社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# アンカーテキストを用いたWebディレクトリの構築

鈴木　祐介†　　松原　茂樹††　　吉川　正俊††

† 名古屋大学大学院情報科学研究科
†† 名古屋大学情報連携基盤センター
E-mail: †suzuki@dl.itc.nagoya-u.ac.jp

**あらまし**　Web 上から必要な情報を効率よく得るために，Web ページがあらかじめ効果的に整理されていることが望まれる．本論文では，複数サイトに散在する Web ページをディレクトリ構造として階層的に分類する手法を提案する．意味的な上位-下位関係にある Web ページをハイパーリンクを手がかりに特定し，その関係を用いてディレクトリの上位-下位関係を作り上げる．階層構造は，ディレクトリの間で統合を実行することにより構築する．名古屋大学の Web サイトを用いてディレクトリ構造の構築実験を行い，提案手法の実現可能性を確認した．
**キーワード**　WWW，階層ディレクトリ，クラスタリング，階層構造，ハイパーリンク

# Construction of Web Directory using Anchor Text

Yusuke SUZUKI†, Shigeki MATSUBARA††, and Masatoshi YOSHIKAWA††

† Graduate School of Information Science, Nagoya University
†† Information Technology Center, Nagoya University
E-mail: †suzuki@dl.itc.nagoya-u.ac.jp

**Abstract**　This paper proposes a method for automatically constructing the hierarchical Web directories from several sites. In order to construct the hierarchical structure of the directories, the method finds the Web pages with the super-sub relations which are connected by the hyperlink, and replaces the relation by the super-sub hierarchical 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 in which the Web pages on several sites are contained were constructed.
**Key words**　WWW，hierarchical directory，clustering，hierarchical structure，hyperlink

## 1. はじめに

Web 上から必要な情報を効率よく得るために，Web 上の情報があらかじめ効果的に整理されていることが重要となる．一般に，Web 上の情報はサイト単位では適切に整理されている場合が多いが，ユーザから見れば，サイト内だけでなくサイト間でも同様に整理されていることが望ましい．複数サイト間の関係を記すものとしてリンク集があり，学会やプロバイダなど，あるテーマに関連したサイトのリンクが集められている．リンク集では多くの場合，関連するサイトのトップページにリンクが張られており，ユーザは検索対象に合致したジャンルのリンク集を利用すればよい．しかし，関連する学会の論文の締め切り日を知りたい，あるいは，プロバイダが提供するサービス内容を比較したい，といったように，複数のサイトにまたがってある特定のページを参照したい場合は，ユーザは各サイトごとにトップページからリンクを辿って目的とするページを探す必要がある．

したがって，サイト間の情報全体が，その内容に従ってより効果的に整理されていることが望まれる．複数サイト内のページをその共通性に基づいてディレクトリとしてまとめ，階層化することは 1 つの方法である．例えば，学会のサイト群に対しては，ディレクトリ構造として図 1 のような構成が考えられ，各学会の論文投稿や大会プログラムに関するページをそれぞれのディレクトリにまとめる．これにより，サイト間で関連したページの閲覧が容易になる，ジャンルのサイト内情報の全体像を把握しやすいといった利点がある．しかし，このディレクトリの階層構造は対象とするジャンルによって異なるため，階層構造の設計やディレクトリへのページの分類には多大な労力を要する．

本論文では，複数サイトのページを分類し，階層的なディレクトリを構築する手法を提案する．Web ページ間の意味的な上位-下位関係を抽出し，内容に従ってクラスタリングすることに

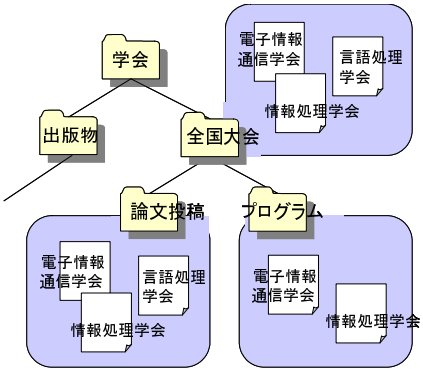


図 1　Web ディレクトリの階層構造

Fig. 1　Hierarchical structure of Web directory

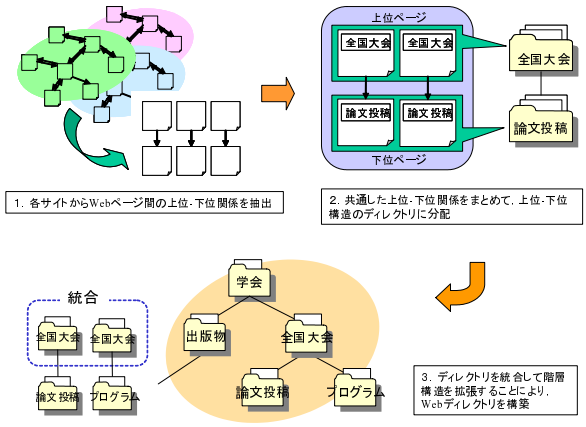


図 2　提案手法の概要

Fig. 2　Overview of the proposed method

より，ディレクトリの上位-下位構造を作り上げる．さらに，内容が類似したディレクトリを統合することにより，ディレクリの構造を階層化する．

Web ページを自動的に整理する研究はいくつか行われている．原田らは，同一ディレクトリ内のページを，ある主題をもったページ群としてまとめ，グループの中心となるページをファイル名やリンクの参照関係からもとめることにより，サイト内のページを整理する手法を提案している [3]．また，小島らは，サイト上の意味的に関連したページをページ間のリンク構造パターンをもとにグループ化して，グループの階層構造を作成することにより，サイト内のページを整理する手法を提案している [4]．しかし，これらの研究は，特定のサイト内ページのグループ化を対象としており，複数サイトのページを整理する我々の研究とは異なる．

本手法の実現可能性を検証するために，評価実験を行った．名古屋大学の 5 つの Web サイトを用いて実験したところ，複数サイトのページが分類された階層ディレクトリ構造が生成され，本手法の実現可能性を確認した．

本論文の構成は以下の通りである．2 章では，Web から階層的なディレクトリ構造を作成する考え方について述べ，3 章でディレクトリ構造の構築手法について述べる．4 章では，提案した手法の評価とその考察を述べる．

## 2.　提案手法の概要

階層的なディレクトリ構造を自動で作成するためには，(1) ディレクトリ間の上位-下位の階層関係を作り出し，(2) ディレクトリに Web ページを分類する必要がある．

図 1 で示したディレクトリ構造の上位-下位関係にある「全国大会」ディレクトリと「論文投稿」ディレクトリに，いくつかの学会サイトの Web ページが分類されたとする．このとき，上位ディレクトリの「全国大会」に分類されたページと下位ディレクトリの「論文投稿」に分類されたページとの間にはあらかじめ意味的な上位-下位関係があると考えられる．このことは，階層構造を作るためには，Web サイト上で意味的に上位-下位関係にあるページを抽出すればよいことを示している．そのような意味的な関係にあるページ間はリンクで結ばれている可能性が高い．例えば,「全国大会」に分類されたページは「論文投稿」のページをハイパーリンクで参照している．このようにハイパーリンクをもとに上位-下位関係にある Web ページを特定できれば，Web ページ間の上位-下位関係をそのままディレクトリの上位-下位構造に置き換えることにより，ディレクトリの上位-下位の階層関係が獲得でき，同時に Web ページの分類も行える (図 2)．

### 2.1　リンクで結ばれた Web ページと上位-下位関係

ハイパーリンクで結ばれているすべての Web ページが意味的に上位-下位関係にあるわけではないため，それを判別することが重要となる．

Web サイトを作成するとき，作成者は Web ページをフォルダごとに整理してサーバ上に配置する．その配置には作成者のある判断が働いており，その点に着目することにより，上位-下位関係にある Web ページを見つけ出すのに利用できると考えられる．

そこで，このような作成者の知識を利用するために，リンクで結ばれた Web ページ間の上位-下位関係とサーバ上の配置との関連性について調査した．調査の方法としては，名古屋大学の 4 つの研究科サイトを対象に，サイト内へのリンクをそれぞれ 200 個抽出し，そのリンク元ページとリンク先ページが上位-下位関係にあるかどうかを判定した．また，リンク元ページに対するリンク先ページのサーバ上における相対的な配置関係を 6 つに分類し，それぞれ上位-下位関係にある割合を調べた．リンク元ページに対するリンク先ページのサーバ上の配置関係の分類を図 3 に示す．図で,「上位フォルダ」,「同位フォルダ」,「下位フォルダ」はそれぞれ，リンク元ページのサーバ上のパスに比べてリンク先ページのパスが浅い，同じ，深いフォルダに位置することを表している．

調査結果を表 1 に示す．リンク数の総計が少ないのは，リンク切れのものを除いたことによる．表で,「上位-下位関係」は各配置のリンク数に対する上位-下位関係にあるリンクの割合を表している．上位-下位関係にあるリンクの 97.5%が子孫フォルダ，もしくは，同一フォルダであった．また，同一フォルダ内ページへのリンクから上位-下位関係にあるリンクを特定するために，ファイル名が “index.html” であるページからのリンク

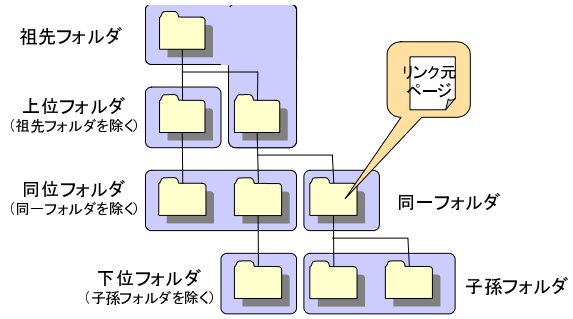


図 3　2 つのページのサーバ上の配置関係
Fig. 3　Location relation of two pages on a server

表 1　上位-下位関係の割合
Table 1　Rate of the super-sub relation

| リンク先ページの配置 | リンク数 | 上位-下位関係 (%) |
|---|---|---|
| 子孫フォルダ | 136 | 91.9 |
| 祖先フォルダ | 151 | 0.7 |
| 同一フォルダ | 246 | 58.1 |
| 下位フォルダ (子孫フォルダを除く) | 3 | 0 |
| 上位フォルダ (祖先フォルダを除く) | 77 | 2.6 |
| 同位フォルダ (同一フォルダを除く) | 152 | 2.7 |
| 総計 | 765 | 36.0 |

を調べたところ，同一フォルダ内ページへのリンクは 41 個存在し，それが上位-下位関係である割合は 85.3%であった．

### 2.2　上位ページと下位ページの表現

本研究では，上位-下位関係の上位ページと下位ページを表現する手段としてアンカーテキストに着目する．アンカーテキストは Web ページの作成者が閲覧者をリンク先ページに案内することを目的として設定しているため，リンク先ページの内容全体を簡潔に表した文字列が付与されていることが多い．そのため，各 Web ページの特徴付けは，ページに掲載される内容の種類数やテキスト量に制限のないページ自体のテキスト情報を用いるよりも，アンカーテキストを用いた方が上位-下位関係をより明確に表わすことができると考えられる．そこで，上位-下位関係の各ページはそのページを参照するリンクに対応するアンカーテキストを用いて表すことにする．なお,「戻る」といった指示語は参照先ページの内容を表さないため，ストップワードとして除外する．

## 3.　Web ディレクトリの構築

複数のサイトからの階層的なディレクトリ構造の構築の流れを図 4 に示す．各手順は以下の通りである．

（1）　各サイトからリンクで結ばれた Web ページ間の上位-下位関係を抽出する．
（2）　共通した上位-下位関係をクラスタリングする．
（3）　Web ページ間の上位-下位関係をディレクトリの上位-下位構造に置換し，ディレクトリ間の統合によりディレクトリの階層構造を構築する．
（4）　各ディレクトリの名前を決定する．

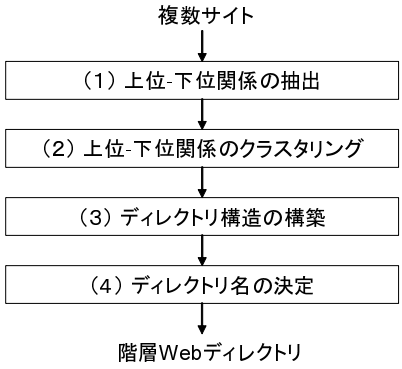


図 4　Web ディレクトリ構築の流れ
Fig. 4　Flow of constructing the Web directory

### 3.1　上位-下位関係の抽出

各サイト内のページ間に張られたリンクに対して，上位-下位関係にある Web ページを対として抽出する．2 つのページが上位-下位関係にあるかどうかは，2.1 節の調査結果に基づいて作成した規則を用いて判定する．すなわち，リンクで結ばれた 2 つのページに対して，以下に示す条件をすべて満たすリンク元ページとリンク先ページを上位-下位関係として抽出する．

（1）　リンク先ページがリンク元ページと同一のサーバ上に存在する．
（2）　リンク先ページがリンク元ページと同一のフォルダ，もしくはその子孫フォルダに存在する．
（3）　(2) でリンク先ページが同一のフォルダに存在する場合，そのフォルダ内に “index.html”が存在すれば，リンク元ページは “index.html”である．存在しなければ，リンク元ページは同一フォルダ内に最も多くリンクしているページである．

以下では，上位側の Web ページを $d_{sup}$，下位側の Web ページを $d_{inf}$ とするとき，上位-下位関係にある Web ページ対を $p = (d_{sup}, d_{inf})$ で表す．また，$d_{sup}$ を上位ページ，$d_{inf}$ を下位ページと呼ぶことにする．

### 3.2　上位-下位関係のクラスタリング

複数のサイトから抽出した上位-下位関係に対して，同じ内容の上位-下位関係同士をまとめる．ここで，上位-下位関係の内容が同じであるとは，Web ページ対の上位ページ間の内容と下位ページ間の内容がそれぞれ類似していることと定義する．Web ページ間の類似度は，そのページを参照するアンカーテキスト間の Dice 係数 [1] を計算することにより求める．Web ページはそのページを参照するアンカーテキストで表現できるため，対象とする 2 つのページをそれぞれ参照するアンカーテキスト間の類似度をすべて求めて，その最大値をページ間の類似度として採用する．すなわち，ページ $d_i$ を参照するアンカーテキストを $a_{i_s}(1 \leq s \leq m)$，ページ $d_j$ を参照するアンカーテキストを $a_{j_t}(1 \leq t \leq n)$ とするとき，$d_i$ と $d_j$ の類似度を式 (1) で定義する．

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

なお，$M_{i_s}$ は $a_{i_s}$ の名詞の数を，$M_{i_s j_t}$ は $a_{i_s}$，$a_{j_t}$ に共通して

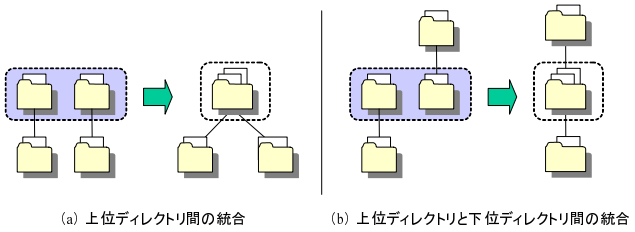


図 5　ディレクトリ対の統合
Fig. 5　Integration of the directories

出現する名詞の数を表す．

上位-下位関係間の類似度は，上位ページ間の類似度と下位ページ間の類似度を用いて表現する．2 つの Web ページ対 $p_i$ と $p_j(i \neq j)$ の上位ページ間の類似度 $sim_{sup}(p_i, p_j)$ と下位ページ間の類似度 $sim_{inf}(p_i, p_j)$ を式 (2)，及び，式 (3) でそれぞれ求める．

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

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

クラスタリングは上位-下位関係間の類似度をもとに行う．まず，初期クラスタとして Web ページ対 $p_i$ 自体からなるクラスタ $C_i$ を作成する．併合するクラスタは，上位ページ間の類似度と下位ページ間の類似度が共に閾値 $\alpha$ 以上で，かつ，その平均が最大となるクラスタであるとし，クラスタ間の類似度の計算は上位ページ間と下位ページ間それぞれに対して，完全連結法 [5] を適用することにより求める．クラスタ $C_k$ と $C_l$ の上位ページ間の類似度 $sim_{sup}(C_k, C_l)$ と下位ページ間の類似度 $sim_{inf}(C_k, C_l)$ を式 (4)，及び，式 (5) でそれぞれ定義する．

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

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

閾値を超えるクラスタが無くなった時点でクラスタリングを終了する．クラスタ内の Web ページ対の数が $m$ 個未満であるクラスタは排除する．

### 3.3　ディレクトリ構造の構築

クラスタリングでまとめた上位-下位関係をディレクトリの上位-下位構造に置き換える．これは，クラスタ $C$ に含まれる Web ページ対 $p_i$ の上位ページ $d_{i_{sup}}$ をディレクトリ $D_{sup}$ に，下位ページ $d_{i_{inf}}$ をディレクトリ $D_{inf}$ にそれぞれ分配して，ディレクトリの上位-下位構造をディレクトリの対 $P = (D_{sup}, D_{inf})$ として表すことにより実現する．以下では，$D_{sup}$ を上位ディレクトリ，$D_{inf}$ を下位ディレクトリと呼ぶことにする．

最終的なディレクトリの階層構造は，ディレクトリを順に統合することにより構築する．ディレクトリ対の上位ディレクトリ間で統合すると，最終的に図 5(a) のような親子関係にあるディレクトリ構造が生成され，上位ディレクトリと下位ディレクトリ間で統合すると，図 5(b) のような 3 代にわたるディレクトリ構造が生成される．

ディレクトリの統合におけるディレクトリ間の類似度は，ベクトル空間モデル [2] を用いて計算する．ディレクトリ $D_i$ に属

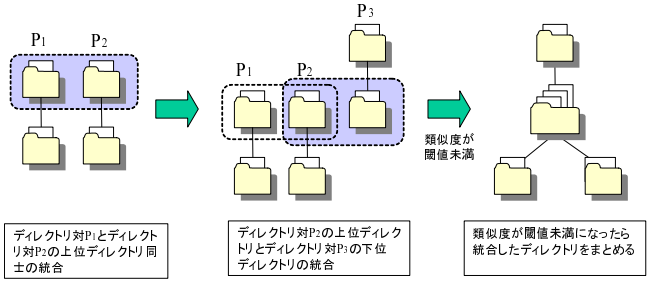


図 6　ディレクトリ構造の構築処理
Fig. 6　Construction of the directory structure

する Web ページを参照しているアンカーテキストの集合を $A_i$ とした場合，ディレクトリ $D_i$ はアンカーテキスト集合 $A_i$ 中の名詞の出現頻度を重みとした特徴ベクトルで表現する．システムが扱う名詞の集合を $\{e_1 \ldots e_N\}$ とし，名詞 $e_j$ の重み $w_{ij}$ を式 (6) で定義すると，ディレクトリ $D_i$ の特徴ベクトルは，

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

で表される．ただし，$F_{ij}$ はアンカーテキスト集合 $A_i$ における名詞 $e_j$ の頻度を表す．式 (6) を用いて，ディレクトリ対 $P_i = (D_{i_{sup}}, D_{i_{inf}})$ の上位ディレクトリ $D_{i_{sup}}$ の特徴ベクトル $\vec{x}_{i_{sup}}$ と下位ディレクトリ $D_{i_{inf}}$ の特徴ベクトル $\vec{x}_{i_{inf}}$ を求める．

ディレクトリ間の類似度には特徴ベクトルの余弦を用いる．ディレクトリ $D_i$ とディレクトリ $D_j(i \neq j)$ の類似度を式 (7) で定義する．

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

式 (7) を用いて，ディレクトリ対 $P_i$ と $P_j$ の上位ディレクトリ間の類似度 $Sim(D_{i_{sup}}, D_{j_{sup}})$ と上位ディレクトリと下位ディレクトリ間の類似度 $Sim(D_{i_{sup}}, D_{j_{inf}})$ を求める．

ディレクトリの統合は，ディレクトリ間の類似度が最大となるものから順に木構造の性質を満たすように統合することで実現する．その処理過程を図 6 に示す．手順としては，ディレクトリ間の類似度をすべて求めたのち，ディレクトリ対 $P_i$ と $P_j$ のディレクトリ間類似度 $Sim(D_{i_{sup}}, D_{j_{sup|inf}})$ が閾値 $\beta$ 以上で，その値が最大となるディレクトリ対 $P_k$，$P_l$ に対して統合妥当性を検証する．妥当であれば，ディレクトリ対 $P_k$ と $P_l$ のディレクトリ $D_{k_{sup}}$ と $D_{l_{sup|inf}}$ を統合する．

ここで，統合妥当性とは，ディレクトリの統合により拡張されたディレクトリ構造が木構造の性質を満たしていることを保証するものであり，最終的なディレクトリ構造が以下の条件を満たすように統合を行うものとする．

（1）　親ディレクトリは高々1 つである．

（2）　ディレクトリ構造は非循環である．

ディレクトリを統合した，もしくは，統合妥当性を満たさない場合は，ディレクトリ間の類似度が次に高いディレクトリ対に処理を移して同様の操作を繰り返す．この操作をすべてのディレクトリ間類似度が閾値 $\beta$ 未満になるまで繰り返して，階

表 2　実験に使用したサイトとデータ量
Table 2　Experimental data and its site

| ID | サイト | ページ数 | リンク数 |
|---|---|---|---|
| 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 |

層的なディレクトリ構造を構築する．

類似度が閾値未満になった時点で，統合されたディレクトリをまとめて新たなディレクトリを作成する．統合するディレクトリを $D_1,\cdots,D_n$ とし，統合後のディレクトリを $D_r$ とすると，ディレクトリ $D_i$ に属する Web ページの集合を $W_i$ とした場合，統合後のディレクトリ $D_r$ に属する Web ページの集合 $W_r$ は式 (8) で定義される．

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

### 3.4　ディレクトリ名の決定

ディレクトリの名前は，そのディレクトリに属する Web ページを参照しているアンカーテキストの集合をもとに決定する．ディレクトリに対応するアンカーテキスト集合に共通して出現し，かつ，ある程度の長さをもった語句をそのディレクトリの名前とする．

まず，ディレクトリ $D_i$ に対応するアンカーテキスト集合 $A_i = \{a_{i_1},\cdots,a_{i_M}\}$ から任意の部分形態素列 $s_{ij}$ を抽出し，これらをディレクトリ名の候補とする．各部分形態素列 $s_{ij}$ に対して，$A_i$ のアンカーテキスト $a_{i_k}$ における包含率 $Cover(s_{ij},a_{i_k})$ を式 (9) より求めたのち，式 (10) で定義する平均包含率 $Cover_{ave}(s_{ij},A_i)$ を計算し，その値が最大となる $s_{ij}$ をディレクトリ名とする．

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

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

なお，$|a_{i_k}|$ はアンカーテキスト $a_{i_k}$ の形態素数，$F^i_{jk}$ は部分形態素列 $s_{ij}$ とアンカーテキスト $a_{i_k}$ に共通して出現する形態素数，$|s_{ij}|$ は部分形態素列 $s_{ij}$ の形態素数，$M$ はアンカーテキスト集合 $A_i$ 中のアンカーテキストの数を表す．

## 4.　評価実験

### 4.1　実験方法

ハイパーリンクに基づいて複数のサイトから階層的なディレクトリ構造を構築する手法の実現可能性を確認するために，評価実験を行った．実験には，表 2 に示した名古屋大学の 5 つの研究科サイトを使用した．実験に使用したサイトの Web ページ数とサイト内リンク数を表 2 に示す．実験では，各サイトのトップページからリンクを辿って上位-下位関係を抽出した．上位ページと下位ページを表現するアンカーテキストは，各サイト内の Web ページから取得した．なお，閾値の設定は，上位-下位関係のクラスタリングに使用する閾値 $\alpha$ を 0.5，ディレクトリ構造の構築に使用する閾値 $\beta$ を 0.6 とし，クラスタリングの結果，Web ページ対が 2 個未満のクラスタは処理の対象外とした．また，形態素解析器には「茶筌」[6] を使用した．

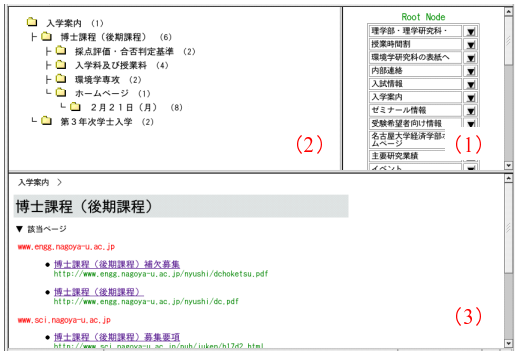


図 7　階層ディレクトリ構造の出力例
Fig. 7　Example of the system output

### 4.2　実験結果

構築したディレクトリ構造の出力例を図 7 に示す．図 7 の結果は，ディレクトリの統合を少なくとも一度は行って生成された 13 個のディレクトリ構造のうちの一部である．なお，図中の (1) は生成されたディレクトリ構造のルートディレクトリ，(2) は特定したディレクトリ構造の全体図，(3) は特定したディレクトリに分類された Web ページへのリンクをそれぞれ表している．

結果の例として，入学情報に関してまとめられたディレクトリ構造を表 3，及び，表 4 に示す．表中で「階層」はディレクトリの階層構造を，「ページ数」はそのカテゴリに分類されたページ数を表している．なお，ページ数は各サイトごとに表しており，それぞれ表 2 の ID と対応する．表より，複数サイトのページが 1 つのディレクトリ構造に分類されていることがわかる．また，ディレクトリの階層構造に対しても，ある程度の上位-下位構造が形成されており，ハイパーリンクの関係を用いて複数サイトから階層的なディレクトリ構造を構築する手法の実現可能性を確認できた．

### 4.3　考　　察

実験結果をもとに以下の項目に分けて考察する．

#### 4.3.1　階層構造の妥当性

生成されたディレクトリの階層構造は，表 3，表 4 からも見られるように上位ディレクトリと下位ディレクトリの間にはある程度の妥当性が確認できる．しかし，構築されたディレクトリ構造の中には，不適切な上位-下位構造をもつディレクトリもいくつか観察された．本手法では，ディレクトリの階層構造を構築するときに，統合するディレクトリをアンカーテキスト集合の類似度により決定している．そのため，ディレクトリ内ページの内容が異なっていても，アンカーテキストの語句が類似していれば統合されてしまう．例えば，ディレクトリが “情報学専攻” と表現された場合，その語句からディレクトリ内の

表 3　生成されたディレクトリ構造の結果 (1)
Table 3　Constructed directory structure (1)

| 階層 | ディレクトリ名 | ページ数 | | |
|---|---|---|---|---|
| | | I | II | IV |
| 1 | 入学案内 | 1 | 0 | 0 |
| 1-1 | 博士課程 (後期課程) | 2 | 2 | 2 |
| 1-1-2 | 採点評価・合否判定基準 | 0 | 2 | 0 |
| 1-1-3 | 入学科及び授業料 | 0 | 4 | 0 |
| 1-1-4 | 環境学専攻 | 0 | 2 | 0 |
| 1-1-5 | ホームページ | 0 | 0 | 1 |
| 1-1-5-1 | 2 月 21 日 (月) | 0 | 0 | 8 |
| 1-2 | 第 3 年次学士入学 | 2 | 0 | 0 |

表 4　生成されたディレクトリ構造の結果 (2)
Table 4　Constructed directory structure (2)

| 階層 | ディレクトリ名 | ページ数 | | |
|---|---|---|---|---|
| | | I | II | V |
| 1 | 入試情報 | 1 | 1 | 1 |
| 1-1 | 博士課程 | 2 | 2 | 10 |
| 1-1-1 | 経済学修士号への道 | 0 | 0 | 2 |
| 1-2 | 募集要項の請求方法 | 0 | 1 | 1 |
| 1-3 | 都市環境学専攻 | 0 | 2 | 0 |

ページの内容を判断することは難しい．現段階では，ディレクトリを統合するときは，統合するディレクトリ間の関係しか考慮していないが，上位-下位関係を構成しているもう一方のディレクトリとの関係も考慮に入れることにより，このような現象を抑えることができると考えられる．

**4.3.2**　ディレクトリ分類の正しさ

各ディレクトリにおける分類の正しさの判断は，ここではディレクトリに分類されたページの内容がそのディレクトリ名に即していれば正解として行った．この基準に基づくと，表 3, 表 4 のディレクトリ構造では，49 文書中 35 文書が正しく分類されていた．例えば，表 4 の「博士課程」は前期・後期課程の募集要項や博士課程への留学生選考に関する内容のページはすべて正しいと判断している．一方，表 3 の「博士課程 (後期課程)」に分類されている前期課程の募集要項に関するページは適切ではないとしている．

次に，再現性の面からみた場合，同じ上位-下位関係にあるページが 1 つのディレクトリにまとめられないことがある．また，本実験で構築された 13 個のディレクトリ構造のうち 8 個は，単一のサイト内ページで構成されている．これらの原因の 1 つとして，意味的に同じ上位-下位関係であっても，アンカーテキストによる表現の違いによりクラスタ化されないことが挙げられる．例えば，入学情報全般と募集要項といった上位-下位関係のページが存在したとしても，前者のページを表現する語句が “入学案内” と “受験希望者向け情報” であったときは同じ上位-下位関係として見なされない．複数のサイトに共通する上位-下位関係をできるだけ多くディレクトリにまとめるためには，アンカーテキストを用いるだけでは限界があり，それ以外の情報も用いた方法を検討する必要がある．

表 5　アンカーテキスト集合からのディレクトリ名の決定
Table 5　Directory names generated by the anchor texts

| 博士課程（後期課程） |
|---|
| 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 |
| 募集要項の請求方法× 2，各種募集要項の請求方法× 3 |

**4.3.3**　ディレクトリ名の正しさ

ディレクトリ名はアンカーテキスト集合をもとに決定している．例えば，表 3 の「博士課程 (後期課程)」や表 4 の「募集要項の請求方法」は，表 5 に示したアンカーテキスト集合から生成されたものである．ディレクトリ名を決定する条件にほぼ合致する語句を抽出することはできているものの,「博士課程 (後期課程)」の例では “後期課程” の記述が “前期課程” よりも多いためにディレクトリ名として選択されてしまうなど，必ずしも正しく表現しているとは言い難い場合もある．また，ディレクトリ名が助詞から始まるなど，文法的におかしなものも存在しており，決定に文法的制約を取り入れるなどの対処が必要である．

## 5. ま と め

本稿では，Web のハイパーリンク構造とアンカーテキストをもとに，複数サイトから階層的なディレクトリ構造を構築して Web ページを分類する手法を提案した．また，複数のサイトを用いて実施した評価実験について述べた．実験では，複数サイトのページが分類されたディレクトリ構造が構築され，本手法の実現可能性を確認できたが，構造的に不適切なディレクトリ構造もいくつかあった．今後は，上位-下位関係にある階層構造を適切に構築するために，アンカーテキスト以外の情報も用いて関係を表現する必要がある．また，実験のデータ量を増やすことにより手法の実用可能性を検証する予定である．

## 文　献


[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] 原田 晶紀，佐藤 進也，風間 一洋: WWW ページ間の階層構造の推定と検索システムへの応用, 情報処理学会研究報告, DBS118-14，pp.105-112 (1999).

[4] 小島 秀一，高須 淳宏，安達 淳: Web ページ群の構造解析とグループ化，NII Journal, No.4, pp.23-35 (2002).

[5] 神嶌 敏弘: データマイニング分野のクラスタリング手法 (1) －クラスタリングを使ってみよう！－, 人工知能学会誌, Vol.18, No.1, pp.59-65 (2003).

[6] 松本 裕治，北内 啓，山下 達雄，平野 善隆，松田 寛，高岡 一馬，浅原 正幸: 形態素解析システム『茶筌』, version2.2.9, 使用説明書 (2002).