meta タグを利用した Web ディレクトリの自動構築手法

佐々木 稔　新納 浩幸
茨城大学工学部情報工学科

1　はじめに

ディレクトリ型の検索サービスはあらかじめ Web ページが項目別にまとめられているので，初心者でも簡単に WWW(World Wide Web) 検索をすることができる．このようなサービスを運営する側は Web ディレクトリへのサイト登録や分類，管理といった作業を人手により行っているため，膨大な Web ページを処理することが困難となる．また，充実した Web ディレクトリや個人によるリンク集を構築することも難しい．このため，Web ページの形式や内容からある視点を定め，Web ページを自動分類する研究が盛んに行われている．このような研究には，複数の Web ページにおいて類似した内容をまとめるもの [2]，リンクの参照共起などを分析してつながりの強い文書をまとめるもの [4]，また，Web ディレクトリにおいてリンクの紹介文を利用して分類器を学習するもの [5] が存在する．

この Web ディレクトリに登録された情報は，多くの場合が企業や個人のホームページであり，Web サイト単位で検索をすることができる．しかし，登録された情報が多くなるほど，リンク切れやカテゴリ分けなどの管理が難しくなる．そのため，Web ディレクトリはロボット検索と同様に網羅的な Web の目録を作ることを目的としているが，現状では厳しい登録審査を通った厳選サイトが登録されている．

そこで，我々はこれまで人手で行っている Web ディレクトリの管理作業を自動で行う研究をしている．オープンディレクトリなどを利用してカテゴリは既に存在するものと考え，まずはディレクトリへのサイト登録を自動的に行うことが課題となる．この課題に対して解決すべき 2 つの問題点を以下に示す．

1. どのようなページを代表的なサイトとしてディレクトリに登録するか決定すること
2. 選ばれたサイトを自動的に適切な Web ディレクトリに分類すること

これまでに，我々はこの課題に対して企業サイトに記載された情報から業種判別を行った [6]．このとき，企業サイトに記載された内容からキーワードを抽出した結果に一般的な単語が数多く含んでいることが原因で，自動分類の精度を大きく下げていることが分かった．分野を特定しやすいキーワード抽出を行うことは，情報検索や文書分類において非常に難しい問題で，この問題を解決するキーワード抽出手法の開発を行ってきた．その結果，サイトの内容語を扱わず，ホームページに記述された meta タグの name 属性値である keyword と description をキーワードとして利用することで，分類精度が向上するのではないかと考えた．本稿では，企業分類よりも大規模な実験を行うために，一般のサイトを対象として，それを Google ディレクトリに採用されている Dmoz 内のカテゴリに自動分類する手法を提案する．

2　Web ディレクトリ

Web ディレクトリは，Web サイトへのリンクをカテゴリ別に分類した階層的なリストである．このような Web ディレクトリの代表的な例として，Yahoo!，Google Directory，goo などの検索エンジンを提供するサイトや Open Directory プロジェクトが世界中のボランティアの協力のもとで作成しているディレクトリ，さらには地域の観光，飲食店情報等を網羅した個人運営によるポータルサイトなどが存在する．ユーザにとってこのような Web ディレクトリが威力を発揮するのは，検索したい分野が

あらかじめ分かっているときに，素早く欲しい情報を見つけられることにある．

Web ディレクトリに登録されている内容は，そのカテゴリに属する Web サイトのタイトル， URL，概要がひとつの組となって，一覧表示される．そこに登録された URL は，多くの場合が企業や個人のホームページであり，1 ページ単位での登録をしているサイトは少ない．ホームページだけであっても，Web ディレクトリの登録や削除などの管理は現在でも人手で行われており，作業の手間がかかってしまう．そのため，できるだけ少ない作業に抑えるために，サイト単位でのディレクトリ設計をしていると考えられる．

現在，Web ディレクトリは上記のような人手による管理の困難さと Web ページの爆発的な増加により作業が追いつかない状態が続いている．また，Google のように，WWW を網羅した精度の高いページ単位での検索が可能となり，現在の検索の主流となっている．そのため，Web ディレクトリはロボット型検索エンジンの検索結果を補完する役割になっている．

3 meta タグを記述する意義

meta タグは，HTML (Hyper Text Markup Language) 形式で書かれた文書の head タグで囲まれた領域において記述されるもので，HTML 文書で利用される文字コードやキーワード，概要などを埋め込むことができる．meta タグの name 属性による keyword の指定は，以下のように記述を行う．

<meta name=”keyword” content=”keyword_1, keyword_2, ··· , keyword_n”>

meta タグの content 属性値には，具体的なキーワード keyword_1, keyword_2, ··· , keyword_n が記述される．また，meta タグの name 属性による description の指定は以下のようになる．

<meta name=”description” content=”掲載内容の概要”>

この場合，content 属性値には，掲載内容の概要を記述する．これらの記述によるキーワードや概要説明は，検索エンジンがクローラーなどを利用して情報収集を行う際に重要なキーワードとして処理される他，検索結果の表示をする際のサマリーに引用されている．

情報を公開するユーザの立場から見て，これらのタグを作成する意味は，検索エンジンに検索されやすく，また，検索結果で表示されるサマリーを記述することである．そのため，ユーザは作成したページを端的に表現できるキーワードを選定し，ページの概要を作成している．

しかし，これらのタグは検索エンジンにとって非常に有効なものであるとは限らず，検索エンジンのシステムに偽った情報を登録させるなどの欠点も存在する．例えば，keyword や description の属性値にページの内容とは異なるキーワードや説明を記述し，様々なキーワードからそのページが上位に検索されるようにインデックスや検索順位の操作を行う．また，キーワードとは微妙に異なる文字列を加えることにより，キーワードの綴りを間違えた場合でも検索されるように操作する方法もある．

4 Web ディレクトリへの分類実験

本節では，Web ディレクトリに URL を自動的に登録するためのディレクトリ判別実験について述べる．Web ディレクトリについては，企業の大規模な Web ディレクトリ構造や個人のリンク集など，さまざまな種類の分類データが存在するが，Open Directory Project においてフリーで利用可能な Dmoz [1] を利用する．この Dmoz の Web ディレクトリには数多くの言語で記述されたカテゴリが登録され，日本のカテゴリ (“Top/World/Japanese”) 以下を考慮するだけでも 17,519 種類のカテゴリが存在し，それらが階層的につながっている．

この Dmoz をそのまま利用すると問題点が存在する．例えば，ディレクトリの中に “地域” というカテゴリが存在する．その “地域” カテゴリ以下には都道府県名，具体的な内容での分類へと続いている．meta タグの中には具体的な地域名が記載されていることが多く，内容を表すカテゴリに分類されずに，地域以下のカテゴリに分類されやすくなる．この問題は “地域” と文書内容の各々のカテゴリに

[1] http://dmoz.org/

重複して分類することで対応が可能であると考えられるが，この問題は今後の課題とする．

以上のような理由により，本稿で行う実験は，日本のカテゴリに所属するスポーツのカテゴリ (“Top/World/Japanese/スポーツ”) 以下を利用し．より簡単なディレクトリを対象として meta タグを利用した分類性能を調査することとする．スポーツ以下のディレクトリには 753 種類 (登録なしのカテゴリを除く) のカテゴリが存在し，約 8,000 件の URL が登録されている．

4.1 ディレクトリ情報の学習方法

Web ディレクトリに登録されたデータから，各ディレクトリに含まれるキーワードの出現確率を計算する．このとき，計算対象となるキーワードは Web ディレクトリに登録されたリンクのタイトルと紹介文から抽出した．ひとつのディレクトリに登録されたリンクのタイトルと紹介文を抽出し，それに対して 茶筌[2] を利用して形態素解析を行い，名詞 (数，非自立は除く)，片仮名語，未知語，アルファベットをキーワードとして取り出す．これらのキーワードに対して，すべてのディレクトリにおけるキーワードの頻度統計を用いて出現確率を計算する．この統計値とディレクトリ名をラベルとして教師あり学習であるナイーブベイズ手法 [3] を利用し，Web ディレクトリの分類モデルとする．

4.2 URL 自動分類の方法

以上の方法で得られた Web ディレクトリの分類モデルを利用して，分類したい Web サイトの URL を入力することで，分類すべきディレクトリを自動的に判定する方法について説明する．分類したい URL を入力すると，それに対応したファイルをダウンロードする．ダウンロードした HTML ファイルから meta タグの name 属性が keyword(または keywords) と description の値を持つ場合の content 属性値を抽出する．抽出した文字列に対して茶筌を利用して形態素解析を行い，ディレクトリ情報の学習時と同じ品詞属性をもつ単語をキーワードとして用いる．得られたキーワードに対して分類モデルにより，ディレクトリを判定する．

5 実験

Dmoz のスポーツ以下に登録されたデータから学習した分類モデルを利用して，そこに含まれていない新しい URL から分類すべきディレクトリの自動判別を行った．テストデータには，Yahoo! カテゴリ[3] のスポーツに登録された URL で， Dmoz に登録されていない 1,044 件の Web ページを利用した．

分類する URL から HTML 文書をダウンロードし，その中に必要とする meta タグの keyword と description の name 属性値が含まれていれば，その内容を抽出する．抽出したデータに対して，分類モデルの作成時と同様に形態素解析を行い，片仮名文字列，未知語，アルファベットや名詞 (数，非自立は除く) をキーワードとして利用する．このとき，キーワードの頻度に重みを加えるために，出現頻度を 2 乗した値をキーワードの重要度とする．このキーワード統計に対してナイーブベイズ手法を利用して，ディレクトリの判定を行う．

meta タグの有効性を実証するため，比較実験として HTML 文書の meta タグを使わず，本文からキーワードを抽出した場合，また，meta タグから取り出したキーワードと本文をともに利用した場合について，分類性能の比較を行う．この 2 つのベースライン精度を求めるとき，meta タグのキーワードと本文のキーワードの重みには出現頻度をそのまま利用することとした．

5.1 実験結果・考察

本実験を行った結果を表 1 に示す．表 1 は，テストデータの適切なカテゴリ名とそのカテゴリに属する文書数，3 種類データに対して分類モデルが正しく分類した数をそれぞれ表している．

まず，本文のみを利用した場合と meta タグを利用した場合の分類結果を比較すると，陸上競技とバドミントンで本文を利用した場合の分類精度が上回っているが，その他では meta タグのみを利用し

[2] http://chasen.naist.jp/hiki/ChaSen/

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

表 1: meta タグと本文を利用したときの実験結果

カテゴリ名	文書数	分類結果		
		タグのみ	タグ+本文	本文のみ
サッカー	179	165	162	147
モータースポーツ	168	119	116	58
野球	124	113	107	93
ゴルフ	66	59	53	37
テニス	65	55	51	34
自転車	58	48	41	20
サーフィン	52	40	40	15
格闘技	46	41	35	27
バスケットボール	37	27	27	15
ラグビー	37	29	25	18
ホッケー	25	17	16	8
乗馬	23	15	12	4
陸上競技	22	15	17	16
弓道	21	20	18	17
格闘技	21	17	18	14
ソフトボール	19	18	17	13
バドミントン	12	4	5	5
武道・武術	11	11	11	11
アーチェリー	10	6	4	0
柔道	10	9	9	8
剣道	10	10	10	9
ハンドボール	9	6	6	5
ビリヤード	8	6	6	2
カーリング	7	5	4	1
ペタンク	4	3	3	2
合計	1044	858	813	579
分類精度		0.82	0.78	0.55

た場合の方が分類精度が向上していることが分かる．本来，meta タグに記述するキーワードや概要は，そのページの作者が検索エンジンのキーワードとして使われやすい，分野特有のキーワードやフレーズを記述していることが多い．そのため，本文を利用するよりも分野特有のキーワードが含まれている可能性が高い．その結果，meta タグを利用した場合の分類性能が大幅に高くなったと考えられる．

次に，タグと本文の両方を利用した場合の分類精度と比較すると，本文のみを利用した結果からは精度が向上しているが，タグのみを利用した場合と比較すると少々分類性能が低かった．これは，分野に重要なキーワードの他に一般的な単語が含まれているために，このような一般的な単語の出現確率が分類に影響を及ぼしているのではないかと考えられる．

6 おわりに

本稿では，Web ページの本文を扱わず，HTML ファイルに記述された meta タグの name 属性値である keyword と description における記述をキーワードとして利用した，Web ディレクトリの自動構築実験を行った．その結果，HTML ファイルの本文のみを利用した場合と比較して，大きく分類性能が向上することが分かった．ユーザの意図するキーワードを抽出できるこのタグを利用することで，検索エンジンだけではなく文書分類にも有効な素性となることが分かった．

今後は，meta タグに記述された単語の分析を行って分類精度を高めるとともに，階層的な分類を行い，より実用的な Web ディレクトリの構築を行うことが課題である．

参考文献

[1] Edie Rasumssen. *Clustering algorithms*, pages 419–442. W. B. Frakes and R. Baeza-Yates, editors, Information Retrieval: Data Structures and Algorithms, Prentice Hall, London, 1992.

[2] 石田 栄美, 久野 高志, 安形 輝, 野末 道子, 上田 修一. 内容的なまとまりをもつＷｅｂページ群の自動判定. 1999 年度三田図書館・情報学会研究大会発表論文集, 三田図書館・情報学会, 1999.

[3] 北 研二. **確率的言語モデル**. 東京大学出版会, 1999.

[4] 原田 昌紀, 風間 一洋, 佐藤 進也. 参照共起分析の web ディレクトリへの適用. 情報学基礎研究会, 情報処理学会, 2001.

[5] 谷津 哲平, 新納 浩幸, 佐々木 稔. Web ディレクトリを用いた検索ナビゲーション. **言語処理学会第** *11* **回年次大会論文集**, pages 1022–1025, 2005.

[6] 佐々木 稔, 新納 浩幸. 文書分類手法を用いた企業 web サイトからの業種分類. **言語処理学会第** *12* **回年次大会論文集**, pages 352–355, 2006.