3C2-02

アンカーテキストとハイパーリンクに基づくWeb文書の階層的分類

Hierarchical Categorization of Web Documents based on Anchor Text and Hyperlink

鈴木 祐介[*1] 松原 茂樹[*2] 吉川 正俊[*2]
Yusuke Suzuki Shigeki Matsubara Masatoshi Yoshikawa

[*1]名古屋大学 大学院 情報科学研究科
Graduate School of Information Science, Nagoya University

[*2]名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University

This paper proposes a method for automatically constructing the hierarchical Web directories and categorizing Web documents into them. The method finds two Web pages, which are a super-sub relation each other, from the hyperlink structure, and describes the relation as a pair of the anchor texts referring their documents. After clustering the same relation, the method constructs the hierarchical directories by connecting the similar anchor texts, which are one of the pairs of the anchor texts. The Web pages are categorized into the directory which consists of the anchor texts referring their pages. As a result of the experiment, the hierarchical directories reflecting information on several sites were constructed.

1. はじめに

ディレクトリ型検索エンジンは，Web ページを内容別に分類して提示することにより，目的とする情報へのアクセスを支援する．現在では，Yahoo![*1]に代表されるように多くのポータルサイトがディレクトリを提供している．しかし，大学内 Web ページのような特定の分野に対して既存のディレクトリを適用することは，それらの間でカテゴリの粒度が異なることもあり難しい．また，新たにディレクトリを設計する場合，粒度が細かいカテゴリでは，上下関係を決めることは容易ではない．このため，サイトごとの特性を踏まえ，対象とする分野に適した階層ディレクトリを自動的に構築する技術が望まれる．さらに，構築したディレクトリに Web ページを自動分類できれば，人手による分類のコスト削減にも繋がる．

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階層的なディレクトリ構造を作成し，Web ページを分類する手法を提案する．上位-下位関係にある 2 つの Web ページをリンク構造から特定し，それらを参照しているアンカーテキストを用いて対の形で表現する．類似したアンカーテキスト同士を連結することにより，階層的なディレクトリ構造を構築し，同時に Web ページの分類を実現する．

2. 提案手法の概要

本研究では，複数のサイトから統一的な階層ディレクトリ構造を自動構築し，ディレクトリ内の各カテゴリに Web ページを分類する．各サイトから統一的な階層構造を構築するために，Web ページ間の上位-下位関係を抽出し，それをもとに階層化する．

Web ページは互いにリンクで連結されており，リンクで結ばれた Web ページ間には何らかの内容上の関連があると考えられる．そこで，Web のハイパーリンクに注目し，ページ間の上位-下位関係を抽出する．例えば，図 1 の場合，ページ “entrance.html” は入学に関する総合的な情報が記述されており，各リンクで右のページを参照している．また，右に示したページはそれぞれ “entrance.html” の詳細情報となっている．

連絡先: 鈴木祐介，名古屋大学大学院情報科学研究科社会システム情報学専攻吉川研究室，〒464-8601 名古屋市千種区不老町，TEL：(052)789-1532，E-mail：suzuki@dl.itc.nagoya-u.ac.jp

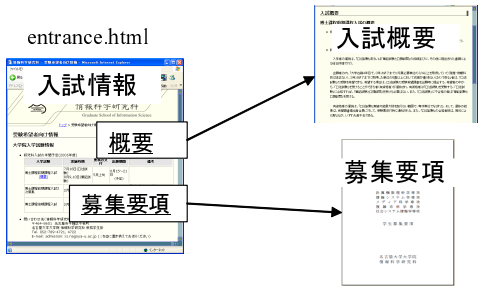

図 1: Web ページの上位-下位関係

このような内容的に包含したページ間の関係を上位-下位関係として捉える．同様な上位-下位関係が各サイトに複数存在していれば，それらを一つにまとめることにより，複数のサイトに適したディレクトリ構造が構築できる．

ここで，上位-下位関係が同様であるとは，リンクで結ばれた参照元ページと参照先ページがそれぞれ同じような内容を表していることをいう．本研究では，上位-下位関係にあたる Web ページの内容を表現するためにアンカーテキストに着目する．アンカーテキストはユーザをリンク先ページに案内する役割を担っており，リンク先ページの内容を端的に表現していることが多い．そのため，Web ページの内容を参照元ページのアンカーテキストを用いて表現することが可能である [1]．そこで，上位-下位関係にある 2 つのページを，各ページを参照しているアンカーテキストを用いて対の形で表現する．例えば，図 2 では，ページ “entrance.html” を，それを参照しているアンカーテキスト「入試情報」で表し，“doctor.html” をアンカーテキスト「博士課程」で表す．そして，それらを対にして “(入試情報，博士課程)” とする．

このような上位-下位関係を Web のリンク構造からすべて抽出し，同じ関係をもつアンカーテキスト対同士をまとめる．これらとの間で同じ内容のアンカーテキスト同士を連結することにより，階層的なディレクトリ構造を構築する．例えば，アンカーテキスト対 “(入試情報，博士課程)” のまとまりと “(入試情報，学部)” のまとまりがあった場合，アンカーテキスト対の親にあたる “入試情報” で両者を連結することにより，親に

*1 http://www.yahoo.co.jp/

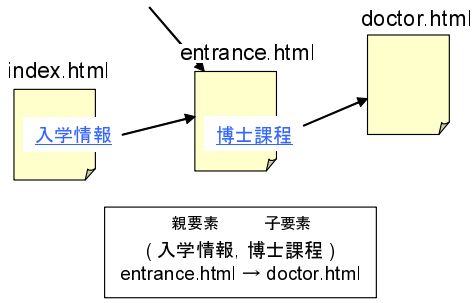

図 2: アンカーテキスト対の抽出例

“入試情報”，子に “博士課程” と “学部” をもつ木構造が形成される．この操作を繰り返し行うことにより，階層的なディレクトリ構造を構築する．

最後に，各カテゴリを形成しているアンカーテキストが本来参照していた Web ページを，そのカテゴリに振り分けることにより Web ページを分類する．

3. Web 文書の階層的分類

本節では，前節で述べた提案手法を実現するための詳細について述べる．

3.1 アンカーテキスト対の抽出

Web のリンク構造に対して起点となる Web ページから順にリンクを辿りながら，上位-下位関係にある 2 つの Web ページを，参照するアンカーテキストを対にして “($Anchor1, Anchor2$)” の形で表す．なお，$Anchor1$ を親要素，$Anchor2$ を子要素と呼ぶことにする．同時に各アンカーテキストが参照しているページの URL も記録する．

2 つの Web ページが上下関係にあるかどうかの判別は，サーバ上における Web ページの配置とそれらの間のリンク構造，及び，アンカーテキストの記述をもとに行う．Web ページをサーバ上にどのように配置して，リンク付けるかにはある程度共通した傾向がある [2]．例えば，大まかな内容ごとにディレクトリを設定し，詳細な情報は 1 つ下にディレクトリを設けて設定することが多い．また，同じディレクトリ内にもトップページのような中心となるページが存在する．そこで，以下の条件のすべてを満たすリンクから得られるアンカーテキスト対を上位-下位関係をもつものとして抽出する．

1. リンク先ページが同一サーバ上に存在する．
2. リンク先ページが同一ディレクトリ，もしくはその子孫ディレクトリに存在する．
3. 2 でリンク先ページが同一ディレクトリに存在する場合，リンク元ページはメインページである．メインページとはファイル名が “index.html” であるページとし，それが存在しなければ，同一ディレクトリ内に最も多くリンクしているページとする．
4. アンカーテキストが内容的に意味のある表現である．つまり,「次へ」や「戻る」といった指示語のリンク以外からなる．

3.2 アンカーテキスト対のクラスタリング

複数サイトから対象分野に適したディレクトリ構造を構築するためには，その分野で多く用いられている Web ページの上位-下位関係を考慮する必要がある．そこで，抽出したアンカーテキスト対のうち，その親要素と子要素が共に類似しているものをクラスタにまとめ，各クラスタをアンカーテキスト対の集合として表現する．ここで，クラスタに含まれるアンカーテキスト対のすべての親要素をクラスタの親要素群，すべての子要素をクラスタの子要素群と呼ぶことにする．クラスタリングの結果，一定数以上のアンカーテキスト対を含むクラスタをディレクトリの構築に使用し，アンカーテキスト対の少ないクラスタは排除する．

まず，クラスタ間の類似度を定義する前に，クラスタを構成するアンカーテキスト対間の類似度を定義する．アンカーテキスト対間の類似度は，アンカーテキスト対の親要素間と子要素間それぞれに対して求める．2 つのアンカーテキスト対 p_i と $p_j (i \neq j)$ の親要素間の類似度 $sim_{parent}(p_i, p_j)$ と子要素間の類似度 $sim_{child}(p_i, p_j)$ を式 (1)，及び，式 (2) でそれぞれ定義する．

$$sim_{parent}(p_i, p_j) = \frac{2F_{ij}(parent)}{F_i(parent) + F_j(parent)} \quad (1)$$

$$sim_{child}(p_i, p_j) = \frac{2F_{ij}(child)}{F_i(child) + F_j(child)} \quad (2)$$

なお，$F_i(parent)$ は p_i の親要素に出現する名詞の頻度，$F_{ij}(parent)$ は p_i と p_j の親要素に共通して出現する名詞の頻度を表す．子要素 $child$ についても同様である．

次に，アンカーテキスト対 p_i 自体からなるクラスタを作成する．クラスタリングは，クラスタの親要素群間の類似度と子要素群間の類似度が共に閾値 α 以上であり，その平均が最大となるものに対して行う．なお，クラスタ間の類似度はクラスタの親要素群間と子要素群間それぞれに対して，最長距離法 [3] を適用することにより求める．クラスタ C_k と C_l の親要素群間の類似度 $sim_{parent}(C_k, C_l)$ と子要素群間の類似度 $sim_{child}(C_k, C_l)$ を式 (3)，及び，式 (4) でそれぞれ定義する．

$$sim_{parent}(C_k, C_l) = \max_{p_i \in C_k, p_j \in C_l} (sim_{parent}(p_i, p_j)) \quad (3)$$

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

クラスタリングの結果，アンカーテキスト対が 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

各クラスタはアンカーテキストの集合であり，クラスタの親要素群と子要素群をそれぞれ 1 つのカテゴリと見なすことにより，親子関係をもつディレクトリ構造を形成できる．さらに，クラスタの各要素群を類似したもの同士で順に統合して木構造を生成することにより，ディレクトリ構造を拡張することが可能になる．例えば，クラスタの親要素群同士を統合すると図 3 のような親子関係が生成され，親要素群と子要素群を統合すると図 4 のような 3 代の関係が生成される．

まず，クラスタの要素群間の類似度に基づいて要素群の統合を行うために，クラスタの各要素群をそれぞれ特徴ベクトルとして表現する．クラスタの親要素群と子要素群それぞれに対し，アンカーテキスト中の名詞の頻度による重みを与えることにより，各要素群の特徴ベクトルを作成する．クラスタ C_k の要素群 c_k^x における名詞 $e_j (j = 1, 2, \cdots, N)$ の重み w_{kj}^x を式 (5) で定義し，クラスタ C_k の要素群の特徴ベクトル

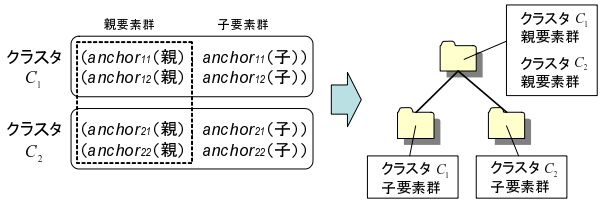

図 3: ディレクトリ構造の構築（親要素群同士の統合）

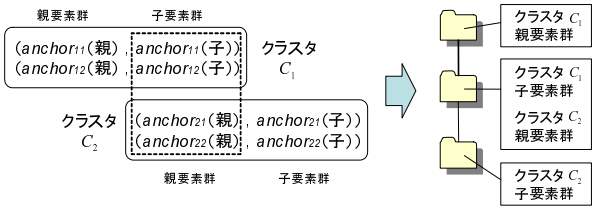

図 4: ディレクトリ構造の構築（親要素群と子要素群の統合）

$\vec{x}_k^x = (w_{k1}^x, w_{k2}^x, \cdots, w_{kN}^x)$ を求める．なお，x は 1，もしくは 2 をとり，1 の場合は親要素群，2 の場合は子要素群を表すことにする．

$$w_{kj}^x = F_{kj}^x \tag{5}$$

ここで，F_{kj}^x はクラスタ C_k の親要素群，または，子要素群に出現する名詞 e_j の頻度を表す．なお，特徴ベクトルはその大きさが 1 になるように正規化する．

クラスタの要素群間の類似度は特徴ベクトルの内積により求める．クラスタ C_k の要素群 c_k^1 とクラスタ C_l の要素群 c_l^x の類似度を式 (6) で定義する．

$$Sim(c_k^1, c_l^x) = \vec{x}_k^1 \bullet \vec{x}_l^x \quad (k \neq l) \tag{6}$$

ディレクトリ構造の構築は，要素群間の類似度が最大の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により実現する．統合の手順としては，まず，要素群 c_k^x 自体からなる要素群集合 $A_k = \{c_k(X)\}$ を作成する．次に，要素群間の類似度 $Sim(c_k^1, c_l^x)$ が閾値 β 以上で値が最大となる要素群 c_m^x，c_n^x のうち，木構造の性質を満たすものから順に統合する．統合は要素群 c_m^x，c_n^x を含む要素群集合 A_m，A_n の和集合 $A_{m\cup n} = \{c_m^x, c_n^x\}$ を求めることにより行う．要素群間の類似度が閾値 β 以上であるものが無くなった時点で統合を終了する．

最後に，構築したディレクトリ構造に Web ページを分類する．これは，カテゴリを形成する要素群のアンカーテキストが本来リンクで参照していた Web ページをそのカテゴリに振り分けることにより実現する．

3.4 カテゴリ名の決定

各カテゴリはアンカーテキストの集合から成り立っているため，それらをもとにカテゴリの名前を決定する．カテゴリに含まれるアンカーテキスト集合の中にできるだけ共通して出現し，かつ，ある程度の長さをもった語句をそのカテゴリの名前とする．

まず，カテゴリ D_i に含まれるアンカーテキスト集合 $B_i = \{anchor_{i1}, \cdots, anchor_{iM}\}$ から形態素単位で N-gram($N \geq 1$) s_{ij} をすべて抽出し，これらをカテゴリ名の候補とする．そして，s_{ij} ごとに，アンカーテキスト集合 B_i の各要素に対する包含率 $Cover(s_{ij}, anchor_{ik})$ を式 (7) より求め，式 (8) で定義する平均包含率 $Cover_{ave}(s_{ij}, B_i)$ を計算する．カテゴリ名は平均包含率を最大とする s_{ij} とする．

$$Cover(s_{ij}, anchor_{ik}) = \begin{cases} \frac{F_{jk}^i}{F_k^i} & (s_{ij} \subseteq anchor_{ik}) \\ 0 & (otherwise) \end{cases} \tag{7}$$

$$Cover_{ave}(s_{ij}, B_i) = \frac{\sum_{k=1}^{M} Cover(s_{ij}, anchor_{ik})}{M} \tag{8}$$

なお，F_k^i はアンカーテキスト $anchor_{ik}$ に出現する形態素の頻度，F_{jk}^i は語句 s_{ij} とアンカーテキスト $anchor_{ik}$ に共通して出現する形態素の頻度，M はアンカーテキスト集合 B_i 中のアンカーテキストの数を表す．

表 1: サイトごとのデータ量

ID	サイト	ページ数	アンカーテキスト対数	
			全て	上位-下位
I	www.engg.nagoya-u.ac.jp	126	822	85
II	www.env.nagoya-u.ac.jp	265	11432	430
III	www.is.nagoya-u.ac.jp	106	1288	236
IV	www.sci.nagoya-u.ac.jp	264	9267	1680
V	www.soec.nagoya-u.ac.jp	578	15840	2393

4. 評価実験

4.1 実験方法

複数のサイトから階層ディレクトリの構築実験を行った．実験には，表 1 に示した名古屋大学の 5 つの研究科サイトを使用した．各サイトのトップページからリンクを辿ることにより，アンカーテキスト対を抽出し，ディレクトリ構造を構築する．なお，アンカーテキスト対のクラスタリングに使用する閾値 α を 0.5，ディレクトリ構造の構築に使用する閾値 β を 0.6 に設定し，クラスタリングの結果，アンカーテキスト対が 2 個未満のクラスタを対象外とした．なお，実験では，形態素解析器に「茶筌」[4] を使用している．

4.2 実験結果

各サイトごとに対象とした Web ページ数と得られたアンカーテキスト対の数を，サイト内のすべてのリンクと上位-下位関係を表すリンクに分けて表 1 に示す．

ディレクトリ構造の構築のときに要素群間の統合を少なくとも一度は行って生成されたディレクトリは 25 個であった．結果の例として，入試情報に関してまとめられた 2 つのディレクトリ構造を表 2，表 3 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分類して表しており，それぞれ表 1 の ID と対応する．カテゴリ名はアンカーテキスト集合をもとに自動生成したものである．表 2 の「平成 14・15・16・17 年度」には平成 14 年度から 17 年度の入学試験実施状況について記述されたページが分類されている．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見て取れる．

表 2: 生成されたディレクトリ構造の結果 (1)

階層	カテゴリ名	ページ数			
		I	II	IV	V
1	受験案内	1	0	1	0
1-1	過去の入試問題情報請求方法	0	0	1	0
1-2	平成 14・15・16・17 年度	0	0	1	0
1-3	入学案内	1	0	1	0
1-3-1	第 3 年次学士入学	2	0	0	0
1-3-2	第 3 年次編入学（高専）	1	0	0	0
1-4	博士課程（前期課程）	0	2	2	1

表 3: 生成されたディレクトリ構造の結果 (2)

階層	カテゴリ名	ページ数			
		I	II	IV	V
1	入試情報	1	1	1	1
1-1	博士課程	4	2	1	5
1-2	推薦入試	0	0	0	1
1-3	高度専門人コース	0	0	0	1
1-4	第 3 年次編入学について	0	0	0	1
1-5	募集要項の請求方法	0	1	0	1
1-6	経済学研究科過去問題の請求方法	0	0	0	1
1-7	平成 16 年度社会人一般コース第 2 次学生募集説明会	0	0	0	1

4.3 考察

実験結果を以下の項目に分けて考察する．

階層構造の妥当性

表 2 の「受験案内」と「入試情報」は同じ内容を表しており，階層としては不適切である．また，入学情報に関するディレクトリ構造は，表 2 や表 3 のようにいくつかのディレクトリ構造として構築されているが，これらは 1 つにまとめることが望ましい．

カテゴリ分類の正しさ

分類の正しさは，主観的に見てカテゴリ名に適したページが分類されているかにより判断する．表 2，表 3 のディレクトリ構造では，37 文書中 33 文書が正しく分類された．例えば，表 3 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を含んでおり，すべて正しく分類されている．一方，表 2 の「博士課程 (前期課程)」は後期課程の募集要項に関するページも含んでおり，これらのページは適切ではない．また,「入学案内」に専攻の案内パンフレットに関するページが分類されていたが，これはアンカーテキスト対をクラスタリングしたときに同時にクラスタ化されたことに起因する．

カテゴリ名の正しさ

カテゴリ名は，そのカテゴリを形成しているアンカーテキストをもとに決められる．例えば，表 2 の「受験案内」と「博士課程 (前期課程)」は，表 4 に示したアンカーテキスト集合から決定されている．しかし,「博士課程 (前期課程)」では “前期課程”の記述が “後期課程”よりも多いためにカテゴリ名として選択されてしまうなど，正しく表現しているとは言い難い部分もある．

不適切なディレクトリ構造

表 2，表 3 に示したディレクトリ構造の結果は，複数サイトのページをもとに構築されたものであるが，このうち，構造的に不適切なものもいくつか見受けられる．本手法では，ディレクトリの階層構造を構築するための要素群同士の統合は，アンカーテキスト集合の類似度から決定している．そのため，語句として類似していても内容が異なれば，不適切な上位-下位関係がディレクトリ構造に生じることがある．例えば，上位-下位関係の片方がアンカーテキストで “概要”と表現されていたために，異なる内容であったとしても，同一の語句であるために統合され，内容の異なるディレクトリ構造が構築される．

表 4: アンカーテキスト集合からのカテゴリ名の決定

受験案内
受験案内，受験案内，受験案内へ，受験案内，大学院受験希望者向け工学研究科案内
博士課程（前期課程）
博士課程 (前期課程) 募集要項，博士課程 (後期課程) 募集要項，入試情報 (大学院 博士前期課程 社会人一般コース)，博士課程 (前期課程)，博士課程 (後期課程)，名古屋大学大学院経済学研究科博士課程 (前期課程) 社会人一般コース募集要項のページ

また，本実験で構築した 25 個のディレクトリのうち 16 個が 1 つのサイト内のページで構成されたディレクトリ構造であった．これは，アンカーテキスト対をクラスタリングするときに，同じ上位-下位関係を表していてもアンカーテキストによる表現の違いにより，1 つにクラスタ化されないことが原因として考えられる．複数のサイトで多く見られる上位-下位関係を抽出し，それらを効果的にディレクトリ構造の構築に利用するためには，Web ページの内容をアンカーテキスト以外からも捉える必要がある．

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，複数サイトの情報を反映したディレクトリ構造が構築できたが，構造的に不適切なディレクトリ構造もいくつか見られ，十分なディレクトリ構造を構築するまでには至っていない．今後の課題として，適切な上下関係にある階層構造を構築するために，各ページの内容をアンカーテキスト以外からも抽出することが考えられる．

参考文献

[1] Daniele Riboni: Feature Selection for Web Page Classification, In *Proceedings of the Workshop of EURASIA-ICT 2002* (2002).

[2] 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).

[3] 長尾 真: 岩波講座ソフトウェア科学 15 自然言語処理. 岩波書店 (1996).

[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).