文書分類手法を用いた企業 Web サイトからの業種分類

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 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にとって，階層構造を持つ Web ディレクトリへのサイト登録や分類，管理といった作業は人手により行われている．しかし，人手による分類では膨大な 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．このような従来研究には，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，我々の研究ではこれまで人手で登録した Web ディレクトリへのサイト登録や管理作業を自動で行うことを目標としている．オープンディレクトリなどといった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. 人手で登録する場合と同じようにリンクの質を確保しなければいけない．そのために，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

本稿では，この課題をより簡単に実験するため，東京証券取引所に登録されている企業名とその業種のペアをもとに企業名からの業種判別を行う．この実験は，企業名から企業サイトを推定し，企業の Web サイトをひとつの文書として捉え，そこに含まれるキーワードを特徴として業種の学習を行う．この学習結果から，文書分類の手法を用いて新たな企業を業種に分類するタスクを行う．

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，検索したい内容が限られた範囲の分野に絞られていることを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．

このような 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，カテゴリの内容を端的に記述したページ単位での登録をしているサイトは少ない．ページ単位での Web ディレクトリの登録やリンク切れ防止，カテゴリ情報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，World Wide Web を網羅した精度の高いページ単位での検索が可能となっている．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 業種判別実験

本節では，東京証券取引所に登録されている企業名とその業種のペアをもとに企業名から業種判別を行う実験について述べる．まず，企業名の抽出には Yahoo! ファイナンス [1] からたどることのできる 32 種類の業種別の企業情報を利用する．そこに登録されている企業名と登録された業種とのペアを抽出した結果，3011 件のデータが得られた．このデータをもとに，企業名からその企業のトップページを推測し，その企業の「顔」となる Web ページのアドレスを見つける．次に，推測した URL のアドレスで公開された Web ページとそこからリンクされた同じドメインのページをダウンロードする．ダウンロードしたページからキーワードとなる単語を抽出し，キーワードを特徴として業種カテゴリを判定するための分類器を構築する．

3.1 トップページの URL 推定

企業名からその企業のトップページアドレスを推測する．トップページの URL を推測するには，Yahoo! の検索エンジンを利用する．検索エンジンのキーワードには，先に企業情報から得られた企業名に加えて，企業名と同時に出現しやすい “株式会社” というキーワードを加えた 2 単語で AND 検索をした．その検索の結果，トップにランクされる URL をトップページ URL と推測した．

3.2 企業サイトからの特徴単語抽出

企業サイトの代表的な URL を推測し，それぞれの企業名に対するホームページが分かれば，次に企業サイトから特徴となる単語の抽出を行う．このとき，企業サイトのホームページだけをダウンロードし，特徴単語を利用する場合，特徴となる単語数が非常に少ないため，学習データとしての利用が難しくなる．そのため，企業サイトからの単語抽出には，ホームページからリンクされた企業サイト内の下位ディレクトリにあるファイルもダウンロードし，そこに含まれた特徴単語も利用することとした．

ファイルのダウンロードには，wget を利用した．wget は URL を指定することにより，そこからたどることのできるファイルを収集することができる Web ページの自動収集ロボットである．この wget を用いて，推定した URL からスタートしてそのサイトに存在するファイルをダウンロードした．このとき，ダウンロードするリンクの深さを深くしすぎると，ダウンロードに時間がかかるため，リンクの深さを 5 と設定した．しかし，それでも特徴単語ファイルのサイズが 1 キロバイトに満たない場合は深さを 10 としてダウンロードした．

得られたファイルの集合に対して，HTML タグを取り除き，テキストファイルだけを取り出す．さらに，取り出したテキストに対して形態素解析を行い，名詞，動詞と形容詞となる単語を最終的に特徴単語として取り出した．ここで，形態素解析器には茶筌を利用した．

このようにして得られた特徴単語には豊富な種類の単語が含まれているので，この単語集合をひとつの文書とみなすと，文書と業種名のペアと考えることができる．このペアの集合を文書分類の手法を利用して業種判別を行うための分類器を構築する．

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

3.3 データ取得が不可能なサイト

上述のように企業サイトから特徴単語を抽出する際，ファイルがダウンロードできずに特徴単語が取得できない企業がいくつかあった．その原因には以下に示す 4 点に大別できる．

- wget などの Web ロボットでファイルを取られないように，robot.txt や meta タグでアクセス制御をしている．
- ホームページにテキストで書かれたリンクやテキストがなく，Flush Player でのみ情報提供している．
- リンクがすべて Javascript の中で記述されている．
- リンクがすべて Java で制御されている．

このうち，Web ロボットのアクセス制御をしているサイトについては，そのサイトの意向に従ってデータの収集を取り止めた．リンクが Flush や Java, Javascript 内で記述されている場合には，wget でリンクの解析が行われないため，現状ではデータを取得することができない．しかし，このようなサイトは 30 件ほどと非常に少ないこともあり，手動でリンク先をたどってデータが取得できる URL に修正し，そのページをトップページとみなした．ただ，それでも特徴単語ファイルのサイズが 1 キロバイトに満たない場合は，そのサイトについてもデータ収集を取り止めた．このようにして，3011 件の企業からデータ取得できないサイトを取り除くことで，2947 件のデータを取得することができた．

3.4 業種の学習方法

得られた特徴単語の集合と業種ラベルのペアから，業種ラベルがどのような特徴単語により構成されているのかを学習させる．機械学習には過去のデータを教師として教師データを満足する特徴や境界面を求める教師あり学習と与えられたデータ集合から同じような特徴分布を持つデータをまとめる教師なし学習 [1] がある．ここでは，業種ラベルという教師データが存在するため，教師あり学習手法であるナイーブベイズ手法 [3] を利用して学習を行う．

まず，得られたデータから機械学習を行うための学習データとその判別モデルを評価するためのテストデータへの分割を行う．今回の実験では，全 2947 件のデータのうちすべての業種ラベルにおいて含まれているデータの約 90% を学習データとして抽出した．その結果，2654 件のデータを学習データとし，残りの 10% にあたる 294 件のデータをテストデータとした．

企業の学習データから，ナイーブベイズ手法を用いて各業種がどのような特徴単語から成り立っているかを学習する．各業種ラベルに属する企業の特徴単語を集計し，業種における各単語の出現確率を計算する．また，全体のラベルから各業種ラベルが出現する確率も求める．これらの確率の積を保存することで業種判別を行うための分類器を構築することができる．

分類器が得られると，テストデータから業種を判別し，評価を行う．1 件の評価データに含まれているすべての特徴単語が，それぞれの業種ラベルについて出現する確率を計算する．その中で，最も高い出現確率となる業種ラベルを判定結果として出力する．

3.5 実験結果・考察

本実験により，テストデータとして用意した 294 件のデータについて評価を行った結果を表 1 に示す．テストデータ 294 件のうち正しい判別ができたものが 123 件で，約 41.8 % という結果となった．この中で，建設業，食料品，化学，銀行業は学習データが 100 件以上存在し，豊富なデータ中に特定分野に出現しやすい専門用語や商品名が数多く存在しているために正解数が多くなったと考えられる．また，学習データ数が少ないが精度の高い，電気・ガス業や証券業などの業種も存在している．学習データ中に文書量が少ない場合でも，その分野に特定する単語が多く含まれていると考えられる．

表 1 において正解率が高い場合でも，その業種がこのモデルを使って判定されている数を考慮する必要がある．このモデルを使った判定先の分布と正解数の表を表 2 に示す．表 2 において最も判定先の数が多い 46 件の学習データがサービス業として判定されている．そのため，正解率は高いにもかかわ

表 1: 業種別の判定結果

業種	正解	不正解	合計
建設業	17	6	23
食料品	14	2	16
化学	13	9	22
機械	12	12	24
電気機器	10	19	29
サービス業	9	3	12
銀行業	7	3	10
鉄鋼	4	1	5
繊維製品	4	3	7
陸運業	4	3	7
輸送用機器	4	6	10
金属製品	4	7	11
電気・ガス業	3	0	3
医薬品	3	2	5
不動産業	3	7	10
証券業	2	1	3
倉庫・運輸関連業	2	2	4
その他金融業	2	4	6
その他製品	2	10	12
鉱業	1	0	1
海運業	1	1	2
ガラス・土石製品	1	6	7
小売業	1	30	31
水産・農林業	0	1	1
石油・石炭製品	0	1	1
保険業	0	1	1
ゴム製品	0	2	2
パルプ・紙	0	2	2
非鉄製品	0	4	4
精密機器	0	5	5
情報・通信	0	7	7
卸売業	0	11	11
合計	123	171	294

表 2: ナイーブベイズ手法による判定先の集計

業種	正解	判定数
サービス業	9	46
化学	13	29
機械	12	28
建設業	17	27
食料品	14	22
電気機器	10	16
繊維製品	4	13
鉱業	1	11
輸送用機器	4	9
倉庫・運輸関連業	2	9
鉄鋼	4	9
金属製品	4	9
その他金融業	2	8
その他製品	2	8
銀行業	7	7
ゴム製品	0	6
医薬品	3	5
不動産業	3	5
陸運業	4	5
ガラス・土石製品	1	4
石油・石炭製品	0	4
証券業	2	3
精密機器	0	3
電気・ガス業	3	3
卸売業	0	2
小売業	1	2
海運業	1	1
情報・通信	0	0
非鉄製品	0	0
パルプ・紙	0	0
保険業	0	0
水産・農林業	0	0
合計	123	294

らず，判定モデルとしては多くの誤判定を引き起こしていることが分かる．これは，サービス業の業務内容が幅広いために他の業種の内容を含んでいる場合がある．テストデータの中で，情報・通信，小売業が 7 件，卸売業，不動産業が 4 件など様々な業種がサービス業として判定されている．サービス業のような業種の幅が広い，その他金融業，その他製品，小売業や卸売業という業種を扱うためには，異なる業種ラベルを考えるなどといった別の手法を考慮すべきではないかと思われる．

次に，判定される数が少ないために正解数も少なくなってしまう業種も存在する．情報・通信，非鉄製品などはテストデータにおいて全く判定されていなかった．その中でも情報・通信は学習データ量が多いにもかかわらず，テストデータではすべてサービス業として判定された．そのため，情報・通信とサービス業の類似性や相異点を分析する必要があると考えられる．また，非鉄製品のようにすべて異なる業種として判定されているものも存在する．非鉄製品も内容に幅が広いためにもう少し細かい分類などが必要ではないかと考えられる．

4 おわりに

本稿では，企業名とその業種のペアをもとに企業名からの業種判別を行った．実験の結果，テストデータ 294 件のうち正しい判別ができたものが 123 件で，約 41.8 % という結果となった．

今後は，誤判定しやすい業種について特徴単語の分析と業種ラベルの改良などを行い，分類精度を向上したいと考えている．

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.