / 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 Masters in Economics	0	0	2
1-2	募集要項の請求方法 / Guidelines for admission charges	0	1	1
1-3	都市環境学専攻 / Department of Urban Environment	0	2	0

V. CONCLUDING REMARKS

In this paper, we have proposed a method for constructing hierarchical directory structures from several sites and categorizing Web pages into them based on hyperlinks and anchor texts. We described an evaluation experiment using five sites. In the experiment, we were able to construct directory structures into which pages on several sites were categorized, and as a result we confirmed the feasibility of our method.

In future, in order to construct more valid super-sub structures, it will be necessary to represent super-sub relations by using additional information other than anchor texts. Furthermore, we will examine the practicality of our method by increasing the amount of experimental data.

ACKNOWLEDGEMENT

The authors would like to thank all members of the information distribution working group in Nagoya University for their helpful comments.

REFERENCES

[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 Hypertext 96*, pp. 53-65, 1996.

[3] S. Sato and M. Sato, “Toward Automatic Generation of Web Directories,” In *Proceedings of International Symposium on Digital Libraries 1999*, pp. 127-134, 1999.

[4] O. Segawa, J. Kawai and K. Sakauchi, “Automatic Generation of Link Collections and their Visualization,” In *Proceedings of The 14th International World Wide Web Conference*, pp. 942-943, 2005.

[5] S. Kozima, A. Takasu and J. Adachi, “Structural analysis and grouping of Web pages,” NII Journal, No. 4, pp. 23-35, 2002.

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

[7] Y. Matsumoto, A. Kitauchi, T. Yamashita and Y. Hirano, *Japanese Morphological Analysis System Chasen version 2.0 Manual*. NAIST Techinical Report, NAIST-IS-TR99009, 1999.