Web ディレクトリ構築のためのハイパーテキストの階層的分類

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

suzuki@dl.itc.nagoya-u.ac.jp

1 はじめに

Web 上には大学サイトや趣味のサイトなど様々な分野のサイトが存在している．Web からこのような関連したサイトを探す方法として，Yahoo!カテゴリ[1]に代表される Web ディレクトリの使用が挙げられる．しかし，このような Web ディレクトリは Web 全体を対象としているため，分類はサイト単位で行なわれることが多く，特定分野の複数のサイトにまたがったページを見つけるためのツールとしては，必ずしも十分ではない．

そこで本論文では，複数の関連サイトのページを整理したサイト間 Web ディレクトリの構築支援のために，Web ディレクトリの断片を生成する手法を提案する．ハイパーリンクをもとに概念的に上位-下位関係にある Web ページを対にして抽出し，階層的に分類する．分類実験により，本手法の実現可能性を確認した．

2 サイト間 Web ディレクトリ

図 1 の Web ディレクトリは，サイト間にまたがった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，階層化したものである．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，その分野のサイト内情報の全体像を把握できる，といった利点がある．本論文では，このような特定分野のサイト群を対象に構築した Web ディレクトリを**サイト間 Web ディレクトリ**と呼ぶ．しかし，その階層構造は対象分野によって異なるため，開発者による階層構造の設計や，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．

本研究では，サイト間 Web ディレクトリの構築支援のために，Web ディレクトリの部分グラフを生成する．部分グラフとは，木構造の Web ディレクトリをグラフとしたとき，その部分を構成するディレクトリ構造である．開発者は複数の部分グラフを組み合わせることにより，対象とするサイト群に適した Web ディレクトリを構築でき，その設計や分類にかかる負担を軽減できる．

3 部分グラフの生成

複数の関連サイトから部分グラフを生成する様子を図 2 に示す．まず，Web ページ間のハイパーリンクをもとに，概念的に上位-下位関係にある Web ページを対にして抽出する．ここで，Web ページの概念的な上位-下位関係とは，ページによる全体と部分，概要と詳細といった関係をいう．同じ内容のページ対をクラスタリングしたのち，これをディレクトリの上位-下位関係とすることにより，ページが分類された上位-下位構造のディレクトリを作成する．内容が類似したディレクトリを統合してその構造を階層化することにより，部分グラフを生成し，各ディレクトリに対してラベルを付与する．以下の節では，手法の詳細について説明する．

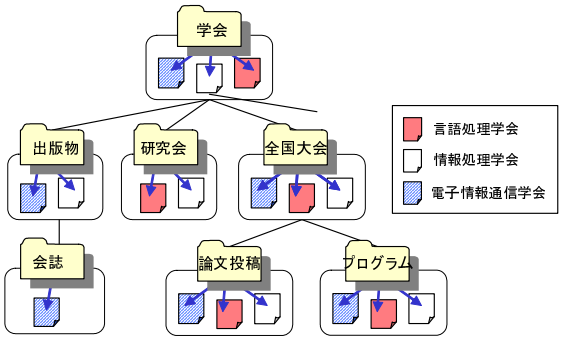

図 1: 学会サイトによるサイト間 Web ディレクトリ

3.1 上位-下位関係の抽出

複数の関連サイトから上位-下位関係にある Web ページを対として抽出する．これは，リンクで結ばれた同一サーバ上の 2 つのページに対して，以下の条件を満たすリンク元ページとリンク先ページを Web ページ対として抽出する．

1. リンク先ページがリンク元ページの子孫フォルダに存在する．
2. リンク先ページがリンク元ページと同一のフォルダに存在する場合，
   (a) リンク元ページが中心ページである．中心ページとはファイル名が “index.html”であるページとし，もし，存在しなければ，同一フォルダ内のページへのリンクが最多のページとする．
   (b) リンク元ページは中心ページからリンクされており，一方，リンク先ページは中心ページからリンクされていない．

[1] http://dir.yahoo.co.jp/

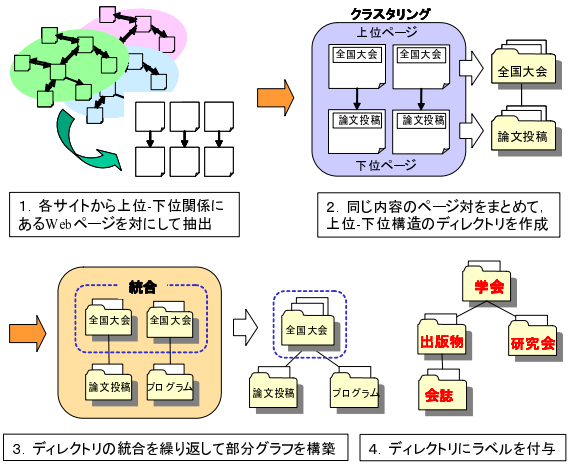

図 2: 提案手法の概要

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} とし，その 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．

3.2 Web ページ対のクラスタリング

サイト間で共通する内容のページをまとめるために，各サイトから抽出したページ対のうち，同じ内容のページ対をまとめる．ここで，ページ対の内容が同一であるとは，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．なお，Web ページは，その内容を端的に表現している箇所として，そのページを参照しているアンカーテキストを用いて表現する．ページ間の類似度は，各ページを参照するアンカーテキスト間の Dice 係数の最大値とする．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，$M_{i_s j_t}$ は a_{i_s}，a_{j_t} に共通して出現する名詞の数を表す．

次に，Web ページ対 p_i 自体からなる初期クラスタ C_i を作成する．クラスタリングは，クラスタの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 α 以上で，その平均が最大となるものから順に行なう．なお，クラスタの類似度の計算には，上位ページ間と下位ページ間それぞれに対して，群平均法 [2] を適用する．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 (2)，及び，式 (3) でそれぞれ定義する．

$$sim_{sup}(C_k, C_l) = \frac{1}{n_k n_l} \sum_{p_i \in C_k} \sum_{p_j \in C_l} (sim(d_{i_{sup}}, d_{j_{sup}})) \quad (2)$$

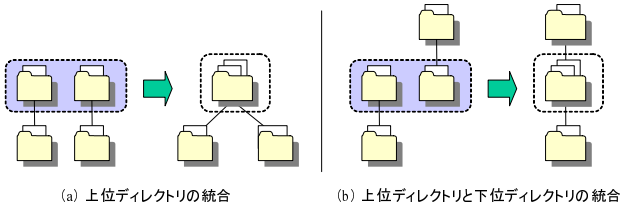

図 3: ディレクトリの統合

$$sim_{inf}(C_k, C_l) = \frac{1}{n_k n_l} \sum_{p_i \in C_k} \sum_{p_j \in C_l} (sim(d_{i_{inf}}, d_{j_{inf}})) \quad (3)$$

なお，n_k は C_k に含まれる Web ページ対の数を表す．

3.3 部分グラフの構築

各クラスタは Web ページ対の集合であり，クラスタ内の上位ページ集合と下位ページ集合をそれぞれ 1 つのディレクトリとすることにより，ディレクトリの上位-下位構造を形成する．上位側のディレクトリを D_{sup}，下位側のディレクトリを D_{inf} とし，そのディレクトリ対を $P = (D_{sup}, D_{inf})$ で表す．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．

部分グラフはディレクトリを統合して構築する．ディレクトリ対の上位ディレクトリを統合すると，図 3(a) のような親子関係にある階層構造が生成され，上位ディレクトリと下位ディレクトリを統合すると，図 3(b) のような 3 代の階層構造が生成される．

ディレクトリ間の類似度にはベクトル空間モデル [3] を適用する．ディレクトリ D_i 内の Web ページを参照するアンカーテキストの集合を A_i とした場合，D_i は 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．出現する名詞の集合を $\{e_1, \cdots, e_N\}$ とし，名詞 e_j の重み w_{ij} を式 (4) で定義すると，D_i の特徴ベクトルは，$\vec{x_i} = (w_{i1}, w_{i2}, \cdots, w_{iN})$ で表される．

$$w_{ij} = F_{ij} \quad (4)$$

なお，F_{ij} は A_i における e_j の頻度を表す．式 (4) 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 D_i とディレクトリ $D_j (i \neq j)$ の類似度を式 (5) で定義する．

$$Sim(D_i, D_j) = \frac{\vec{x_i} \bullet \vec{x_j}}{|\vec{x_i}||\vec{x_j}|} \quad (5)$$

式 (5) 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閾値 β 以上で，その値が最大となるものから順に，部分グラフが木構造になるように統合することにより実現す

表 1: 実験に使用したサイトとデータ量

サイト	ページ数
北海道大学大学院情報科学研究科	118
東北大学大学院情報科学研究科	256
名古屋大学大学院情報科学研究科	140
京都大学大学院情報学研究科	69
九州大学システム情報科学研究院	524
電気通信大学情報システム学研究科	186
筑波大学システム情報工学研究科	222
奈良先端科学技術大学院大学情報科学研究科	148

表 2: 正解 Web ディレクトリ

ID	ディレクトリ名	ページ数
A	情報系研究科	8
A-1	研究科長挨拶	6
A-2	研究科紹介	10
A-3	専攻・講座一覧	11
A-4	教員一覧	9
A-5	講義一覧	10
A-6	入試情報	16
A-6-1	博士課程 (前期) 募集要項	18
A-6-2	博士課程 (後期) 募集要項	17
A-7	アクセス	8
A-8	サイトマップ	6

る．ディレクトリの統合により木構造が崩れる場合は，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行なう．これをすべてのディレクトリ間類似度が閾値 β 未満になるまで繰り返す．

類似度が閾値未満になった時点で，統合する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 $D_1, \cdots, D_n$，新たに生成するディレクトリを D_r とし，ディレクトリ D_i に属する Web ページの集合を W_i とすると，D_r に属する W_r は式 (6) で定義される．

$$W_r = \bigcup_{1 \leq i \leq n} W_i \tag{6}$$

3.4 ディレクラベルの付与

ディレクトリラベルは，そのディレクトリ内の Web ページを参照するアンカーテキスト集合から作成する．まず，ディレクトリ D_i に対するアンカーテキスト集合 $A_i = \{a_{i_1}, \cdots, a_{i_M}\}$ からラベルとして適切な任意の部分形態素列 s_{ij} を抽出し，これらを D_i のラベルの候補とする．ここで，適切なラベルとは，形態素列の先頭や末尾に助詞や副詞などの品詞が出現しないものとする．抽出された 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 (7) より求め，式 (8) で定義する平均包含率 $Cover_{Ave}(s_{ij}, A_i)$ を計算する．平均包含率が最大となる s_{ij} を D_i のラベルとする．

$$Cover(s_{ij}, a_{ik}) = \begin{cases} \frac{F^i_{jk}}{|a_{i_k}|} & (|s_{ij}| \leq F^i_{jk}) \\ 0 & (otherwise) \end{cases} \tag{7}$$

$$Cover_{Ave}(s_{ij}, A_i) = \frac{\sum_{k=1}^{M} Cover(s_{ij}, a_{i_k})}{M} \tag{8}$$

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

4 実験と評価

4.1 実験方法

本研究で対象とするサイト間 Web ディレクトリは，特定分野における関連サイトのページを分類したものである．そこで，実験には，分野の同じ関連サイトとして 8 つの国立大学の情報系研究科サイトを使用した．サイト名とその Web ページ数を表 1 に示す．実験では，各ページを表現するアンカーテキストは，各サイト内の Web ページから取得する．なお，閾値の設定は，$\alpha = 0.6$（3.2 節参照），$\beta = 0.8$（3.3 節参照）とし，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」[4] を使用した．

実験の評価値として分類の精度と再現率を定義する．これは，あらかじめ正解の Web ディレクトリを作成しておき，各正解ディレクトリに対して，すべての部分グラフのディレクトリと比較することにより求める [1]．正解ディレクトリ A_i に対して，部分グラフの各ディレクトリ D_j における F 値を求め，その最大値を A_i に対する評価値とする．A_i に対する D_j の精度 P_{ij} と再現率 R_{ij} を式 (9) で定義し，その F 値を式 (10) で定義する．

$$P_{ij} = \frac{x_{ij}}{d_j}, \qquad R_{ij} = \frac{x_{ij}}{a_i} \tag{9}$$

$$F(A_i, D_j) = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}} \tag{10}$$

ここで，d_j は D_j に含まれるページ数，a_i は A_i に含まれるページ数，x_{ij} は A_i と D_j に含まれるページ数を表す．さらに，正解ディレクトリの評価値をすべて足したものを正解 Web ディレクトリ全体の評価値とする．これを式 (11) で定義する．また，このときの正解 Web ディレクトリに対する精度 P と再現率 R を式 (12) で定義する．

$$TotalScore = \sum_i \left(\frac{a_i}{\sum_i a_i}\right) \max_j F(A_i, D_j) \tag{11}$$

$$P = \sum_i \left(\frac{a_i}{\sum_i a_i}\right) \frac{x_{ik}}{d_k}, \qquad R = \sum_i \left(\frac{a_i}{\sum_i a_i}\right) \frac{x_{ik}}{a_i} \tag{12}$$

このとき，d_k は対象とする A_i に対して F 値を最大とするディレクトリ D_k に含まれるページ数とする．

実験に使用した 8 つのサイトから人手で正解 Web ディレクトリを作成した．作成した正解 Web ディレクトリを表 2 に示す．なお，表中で「ID」はディレクトリの階層を表している．

図 4: 部分グラフの出力例

4.2 実験結果

実験の結果，生成された部分グラフの数は 33 個であった．また，部分グラフあたりの平均ディレクトリ数は 4.7 個，ディレクトリあたりの平均ページ数は 3.4 ページであった．生成された部分グラフの出力例を図 4 に示す．図中の (1) は各部分グラフのルートディレクトリ，(2) は選択された部分グラフの全体図，(3) は選択され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．

正解 Web ディレクトリ全体に対する評価結果を表 3 に示す．再現率は必ずしも高くはない．生成された部分グラフを観察すると，同じ内容のページでまとまってはいるものの，複数のディレクトリに分散して存在していることが多く見られた．しかし，“A-8 サイトマップ”に対応するディレクトリには 4 サイトの正解ページが，“A-1 情報系研究科”では 3 サイトの正解ページが分類されているように，複数の関連サイトのページが正しく分類されたディレクトリも存在する．

4.3 考察

生成された部分グラフの妥当性を評価するために，「階層構造の適切さ」と「ラベルの正確さ」についても評価した．これらの評価は，0.7 以上の精度をもつディレクトリを対象に行った．

4.3.1 階層構造の適切さ

評価方法 すべての部分グラフで親子関係にあるディレクトリ内のページ内容を比較し，それらが適切な上位-下位関係を形成しているかを判定する．判定は，表 4 の判定基準のいずれかに分類することにより行う．

評価結果 評価結果を表 4 に示す．上位-下位関係があると判定された割合が最も多いが，同位の関係であると判定された割合も多かった．同位の関係とは，親ディレクトリと子ディレクトリが共に同じ内容のページを含む関係である． この結果は，同じ内容のページが 1 つのディレクトリにまとまってない場合でも，それらは近くのディレクトリに位置している傾向があることを示している．これらのディレクトリのページをまとめることができれば，分類の再現率向上に繋がると考えられる．

表 3: ディレクトリへの分類結果

適合率 (%)	再現率 (%)	F 値
72.9	30.3	40.5

表 4: 階層構造の適切さの評価結果

判定	個数	割合 (%)
上位-下位関係がある	34	43.6
同位の関係である	29	37.2
上位-下位関係がない，逆転する	15	19.2

表 5: ラベルの正しさの評価結果

判定	個数	割合 (%)
ラベルが正確で理解できる	64	50.4
ラベルが一部異なるが理解できる	36	28.3
ラベルが異なるため理解できない	27	21.3

4.3.2 ディレクトリラベルの正確さ

評価方法 すべての部分グラフの各ディレクトリに対して，ラベルがそのディレクトリ内のページ内容を適切に表現しているか，ラベルからページ内容を理解できるかを判定する．判定は，表 5 の判定基準のいずれかに分類することにより行なう．

評価結果 評価結果を表 5 に示す．ラベルからページ内容が分かると判定されたものは 78.7%であり，多くがページ内容を反映したものであることが分った．

5 まとめ

本論文では，サイト間 Web ディレクトリの部分グラフを生成する手法を提案した．また，実験により本手法の実現可能性を確認した．今後の課題としては，現在，上位-下位関係にある Web ページはリンクにより推定しているが，サイトによるリンク構造の違いから抽出できないページ対も存在する．テキスト情報も用いて上位-下位関係を判定できれば，多くのページ対を抽出でき，分類の再現率の改善にも繋がると考える．

参考文献

[1] 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.

[2] Pang-Ning Tan, Michael Steinbach and Vipin Kumar: Introduction to Data Mining. Addison-Wesley, 2005.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto: Modern Information Retrieval. Addison Wesley, 1999.

[4] 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).