企業ウェブページからの業種情報の抽出と分類

北陸先端科学技術大学院大学
先端科学技術研究科

安道 健一郎

平成 30 年 3 月

修 士 論 文

企業ウェブページからの業種情報の抽出と分類

1610224 安道 健一郎

主指導教員 白井 清昭

審査委員主査 白井 清昭
審査委員 池田 心
東条 敏
長谷川 忍

北陸先端科学技術大学院大学

先端科学技術研究科 [情報科学]

平成 30 年 2 月

概 要

近年，ブログ，SNS，キュレーションサイトなどの普及により個人が気軽に情報を発信できるようになったため，ウェブ上の情報は爆発的に増えている．そのため，ウェブ上には有用な情報が数多く存在するが，誤った情報も多数存在する．したがって，ユーザは情報を検索する際に，ウェブ上の情報が正しいかどうかを判断する必要がある．情報の信頼性を確かめる一つの手段として，その情報が誰によって発信されたかを確認する方法がある．専門的な情報は専門家が発信したもののほうが一般人が発信したものより信頼できるだろうという考えに基づき，発信者情報を参考にすることで信頼性を判断する．例えば，法律のことを調べる際には，法律事務所のウェブサイトに書かれている情報や法律家が書いている記事に記載されている情報などは信頼性が高いといえる．このようなウェブサイトの信頼性を判断するための情報は，ウェブの情報量の増加に伴い，今後さらに重要性が増してくると思われる．

本研究では，検索エンジンでヒットすることの多い企業のウェブページに注目し，ウェブページから業種情報を自動抽出し，また抽出した業種情報を基に企業のウェブサイトを業種によって自動分類することを目的とする．業種情報とは，企業が展開する事業を書き表わした情報と定義する．業種情報は企業のプロフィールに相当する情報といえるため，本研究ではこれを企業ウェブサイトの作成者情報として扱う．作成者情報を検索エンジンにおける検索結果とともに提示することで，ユーザが信頼性の高い情報を選別する作業をサポートすることを狙う．この際，作成者情報 (業種情報) は一般に長いテキストであるため，作成者情報そのものではなく，あらかじめ定義した業種のカテゴリを提示することで，ユーザの視認性を高める．

本研究は，業種のカテゴリを自動分類する先行研究と比べて，ウェブサイトから業種情報を抽出し，そこに出現する単語に高い重みを与えて分類器を学習する点に特徴がある．また，ウェブサイトの作成者名を抽出する研究や，ブログから著者のプロフィールの抽出を試みる研究はあるが，一般のウェブサイトから作成者 (企業) のプロフィールに相当する情報 (業種情報) を抽出する点に本研究の新規性がある．

本研究の提案手法は以下の通りである．まず，業種情報を自動抽出するために，企業サイトの HTML ソースを Document Object Model(DOM) で解析し，HTML タグの階層構造を表わす DOM ツリーを得る．次に，DOM ツリーから，Description，Keywords，業種説明，事業説明を含んだ DOM ノードをテキスト中の単語，リンク先の URL を条件としたいくつかのルールによって抽出する．Description，Keywords とは head タグ内の name 属性がそれぞれ Description，Keywords となっている meta タグのテキストである．業種説明とは，企業の概要がまとめられているページに存在する，表形式で記述されたその企業の業種に関する情報である．事業説明とは，独立したページにまとめて記述されている企業の事業内容を説明したテキストである．

Description と Keywords は HTML のタグにより機械的に抽出出来る．業種説明を抽出する際，表形式で表されていることを想定しているため，まず表において業種説明の見出しに当たる DOM ノードを検出し，次にその近傍にあるノードを業種説明として抽出する．一方，事業説明を抽出する時は，まずそれが書いてある独立したページへのリンクを企業のトップページから検出する．次に，広告や目次など事業説明以外のテキストを除外するため，検出したウェブページのメインコンテンツを同定し，そのテキストを事業説明として抽出する．メインコンテンツの検出アルゴリズムは加藤らの手法を用いた．

次に，機械学習を用いて企業のウェブページを業種カテゴリ分類する方法について述べる．これまでに抽出した業種情報を含む DOM ノード内のテキストから自立語を抽出し，それを機械学習の素性とする．これと合わせて，企業のウェブサイトのトップページ中の自立語も素性として用いる．ただし，素性（自立語）が業種情報に出現するときには素性ベクトルにおいて高い重みを与える．業種カテゴリの分類モデルはナイーブベイズ（NB）とランダムフォレスト（RF）で学習する．業種カテゴリは，ウェブディレトリーサービスの一つである Open Directory Project(ODP) の日本語サイトで定義されているウェブサイトのカテゴリを参考に，28 個の業種カテゴリを設定した．

提案手法の有効性を評価する実験を行った．まず，業種情報抽出の精度，再現率，F 値を求めた．実験データとして ODP から取得した 100 件に対し，提案手法で業種情報を抽出し，人手でタグ付けした正解の業種情報と比較して，その精度，再現率，F 値を求めた．結果は，Keywords，Description の抽出については，精度が 1，再現率が 1，F 値が 1 となった．業種説明の抽出については，精度が 1，再現率が 0.95，F 値が 0.97 となった．事業説明の抽出については，精度が 1，再現率が 0.91，F 値が 0.95 であった．いずれの業種情報も十分正確に抽出できることが確認された．また，100 件のウェブサイトの中にどれくらい抽出対象の業種情報が含まれているかについても調査した．その割合は，Keywords が 0.8，Description が 0.85，業種説明が 0.7，事業説明が 0.36 であった．事業説明を含む企業のウェブサイトはそれほど多くないが，それ以外の業種情報は多くの企業ウェブページに存在することがわかった．

次に，ODP から 29364 件の企業ウェブサイトを取得し，これを訓練データとテストデータに 9:1 の割合で分割し，業種カテゴリの自動分類の正解率を算出した．ベースラインはトップページからのみ素性を抽出する手法とした．また，自動分類の正解率の上限を調べるため，ODP からランダムで取得した 300 件のウェブサイトを対象に人手でウェブサイトをカテゴリ分類したときの正解率も調べた．ベースラインの正解率は NB が 0.252，RF が 0.493 だったのに対して，提案手法の正解率は NB が 0.270，RF が 0.508 だった．また，人手による分類の正解率は 0.717 だった．提案手法はベースラインをわずかに上回った．機械学習アルゴリズムの比較では，RF は NB を大きく上回った．上記の結果は提案手法の有効性を示してはいるが，人による判定との差は大きく，改善の余地が大きい．人が業種カテゴリを判定する際には，カテゴリをすぐに決定できる特定の単語や特徴 (URL 内の「.ac」,「会計」,「税理」,「商工会」など) を見つけて判定することが多かった．このような特徴的な単語を自動的に特定できれば業種判定の正解率が向上すると考えられる．

今後の課題として，ナイーブベイズやランダムフォレストのパラメータを開発データを用いて最適化し，業種カテゴリの自動分類の精度を向上させることが挙げられる．また，実際に抽出した業種情報をユーザに提示できるようなシステムを開発することも重要な課題である．

目 次

第1章　はじめに

1.1　研究の背景

近年，ブログ，SNS，キュレーションサイトなどの流行で個人が気軽に情報を発信できるようになったため，ウェブ上の情報は爆発的に増え，様々な情報が存在している．多くの人は検索エンジンを用いて自分が知りたい情報を取得しようとするが，膨大な情報を含むウェブを情報源とする検索においては，知りたい情報を取得することが困難な場合も多い．一方，ウェブ上には誤った情報も多数存在する. 先に述べたように個人のユーザが気軽に情報発信を行えるようになったため，信頼性の低い情報はウェブ上に多く流通している．したがって, ユーザは情報を検索する際, ウェブ上の情報が正しいかどうかを判断する必要がある. 情報の信頼性を判定する方法はいくつか考えられる．例えば，該当ウェブページに書かれている内容が他の複数のウェブページでも書かれていることを確認することによって，その情報が正しいと判定することが考えられる．しかし，この方法は事実確認のために多くのウェブページを確認しなければならないという問題点がある．また，デマなどのように誤った情報が流布しているときは，複数のウェブページに同じ情報が載っていてもその情報が必ずしも正しいとは限らないという問題点もある．もう一つの手段として，その情報が誰によって発信されたかを確認する方法がある．専門的な情報は専門家が発信したもののほうが一般の人が発信したものより信頼できるだろうという考えを基に，発信元の情報を参考にすることで信頼性を判断する．例えば，法律のことを調べる際には，法律事務所のウェブサイトに書かれている情報や法律家が書いている記事に記載されている情報などは信頼性が高いといえる．このようなウェブサイトの信頼性を判断するための情報は, ウェブの情報量の増加に伴い, 今後さらに重要性が増してくると思われる.

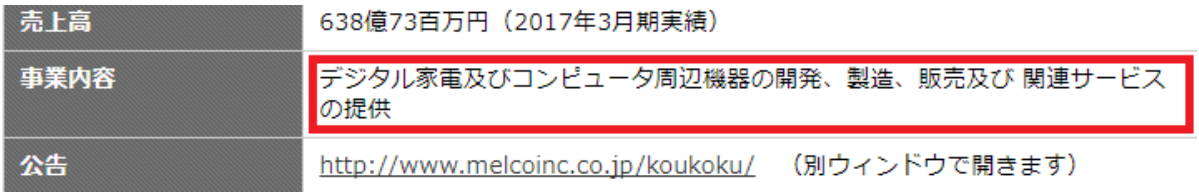
1.2　研究の目的

本研究では, 検索エンジンでヒットすることの多い企業のウェブページに注目し, ウェブページから業種情報を自動抽出し, また抽出した業種情報を基に企業のウェブサイトを業種によって自動分類することを目的とする. 業種情報とは, 企業が展開する事業を書き表した情報と定義する. 業種情報は企業のプロフィールに相当する情報といえるため, 本研究ではこれを企業ウェブサイトの作成者情報として扱う. 作成者情報を検索エンジンにおける検索結果とともに提示することで，ユーザが信頼性の高い情報を選別する作業をサポートすることを狙う．この際，作成者情報 (業種情報) は一般に長いテキストであるため，作

成者情報そのものではなく，あらかじめ定義した業種のカテゴリを提示することで，ユーザの視認性を高める．例として，図 1.1，図 1.2，図 1.3 に株式会社バッファローの業種情報を示す．一般に，業種情報はウェブページ上で様々な様式で記述されている．図 1.2 は，トップページの HTML のヘッダの中の Description と Keywords に会社の業種に関係する記述してある．図 1.2 は，企業に関する様々な情報が表形式で書かれているページの中で，その企業の業種情報が書かれている．図 1.3 は，独立したウェブページに会社の事業が詳細に書かれている．これらの情報を参照すると，この企業がコンピュータ関連の電子機器の製造及び販売をしている会社だと判断できる．既に述べたように，本研究では業種情報を基に企業ウェブページの業種カテゴリに分類し，その業種カテゴリを検索エンジンの結果とともに提示する．図 1.4 は本研究が想定する検索エンジンの出力である．「小売」「電機・エレクトロニクス」は企業の業種カテゴリである．SSD の技術的なことを調べる際には，小売よりも電機・エレクトロニクスの企業のウェブサイトを調べた方がよさそうといえる．

(URL http://buffalo.jp/ ; 取得日 2017/02/02)

図 1.1: 業種情報の例 1



(URL http://buffalo.jp/company/outline/profile.html ; 取得日 2017/02/02)

図 1.2: 業種情報の例 2

1.3 本論文の構成

本論文の構成は以下の通りである．第 2 章では，先行研究を概観し，また先行研究と本研究との違いを論じる．第 3 章では，企業のウェブページから企業情報を抽出し，また抽

会社の事業分野

インターネットを快適に活用するための、パソコン及びブロードバンド関連機器の開発・製造・販売及び関連サービスの提供を行っています。
パソコンに追加増設するメモリー、ストレージ、ネットワークといった各種周辺機器はパソコンの快適環境を構築します。当社では常にお客様の利便性向上を目指した製品開発に取り組み、使い勝手の良い製品を発売し、お客様にソリューションを提案しています。また、無線LANの設置などを行うインターネット訪問設定・設置サービス事業も行っています。
さらに、パソコンとデジタル家電を融合するホームネットワーク事業に注力しています。

メモリー製品

メモリーモジュールはパソコンの記憶容量を増加させる部品です。容量が大きいほど一度に扱えるプログラムやデータの量が大きくなり、効率良く速く処理することができます。メモリー製品には、メモリーモジュールの他に、手軽にデータが持ち運べる「USBメモリー」、携帯電話やデジタルカメラで利用する「microSD」「コンパクトフラッシュ」などがあります。

‣メモリー

‣USBメモリー

‣メモリーカード

ストレージ製品

ストレージはソフトウェアやデータを保存する外部記憶装置です。主力の製品であるハードディスクにはパソコンに内蔵するタイプやパソコンの横に置いて使用する外付けタイプ、また、複数のパソコンから利用できるネットワークハードディスクがあります。パソコンで動画を利用する機会が増えていることから、必要な記憶容量も増加傾向にあるため、ハードディスクの増設需要が高まっています。この他ブルーレイやDVDなど様々な記憶装置があります。

‣SSD

‣ハードディスク(HDD)

‣ブルーレイ/DVD/

‣増設インターフェース

(URL http://buffalo.jp/company/jigyo/ ; 取得日 2017/02/02)

図 1.3: 業種情報の例 3

図 1.4: 本研究が想定する検索エンジンの出力例

出した情報を基に企業を業種カテゴリに分類する手法を提案する．第４章では，提案手法の評価実験について報告し，またその結果を考察する．第５章では，本論文のまとめと今後の課題について述べる．

第2章　関連研究

本章では提案手法の関連研究について述べる．本研究の最終的な目的は，ウェブから信頼性の高い情報を効率よく取得する技術を確立することである．したがって，2.1 節では，ウェブページの信頼性を自動推定する関連研究を紹介する．本研究では，企業のウェブページを対象に，それからウェブページの作成者名 (企業名) やそれに関する情報を抽出する手法を提案する．2.2 節では，ウェブページの作成者の情報を自動抽出する手法を概観する．また，本研究は企業ウェブページを業種によって自動分類する手法も提案する．2.3 節ではその関連研究を紹介する．最後に，2.4 節では先行研究と本研究の違いについて論じる．

2.1　ウェブページの信頼性の評価

ウェブページの信頼性を自動的に評価する研究について述べる．Kakol らはメタ情報やテキスト情報など数多くの素性を用いてウェブページの信頼性をスコアリングした [2]. 彼らは，被験者にウェブページの信頼性を判定させ，信頼性情報が付与されたデータセットを構築した．そのデータセットを訓練データとし，メタ情報やテキスト情報などを学習素性として，ウェブページの信頼度を推測するモデルを学習した．

福島と内海は，ユーザがウェブページの信頼性を判断する際に用いる要素を調べ，それを基にしてウェブページの信頼性を測定する手法を提案した [6]. 彼らは，ウェブページの信頼性を判断する要素をアンケート調査し，40 種類の要素 (素性) を明かにした．さらに，これらの素性を用いてウェブページの信頼度をスコアリングした．

2.2　ウェブページからの作成者情報の抽出

ウェブページから作成者情報を抽出する関連研究について述べる．Changuel らは人名辞書をもとにウェブページの作成者を抽出する手法を提案した [1]. この手法は, 人名辞書を用いてウェブページから網羅的に人名を抽出した後，それらがウェブページの作成者に該当するか否かを判定する分類モデルを教師あり機械学習する．機械学習の素性として用いるのは，人名の前後１５ワード中に E メールや日付などが存在するかという言語的情報と，人名が所属する DOM ノードが DOM ツリー中のどの位置に存在するかという空間情報である．Document Object Model (DOM) はウェブページの HTML ファイルの構

造を表現するためのモデルであり，HTML タグの入れ子構造を表わす木を DOM ツリー，その中のノードを DOM ノードと呼ぶ．DOM ノードはひとつの HTML タグに対応する．機械学習アルゴリズムとして決定木を用いている．抽出対象となる人名の例を図 2.1 に示す．人名 (Mircea NICOLESCU) の近傍で灰色で強調された部分のテキストが言語的情報として扱われる．また，人名を含む HTML タグ (DOM ノード) がウェブページの末尾に存在するという情報が位置情報として扱われる．この手法は多様なウェブページから作成者名を抽出できる．しかし，人名以外は抽出できないという問題がある．また，この手法では人名辞書を事前に用意しておく必要があるが，人名は種類が非常に多く，また新しい人名も今後生成され続けることから，日常世界における全ての人名を網羅した辞書を事前に用意することは難しいという問題点もある．．

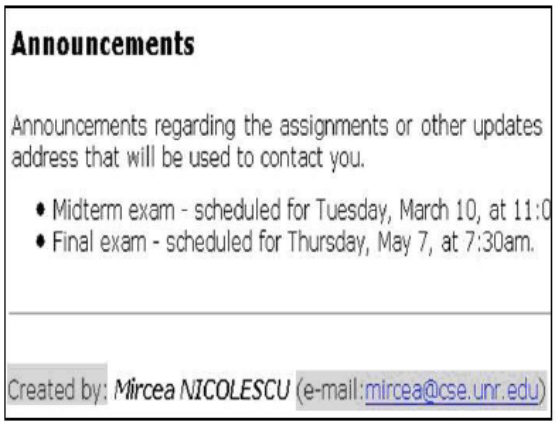


図 2.1: 抽出対象となる作成者名の例 [1]

百瀬らは，ウェブページのレイアウトを用いて情報発信者情報を抽出する手法を提案した [7]．この手法では，まずウェブページをグリッドに分割し，どのグリッドに発信者情報が属するか判定する．次に，そのグリッドに所属する DOM ノードに含まれる言語情報や位置情報を用いて，機械学習で発信者情報が属する DOM ノードを特定する．さらに，その DOM ノードから情報発信者名を抽出する．実際にウェブページをグリットに分割し，発信者情報に関する記述が存在する位置を特定した例を図 2.2 に示す．この手法は幅広いウェブページから情報発信者名を抽出できるが，情報発信者名が属するグリッドが特定できなかった場合に抽出精度が大きく下がる問題点がある．

Kato らは，ウェブページの情報発信元を特定するためのサブタスクとして，ウェブページの作成者名を抽出した [3]．まず，ウェブページから言語情報などを用いてルールベースで作成者の候補を抽出する．これらの候補が真の作成者か否かを判定する分類器を機械学習によって自動獲得する．機械学習には，ウェブページにおける作成者候補とメインコンテンツとの距離や，個人名や組織名などの名詞が含まれているかなどの言語情報を素性として用いる．この手法は事前に辞書を用意しなくても作成者名が抽出できるという利点が

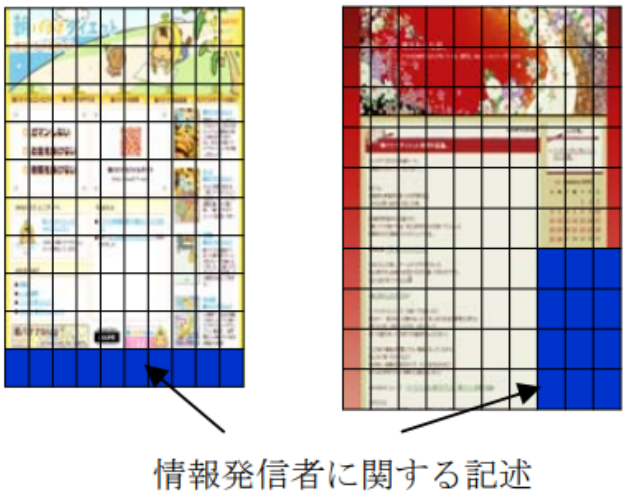


図 2.2: 情報発信者のグリッドを特定した例 [7]

ある. しかし,Changuel らの手法 [1] や百瀬らの手法 [7] と同様に，作成者名以外で情報の信頼性の判定に有用な情報，例えば作成者のプロフィールなどは抽出されない.

堀と白井は, ブログからサイト作成者 (ブロガー) を抽出する手法を提案した [4]. この手法は, ブロガーの名前だけでなく, その人の年齢, 性別, 職業などのプロフィールも合わせて抽出する. しかし, 抽出対象がブログに限定されるため, 一般のウェブページから同様に作成者情報を取得できるかは不明である.

2.3 企業ウェブページの業種分類

企業のウェブページを業種別に自動分類することも試みられている．佐々木と新納は，企業のウェブページを３２個の業種カテゴリに自動分類する手法を提案している [5]．まず，企業のトップページ，およびトップページから深さ 5 までのサイト内の下位ページを取得する．これから名詞，動詞，形容詞を抽出し，Bag-of-words モデルの素性集合を作成する．最後に，これらの素性を基にナイーブベイズモデルを学習し，企業を業種カテゴリに分類する．データセットとして Yahoo!ファイナンス[1]から 2947 件の企業ウェブサイトを取得，これを 9:1 に分けて訓練データとテストデータとし，これを用いて評価実験を行った．分類の正解率は 41.8%であった．

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

2.4 本研究の特徴

ウェブから正しい情報を取得するためには，情報の信頼性を自動的に推定することが理想的だが，情報の真偽を判定するためにはテキストだけでなく様々な要因を考慮する必要があるため，一般には難しい．2.1 節で述べた先行研究とは異なり，本研究では，ウェブ上の情報の信頼性はユーザが判断し，その判断の手助けとなるような情報 (具体的には作成者情報) を提示するという立場を取る．

本研究は作成者情報をウェブページから抽出するという点では 2.2 節で述べた先行研究と共通しているが，多くの先行研究が人名のみを抽出するのに対し，本研究ではプロフィールのような詳細な作成者情報の抽出を試みる．また，堀と白井の手法 [4] では作成者のプロフィールを抽出しているが，その対象はブログに限られる．ブログではプロフィールの書式が比較的固定されていると考えられるため，その抽出も比較的容易であると予想される．一方，本研究は，ブログではなく一般のウェブページ (企業のウェブページ) から作成者情報 (企業のプロフィールに相当する情報) を抽出する点に特徴がある．

佐々木と新納の手法 [5] と同様に，本研究も企業ウェブページを業種に基づいて自動分類する手法を提案する．この先行研究ではウェブページ全体から業種の自動分類のための素性を抽出しているのに対し，本研究ではまず企業の業種情報を抽出し，これから抽出される素性とそれ以外のテキストから抽出される素性を区別して扱う点に特徴がある．企業の業種情報は業種の内容をよく表わしていると考えられるため，業種情報内の素性を重視して分類モデルを学習することで業種の自動分類の精度が向上することが期待できる．

第3章　提案手法

3.1　概要

1 章で述べたように，本研究では，与えられた企業ウェブサイトに対し，まずその企業の業種を説明するテキスト (業種情報) を抽出し，それを基に企業の業種をあらかじめ決められたカテゴリに教師あり機械学習を用いて分類する．その処理の概要を図 3.1 に示す．この図に示したように，本研究では，企業を業種によって分類するため，企業のウェブサイトから以下の 3 種類の業種情報を抽出する．

Description, Keywords
: HTML ファイルのヘッダにおいて，name 属性が description ならびに keywords である ⟨meta⟩ タグでマークアップされているテキスト．例を図 3.2 に示す．図中の赤い枠で囲まれた部分が抽出するべき業種情報である．一般に，HTML 文書において，description にはウェブページの説明文，keywords にはそのウェブページに関連するキーワードが書かれている．企業のウェブページにおいては，これらは会社の業種を説明する文やキーワードであることが多い．

業種説明
: 企業の業種を説明したテキスト．企業の概要がまとめられているページに存在すると仮定する．例を図 3.3 に示す．この例では，企業の概要が表にまとめて掲載されている．その中のひとつに「事業内容」があり，赤い枠で示した箇所に企業の業種に関する情報が書かれている．

事業説明
: 企業の事業内容を説明したテキスト．ここでは，事業説明はトップページとは別の独立したページにまとめて記述されていると仮定する．例を図 3.4 に示す．このページでは企業の展開する事業情報がまとめて掲載されている．赤い枠で囲まれたテキスト，つまりページ全体が事業説明に該当する．

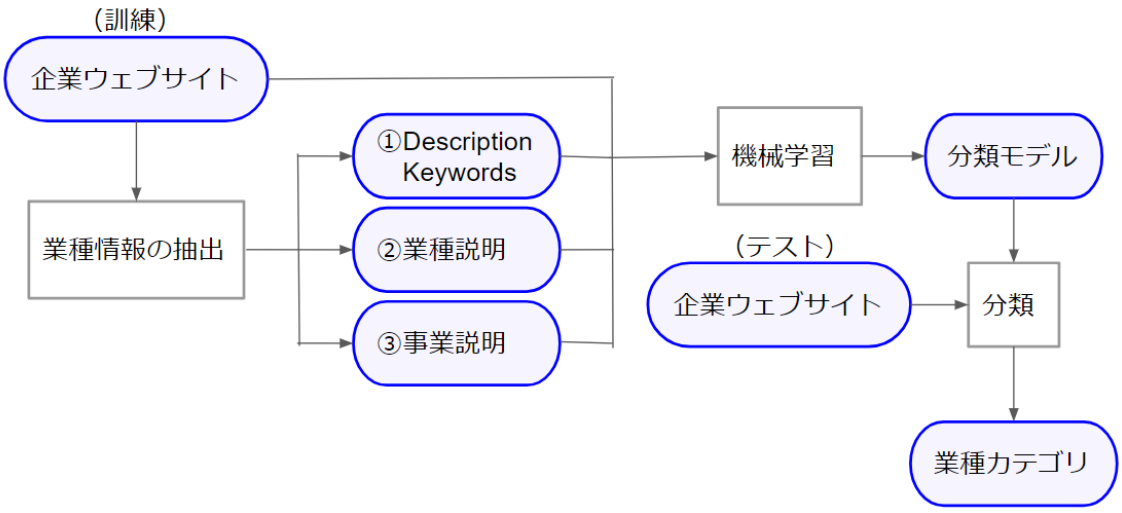


図 3.1: 提案手法の概要

```
<html lang="ja" class="no-js">
<head>
<meta charset="utf-8">
<title>株式会社富士薬品</title>
<meta name="description" content="富士薬品は、複合型医薬品企業として、置き薬・配置薬を中心に、医薬品の製造・販売・研究開発を行います。">
<meta name="keywords" content="富士薬品,医薬品,配置薬,置き薬,常備薬,救急箱,セイムス,ドラックストア,薬局">
```

(URL http://www.fujiyakuhin.co.jp/ ; 取得日 2017/02/02)

図 3.2: Description,Keywords の例

会社概要

商号	株式会社富士薬品
創業	1930年（昭和5年）2月
会社設立	1954年（昭和29年）4月
代表者	代表取締役社長　高柳 昌幸
本社	〒330-0854 埼玉県さいたま市大宮区桜木町4丁目383
資本金	314,559,500円
事業内容	医薬品等の配置薬販売事業、薬局販売等、製造
	【取締役】 代表取締役社長　高柳 昌幸

(URL http://www.fujiyakuhin.co.jp/company/company.php ; 取得日 2017/02/02)

図 3.3: 業種説明の例

(URL http://www.fujiyakuhin.co.jp/business/ ; 取得日 2017/02/02)

図 3.4: 事業説明の例

図 3.5 は企業ウェブページから業種カテゴリの分類器を学習するための素性を抽出する処理の流れを示す．まず，企業のトップページから，3 種の業種情報を抽出する．業種説明を抽出する際には「会社概要ページ」を，事業説明を抽出する際には「事業説明ページ」を，それぞれ検出する．これらは企業ウェブサイト内のページであり，トップページからリンクを辿って検出できるものとする．次に，これらのテキストから自立語を抽出する．さらに，企業ウェブサイトのトップページ内のテキストからも自立語を抽出する．これらを素性とし，さらにそれぞれの重みを決定して，素性ベクトルを得る．

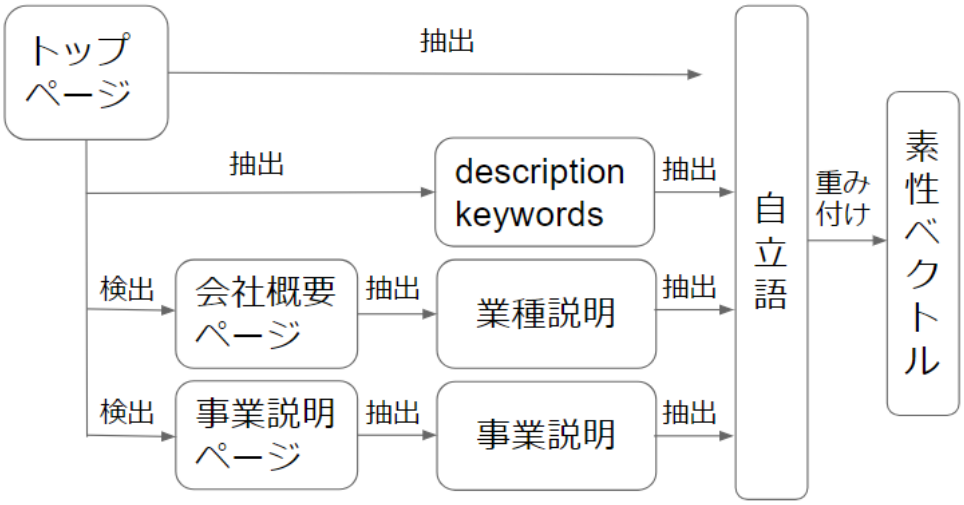


図 3.5: 提案手法の概要 (素性の抽出)

3.2 業種情報の抽出

本節では，企業のウェブサイトから業種情報を抽出する手法の詳細を説明する．業種情報を抽出する際には，企業ウェブサイトの HTML ファイルを解析し，そこから必要なデータを取得する処理が必要である．本研究では，HTML の構文解析は Beautiful Soup[1] を用いて行った．Beautiful Soup は HTML や XML ファイルからデータを取得する Python のライブラリであり，ファイルの構文解析や構文木の探索，検索，修正を比較的簡単に行うことができる．

3.2.1 Description，Keywords の抽出

Description と Keywords は企業のトップページの HTML ファイルから機械的に抽出できる．その際，Description や Keywords が複数ある場合は，最初に取得されたもの

[1]https://www.crummy.com/software/BeautifulSoup/bs4/doc/

表 3.1: 会社概要ページを示唆するキーワード (Rule T1)

Kw-Ta
グループ概要，会社概要，企業概要，法人概要，企業情報，法人情報，会社情報，グループ情報，会社案内，法人案内，企業案内，グループ案内，わたしたち，私たち

(HTML ファイルにおいて先頭に近い位置にあるもの) を業種情報として抽出する．また，Description や Keywords が存在しないときは情報を抽出しない．

3.2.2 業種説明の抽出

業種説明の抽出は以下の 3 つのステップからなる．

ステップ 1: 会社概要ページのリンクの検出

会社の概要を説明しているページを「会社概要ページ」と定義し，企業のトップページの中からこれへのリンクを検出する．以下の 2 つのルールを設定し，そのいずれかに当てはまる ⟨a⟩ タグを全て検出する．

Rule T1

会社概要のページであることを示唆するキーワード (本研究では業種説明を抽出するためにいくつかの種類のキーワードを用いる．区別のため，キーワードのタイプを記号で表記する．Kw-Ta[2]) を含む ⟨a⟩ タグを検出する．用意したキーワードの総数は 14 個である．その一覧を表 3.1 に示す．キーワードは主に「グループ」「会社」「企業」「法人」と「概要」「情報」「案内」の組み合わせで構成されている．また,「私たち」のような一人称のキーワードも含む．

Rule T2

ナビゲーションを示すキーワード (Kw-Tb) をテキストに含み，かつ会社を示唆するキーワード (Kw-Tc) をリンク先 URL に含む ⟨a⟩ タグを検出する．Kw-Tb の数は 6，Kw-Tc の数は 11 である．それぞれの一覧を表 3.2 と表 3.3 に示す．ナビゲーションを示唆するキーワードとしては,「〜について」や「〜はこちら」などリンクテキストによく使われる単語を選定した．URL が含むべきキーワードについては，会社，プロフィール，概要を表わす英単語を選定した．

[2]本研究では業種説明を抽出するためにいくつかの種類のキーワードを用いる．区別のため，キーワードのタイプを記号で表記する．Kw はキーワードを，T は業種 (Type の T) を表わす．a はキーワードのタイプの識別子である．

表 3.2: ナビゲーションを示唆するキーワード (Rule T2)

Kw-Tb
ついて，こちら，詳細，特色，とは，概要

表 3.3: 企業概要ページの URL が含むべきキーワード (Rule T2)

Kw-Tc
about，ABOUT，company，COMPANY，corporate，CORPRATE，profile，PROFILE，gaiyou

表 3.4: 業種説明の見出しの検出に用いるキーワード

Kw-Td
事業内容，事業種目，業務内容，業務種目，営業内容，業種，事業案内，営業品目，業務案内，主要業務，営業案内，主要事業，主な事業，事業の内容，事業概要，業務の内容，業務内容，事業目的，営業内容，業務目的，営業種目，営業目的

ステップ 2: 業種説明の見出しの検出

ステップ 1 で検出したリンクのリンク先ページの HTML ファイルを取得し，その中から業種説明の見出しを含む HTML タグを検出する．具体的には，業種説明の見出しであることを示唆するキーワード (Kw-Td) を含む ⟨th⟩, ⟨td⟩, ⟨dt⟩ タグを検出する．用意した Kw-Td の数は 22 個である．その一覧を表 3.4 に示す．キーワードは主に,「事業」「営業」「業務」「種目」「内容」「案内」「目的」などの組み合わせで構成されている．

ステップ 3: 業種説明の抽出

業種説明を含む HTML タグを検出し，その中のテキストを業種説明として抽出する．ステップ 2 で検出した HTML タグ (業種説明の見出し) を H とし，本ステップで検出するべき業種説明を含む HTML タグを T とすると，T は表 3.5 に示した条件にしたがって検出する．また，T が空白や記号のみしか含まなかったとき，その HTML タグをスキップして，次に条件を満たすものを探して T を検出する．

表 3.5: 業種説明を含む HTML タグの条件

H	T の抽出条件
⟨th⟩	H の次に出現する ⟨td⟩ タグ
⟨td⟩	H の次に出現する ⟨td⟩ タグ
⟨dt⟩	H の次に出現する ⟨dd⟩ タグ

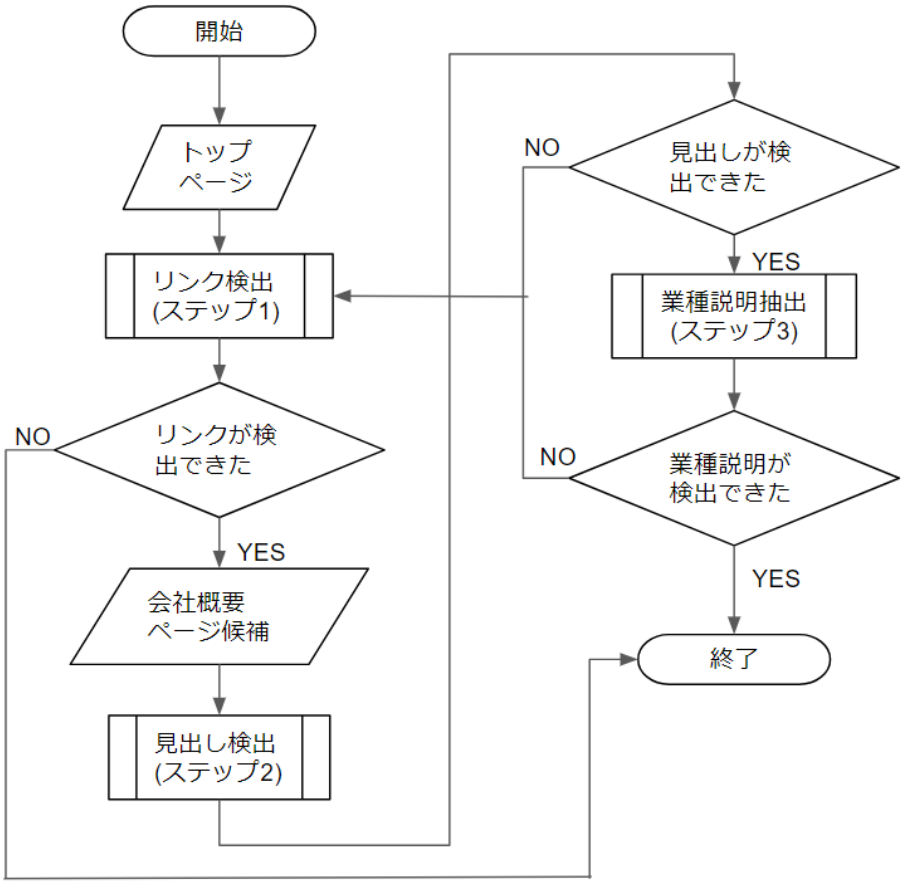


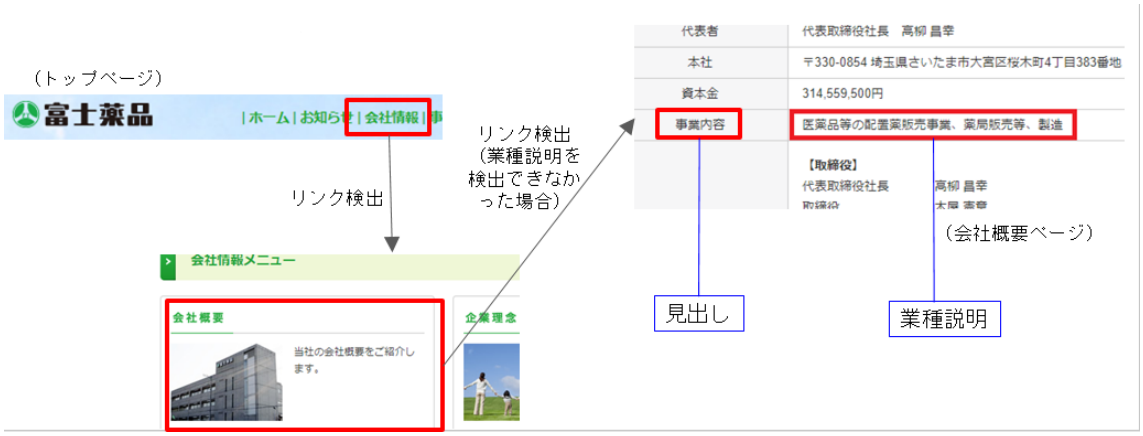
図 3.6: 業種説明の抽出処理のフローチャート

ルールの再帰的適用

上記のステップ 1 には成功したが，ステップ 2 や 3 は失敗したとき，ステップ 1 で検出したページを基点として，ステップ 1〜3 を再度実行する．業種説明が抽出されるか，ステップ 1 で企業概要ページのリンクの検出に失敗するまで，同じ処理を繰り返す．その処理のフローチャートを図 3.6 に示す．

業種情報の抽出例

業種説明を抽出する処理の例を図 3.7 に示す．まず，トップページから会社概要ページへのリンクとして図中の「会社情報」というリンクを検出する．次に，リンク先のページから業種説明の抽出を試みるが，失敗する．そこで，このページから次の会社概要ページのリンクの検出を試み，図中の「会社概要」というリンクを検出する．リンク先の会社概要ページから,「事業内容」という見出しを検出し，その近傍にある「医薬品等の配置薬販売事業，薬局販売等，製造」というテキストを業種説明として抽出する．



左上の図: URL http://www.fujiyakuhin.co.jp/ ; 取得日 2017/02/02
左下の図: URL http://www.fujiyakuhin.co.jp/company/ ; 取得日 2017/02/02
右の図: URL http://www.fujiyakuhin.co.jp/company/company.php ; 取得日 2017/02/02

図 3.7: 業種説明の抽出例

3.2.3　事業説明の抽出

本節では，企業のウェブサイトから事業説明を抽出する手法の詳細を説明する．まず，事業の内容を紹介するページを事業説明ページと定義し，トップページから事業説明ページへのリンクを検出する．リンクの検出は Rule B1 と Rule B2 という 2 つのルールで実現する．Rule B1 をまず適用し，検出できなかったときには Rule B2 を適用する．以下，これらのルールの詳細を説明する．

Rule B1

事業説明を示唆するキーワード (Kw-Ba[3]) を含む ⟨a⟩ タグを検出する．用意した Kw-Ba の数は 34 である．キーワードの一覧を表 3.6 に示す．キーワードは主に,「事業」「営業」「業務」「種目」「内容」「案内」「目的」「商品」「製品」などの組み合わせで構成されている．

Rule B2

特定のキーワード (Kw-Bb) をテキストに含み，かつ事業を示唆するキーワード (Kw-Bc) をリンク先 URL に含む ⟨a⟩ タグを検出する．Kw-Bb の数は 2，Kw-Bc の数は 6 である．それぞれのキーワードの一覧を表 3.7，表 3.8 に示す．

次に，リンク先の事業説明ページの HTML ファイルを取得する．広告やメニューなど，事業説明と関係のないテキストを除くため，そのページのメインコンテンツに相当する

[3]業種説明と同様に，本研究では事業説明を抽出するためにいくつかの種類のキーワードを用いる．区別のため，キーワードのタイプを記号で表記する．Kw はキーワードを，B は事業 (Business の B) を表わす．a はキーワードのタイプの識別子である．

表 3.6: 事業説明ページの検出に用いるキーワード (Rule B1)

Kw-Ba
事業内容，事業の内容，業務内容，業務の内容，営業内容，事業目的，事業案内，業務目的，業務案内，営業目的，営業案内，事業情報，主な事業，業務情報，事業概要，事業紹介，業務内容，業務紹介，営業内容，施設紹介，営業種目，施設案内，事業種目，製品案内業務種目，商品案内，業種，商品一覧，営業品目，製品一覧，主要業務，取扱製品，主要事業，取扱商品

表 3.7: 事業説明ページの検出に用いるキーワード (Rule B2)

Kw-Bb
事業，業務

HTML タグを検出し，それが包含するテキストを事業説明として抽出する．メインコンテンツは加藤らの手法 [3] を用いて検出した．アルゴリズムを図 3.8 に示す. およそ以下の手続きでメインコンテンツを検出する．

1. HTML ファイルの Document Object Model を作成し，変数 *DOM* に代入する．
2. DOM ツリーから body タグに相当するノードを取得し，変数 *body* に代入する (下から 2 行目)．
3. *body* を引数として関数 MAINBLOCK を呼び出す (下から 1 行目)．MAINBLOCK は，与えられた DOM ノード n の下位のノードの中からメインコンテンツに相当する DOM ノードを探索する関数である．
4. 関数 MAINBLOCK の中では，まず n のテキスト長を l_n とおく (3 行目)．また，n の子ノードの集合を C とする (4 行目)．
5. 6 行目以降のループでは，C 中の子ノード c_i の中からメインコンテンツに相当する DOM ノードの候補を探索し，変数 $main$ に格納する．具体的には，c_i のテキスト長 l_i を調べ，もし $l_i/l_n > t_m$ という条件を満たすなら，つまり c_i が包含するテキ

表 3.8: 事業説明ページの URL が含むべきキーワード (Rule B2)

Kw-Bc
business，BUSINESS，project，PROJECT，Business，Project

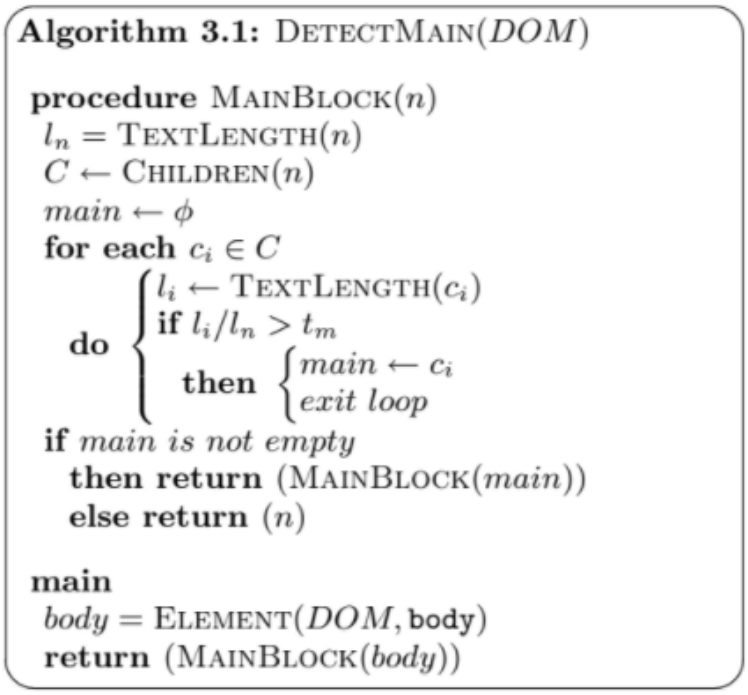


図 3.8: メインコンテンツ領域検出アルゴリズム (Kato et al.(2008) p.39 Figure 4)

スト量が親ノード n が包含するテキスト量の大部分 (閾値 t_m 以上) を占めるなら，c_i を $main$ とする．

6. 前述の処理で $main$ が見つかれば，これを引数として関数 MAINBLOCK を呼び出す (下から 5 行目)．ノード n の子ノードの中には，テキストの大部分を占める子ノード $main$ がある一方，広告や目次のようにメインコンテンツに比べてはるかにテキスト量の少ない子ノードも存在するため，$main$ はメインコンテンツに該当する DOM ノードとはみなさず，$main$ の下位のノードの中からメインコンテンツに相当する DOM ノードを探索する．一方，もし $main$ が見つからなければ，n をメインコンテンツとして返す (下から 4 行目)．

本研究では閾値 T_m を 0.5 と設定している．

事業説明の抽出例

事業説明を抽出する処理の例を図 3.9 に示す．左上のトップページから事業説明ページへのリンクとして「事業内容」というリンクを検出する．次に，リンク先の事業説明ページにメインコンテンツ抽出のアルゴリズムを適用し，赤枠で囲われている部分のテキストを事業説明テキストとして抽出する．

左上の図: URL http://www.fujiyakuhin.co.jp/ ; 取得日 2017/02/02
右，左下の図: URL http://www.fujiyakuhin.co.jp/business/ ; 取得日 2017/02/02

図 3.9: 事業説明の抽出例

3.2.4 キーワードの選定

これまで述べてきた提案手法は，基本的にはキーワードに基づくルールベースの手法である．ここではキーワードの選定方法を説明する．まず，Open Directory Project (ODP)[4] のサイトから，業種説明が含まれている会社概要ページ，事業説明が含まれている事業紹介ページをそれぞれ 100 件づつ人手で収集した．次に，会社概要ページや業種説明ページへのリンクにおけるリンクテキストやリンク先 URL，あるいは業種説明の見出しに共通して出現する単語や文字列を人手で選定し，会社概要ページや業種説明ページの抽出条件として用いるキーワードとする．業種説明，事業説明のリンク検出については，会社名などの固有名詞やそのサイト特有の言い回しを除外し，多くのウェブページに共通すると思われるキーワードのみを選定した．また，業種説明の見出し検出についても同様にサイト特有の言い回しを除外し，キーワードを選定した．除外したキーワードの例を示す. 図 3.10 の「足立はこんな会社」のリンク先には会社概要ページが存在するが，このサイト特有のキーワードであり，他の企業のウェブサイトで使われるとは考えにくいため，キーワード Kw-Ta として用いない．図 3.11 の「八千代商工会議所とは」というリンクの先には会社概要ページが存在するが，URL が「/yachiyocci」となっている.「yachiyocci」は明らかに固有名詞なのでキーワード Kw-Tc として用いない．また，図 3.12 の「製造設備と研究開発」のリンク先には事業説明ページが存在するが，この業種特有の表現であるためキーワード Kw-Ba として用いない．

[4]ODP については 3.3 節で詳しく説明する．

(http://www.adachi-bag.co.jp/ ； 取得日 2017/02/02)

図 3.10: キーワード Tw-Ta として採用しなかった例

(http://www.yachiyocci.jp/ ； 取得日 2017/02/02)

図 3.11: キーワード Kw-Tc として採用しなかった例

(http://www.juzen-chem.co.jp/ ； 取得日 2017/02/02)

図 3.12: キーワード Kw-Ba として採用しなかった例

3.3 業種カテゴリの定義

本節では業種カテゴリに定義について述べる．本研究では，ウェブディレトリーサービスの一つである Open Directory Project(ODP) の日本語サイト[5]で定義されているウェブサイトのカテゴリを参考に，28 個の業種カテゴリを設定した．ODP は様々なウェブサイトの URL がカテゴリ毎に分類されている．基本的には，企業のウェブサイトを多く含む ODP のカテゴリを本研究における業種カテゴリと定義する．企業のウェブサイトを含むカテゴリとして「ビジネス」「ニュース/メディア[6]」「各種資料/教育」の 3 つを選択した．カテゴリ名の中の / は ODP における階層を表わす.「ニュース/メディア」は，新聞社や出版社などの企業が多く含まれるため,「各種資料/教育」は大学，専門学校，予備校などの企業が多く含まれるために選定した．ただし,「ビジネス」「ニュース/メディア」「各種資料/教育」の下位の ODP カテゴリをそのまま業種カテゴリとするのではなく，必要に応じて ODP カテゴリの修正や業種カテゴリとして採用する ODP カテゴリの選別を行った．具体的には，前述した企業のウェブサイトを含む 3 つのカテゴリの下位の ODP カテゴリの中で，業種カテゴリとしてふさわしくないものを人手で除外した．また，それぞれの業種カテゴリに分類される企業の数がだいたい同じになるように業種カテゴリのセットを定義するという指針を設け，ODP の各カテゴリに登録されているウェブサイトの数を参照し，登録されているウェブサイトの少ない ODP カテゴリを別の業種カテゴリに併合したり，登録ウェブサイト数の多い ODP カテゴリは複数の業種カテゴリに分割するなどの処理を行った．

ODP のカテゴリと本研究で設定した業種カテゴリの対応関係の一部を図 3.13 に示す．全ての対応関係は付録 A に記す．図 3.13 で示されている木構造は，ODP における「ビジネス」をルートノードとした階層構造である．以下，図 3.13 で使われている記号の意味を説明する．

- 〈〉は ODP カテゴリを示す．( ) 内の数値は ODP における各カテゴリの登録ウェブサイトの数を示す．
- 《》は，ODP カテゴリのうち，本研究で業種カテゴリのひとつとして採用したカテゴリを示す．{} は業種カテゴリの識別番号 (1〜28) である．
- 《》で示した業種カテゴリは，ODP の階層構造の下位にあるカテゴリを原則として全て含むものとする．例えば，図 3.13 における「薬品・バイオテクノロジー」の下位に位置する<ベンチャーキャピタル>，<団体>，<薬品>，<雇用・スタッフ> というカテゴリは，全て《薬品・バイオテクノロジー》という業種カテゴリに属するとみなす．
- 【】は ODP カテゴリに対する修正作業を示す．

[5]http://dmoztools.net/World/Japanese/
[6]正式なカテゴリ名は「オンラインメディア, ラジオ, 新聞, 雑誌, テレビ, 放送, 通信社」である．

- 【→ category/】 は，その ODP カテゴリを category が示す別の業種カテゴリに併合することを表わす．
- 【← category/】 は，その ODP カテゴリが，category が示す別の上位の ODP カテゴリ (category) から移動し，新しい上位の ODP カテゴリに属することを表わす．
- 【×】は，業種カテゴリとしてふさわしくないため，業種カテゴリとして採用しなかった ODP カテゴリを示す．
- 【新設】は，ODP カテゴリとしては存在しないが，いくつかの下位の ODP カテゴリをマージして新設した業種カテゴリを表わす．例えば，図 3.13 における《アパレル・装飾品》は，<服飾・アパレル>，<かばん・スーツケース>，<宝飾・貴金属>，<時計> の 4 つの ODP カテゴリをマージして作成した新設の業種カテゴリである．

上記の手続きで決定した業種カテゴリの一覧を表 3.9 に示す．

表 3.9: 業種カテゴリの一覧

1	IT	15	環境・資源
2	食品	16	投資
3	教育・受験	17	建設・土木
4	電機・エレクトロニクス	18	広告・マーケティング
5	雇用	19	小売
6	金融サービス	20	宿泊・飲食・接客
7	運輸・物流	21	団体
8	農林・水産	22	印刷・出版
9	財務・会計	23	化学
10	製品・サービス（産業向け）	24	企業向けサービス（法律など）
11	アパレル・装飾品	25	不動産
12	薬品・バイオテクノロジー	26	医療・ヘルスケア
13	自動車	27	ニュース・メディア
14	素材	28	アート・娯楽

3.3.1 業種カテゴリの分類器の学習

正解の業種カテゴリが付与された企業ウェブページの集合を用意し，これを訓練データとする．ODP における企業に関連するカテゴリに登録されている企業のウェブページは，ODP カテゴリと業種カテゴリの対応表を用いれば，その正解の業種カテゴリを自動的に決めることができるため，訓練データは比較的容易に構築できる．詳細は 4.2.1 項で後述

<ビジネス>
├ 《アパレル・装飾品》{11} (969)【新設】
│ ├ <服飾・アパレル> (666) 【←製品・サービス（一般消費者向け）/】
│ ├ <かばん・スーツケース> (96)【←製品・サービス（一般消費者向け）/】
│ ├ <宝飾・貴金属> (147) 【←製品・サービス（一般消費者向け）/】
│ ├ <時計> (60) 【←製品・サービス（一般消費者向け）/】
├ 《薬品・バイオテクノロジー》{12} (421)
│ ├ <ベンチャーキャピタル> (5)
│ ├ <団体> (5)
│ ├ <薬品> (421)
│ └ <雇用・スタッフ> (10)
└ <製品・サービス（一般消費者向け）> (3,202)【X】
├ <おもちゃ・遊具> (91)【X】
├ <かばん・スーツケース> (96)【→アパレル・装飾品/】
├ <オフィス・文房具> (180)【X】
├ <システムとソフトウェア> (0)
├ <スポーツ用品> (339) 【→アート・娯楽/】
├ <ニュースとメディア> (0)
├ <ペット・動物> (134)【X】
├ <健康・美容> (506) 【→医療・ヘルスケア/】
├ <冠婚葬祭> (187)【X】
├ <卸売・輸出入> (7)【X】
├ <団体> (0)
├ <宗教・儀典> (107)【X】
├ <宝飾・貴金属> (147) 【→アパレル・装飾品/】
├ <家庭・園芸> (545) 【→アート・娯楽/】
├ <家電・カメラ> (28) 【→電機・エレクトロニクス/】
├ <小売> (5)【→小売/】
├ <探偵・調査> (5)【X】
├ <教育産業> (60) 【→教育・受験/】
├ <時計> (60) 【→アパレル・装飾品/】
├ <服飾・アパレル> (666) 【→アパレル・装飾品/】
├ <育児・子供> (28)【X】
├ <資格・スキル> (1)【X】
└ <雇用・スタッフ> (0)

図 3.13: ODP カテゴリと業種カテゴリの対応 (一部)

する．訓練データにおける個々のウェブページから学習のための素性を抽出し，素性ベクトルを作成する．

まず，3.2 節で説明した手法で抽出した業種情報を形態素解析する．また，企業ウェブサイトのトップページのテキストも同様に形態素解析する．形態素解析器として JUMAN[7] を用いる．次に，形態素解析結果から自立語のみを学習素性として抽出する．具体的には，品詞が「助詞」「助動詞」「記号」以外の単語を自立語として抽出する．

次に，各素性 (自立語) の重みを設定する．重みの定義を式 (3.1) に示す．

$$w^i = \alpha \times f^i_{profile} + f^i_{other} \tag{3.1}$$

ここで，w^i は単語 i の重み，$f^i_{profile}$ は企業プロフィール (業種情報) における単語 i の出現頻度，f^i_{other} は企業プロフィール以外のテキストにおける単語 i の出現頻度，α は業種情報に高い重みを与えるパラメータである．業種情報から抽出した素性に高い重みを与えるのは，業種情報は企業の業種の種類を表わすテキストであり，それに含まれる単語は業種の分類に有効であると考えられるためである．本研究では直観に基づいて α を 4 と設定する．

学習データから得られた素性ベクトルの集合を用いて，業種カテゴリを分類するモデルを機械学習する．学習アルゴリズムとして，ナイーブベイズモデルとランダムフォレストを用いる．学習には機械学習ライブラリである Scikit Learn[8]を用いた．ナイーブベイズとランダムフォレストの学習パラメータはデフォルト値を用いた．ナイーブベイズには以下の 3 種類の学習パラメータが存在する．

alpha
: 平滑化処理をする際の小数値を指定する．デフォルト値は 1.0 で，ラプラススムージングを行う．

fit_prior
: True または False で指定する．True でクラスごとに事前確率を算出する．False では事前確率に一様分布を使用する．デフォルトは True.

class_prior
: 小数のタプルまたは None を指定する．指定したクラスの事前確率に任意の値を設定できる．デフォルトは None.

ランダムフォレストには 1 7 種類の学習パラメータが存在する．主な 3 つを以下に述べる．

n_estimators
: 整数値を指定する．値に応じて部分木の数を変更する．デフォルト値は 1 0 .

[7] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[8] http://scikit-learn.org/stable/documentation.html

max_features

整数値，小数値，auto，sqrt，log2，None を指定する．決定木において，最適な分割を探す際に考慮する素性の数を指定する．整数値の場合，指定した数の素性を考慮する．小数値の場合，(指定した値) × (全素性数) の個数を使用する．auto の場合，$\sqrt{(\text{全素性数})}$ の個数を使用する．sqrt の場合，auto と同じく $\sqrt{(\text{全素性数})}$ の個数を使用する．log2 の場合，$\log 2(\text{全素性数})$ の個数を使用する．None の場合，全素性を使用する．デフォルトは auto.

min_samples_split

整数値または小数値を指定する．ノードを分割するときに必要な最小のサンプル数を指定する．小数値の場合，(指定した値) × (全サンプル数) の個数が必要となる．デフォルトは 2 .

第4章　評価実験

本章では，提案手法の評価実験について述べる．提案手法は，3.2 節で述べた業種情報の抽出と，3.3 節で述べた業種カテゴリの分類に分けられる．4.1 節では業種情報抽出の評価実験について，4.2 節では業種カテゴリ分類の評価実験について報告する．

4.1　業種情報抽出手法の評価

本研究で提案する手法によって企業のウェブサイトから業種情報をどれだけ正確に抽出できるかを評価する．人手で正解の業種情報を付与したテストデータを用意し，正解の業種情報と自動抽出した業種情報と比較する．

4.1.1　実験データ

実験ではテストデータ A とテストデータ B という 2 つのデータを用意した．テストデータ A として，3.2.4 項で説明したキーワードの選定に用いた 100 件のウェブサイトとは別に，ODP に登録されているウェブサイトをランダムに 100 件選択した．この際，28 個の業種カテゴリからなるべく均等にウェブサイトを選択した．表 3.9 における番号 1〜16 の業種カテゴリから 4 つずつウェブサイトを選択し，17〜28 のカテゴリからは 3 つずつウェブサイトを選択した．テストデータにおける業種カテゴリ毎のウェブサイト数の内訳を表 4.1 に示す．テストデータ A は業種情報がどの程度頻繁に企業のウェブページに記載されているかを調べるために用いる．

一方，ほぼ同じやり方でテストデータ B を用意する．ODP に登録されているウェブサイトから，それぞれの業種カテゴリに対し，表 4.1 に示した件数のウェブサイトを選択する．この際，テストデータ A とは異なり，Description，Keywords，業種説明，事業説明の全てが含まれるウェブページを選択する．次に，これら 100 件のウェブサイトに対し，4 種類の業種情報，すなわち Description, Keywords, 業種説明，事業説明を人手で選別し，正解の業種情報としてテストデータに付与した．正解の業種情報の付与は 1 名の作業者が行った．テストデータ B は業種情報抽出手法の評価に用いる．

表 4.1: 業種情報抽出のテストデータにおける業種カテゴリの内訳

1 IT	4	15 環境・資源	4
2 食品	4	16 投資	4
3 教育・受験	4	17 建設・土木	3
4 電機・エレクトロニクス	4	18 広告・マーケティング	3
5 雇用	4	19 小売	3
6 金融サービス	4	20 宿泊・飲食・接客	3
7 運輸・物流	4	21 団体	3
8 農林・水産	4	22 印刷・出版	3
9 財務・会計	4	23 化学	3
10 製品・サービス（産業向け）	4	24 企業向けサービス（法律など）	3
11 アパレル・装飾品	4	25 不動産	3
12 薬品・バイオテクノロジー	4	26 医療・ヘルスケア	3
13 自動車	4	27 ニュース・メディア	3
14 素材	4	28 アート・娯楽	3

4.1.2 評価基準

まず，4 種類のそれぞれの業種情報に対し，それを含むウェブサイトの割合を「業種情報存在率」と定義し，テストデータ A を用いてその値を調べた．業種情報存在率の定義を式で表わすと以下のようになる．

$$\text{業種情報存在率} = \frac{\text{業種情報が含まれるウェブサイトの数}}{\text{ウェブサイトの数 (100 件)}} \tag{4.1}$$

業種情報存在率は，業種情報を抽出し，その情報を基に業種カテゴリを分類するという本研究のアプローチの有効性を評価するものである．業種情報存在率が低ければ，すなわち多くの企業が業種情報をウェブサイトに掲載していないような状況では，本研究のアプローチはあまり有効でないといえる．

次に，テストデータ B に対し，3.2 節で述べた提案手法で業種情報を自動抽出した．抽出した業種情報を人手で付与した正解と比較し，その性能を評価した．評価基準は，精度，再現率，F 値である．それぞれの定義を式 (4.2), (4.3), (4.4) に示す．

$$\text{精度} = \frac{\text{抽出された正解の業種情報の数}}{\text{抽出された業種情報の数}} \tag{4.2}$$

$$\text{再現率} = \frac{\text{抽出された正解の業種情報の数}}{\text{正解の業種情報の数}} \tag{4.3}$$

$$F \text{ 値} = \frac{2 \cdot \text{精度} \cdot \text{再現率}}{\text{精度} + \text{再現率}} \tag{4.4}$$

4.1.3 実験結果と考察

4 種の業種情報それぞれの精度，再現率，F 値を表 4.2 に示す．表における括弧内の数値は，精度，再現率の分子，分母に相当する業種情報の実数を示している．

表 4.2: 業種情報抽出の評価結果

	精度	再現率	F 値
Keywords	1(100/100)	1(100/100)	1
Description	1(100/100)	1(100/100)	1
業種説明	1(95/95)	0.95(95/100)	0.97
事業説明	1(91/91)	0.91(91/100)	0.95

表 4.2 の結果から，4 種類の全ての業種情報について，精度，再現率ともに十分高いことがわかった．ただし，Description と Keywords は，HTML のタグと属性によって自動的に抽出できるため，F 値が 1 となるのは自明である．業種説明と事業説明については，F 値は 0.95 を越えている．このことから，提案手法は企業ウェブサイトから正確に業種情報を抽出できることがわかった．

次に，4 種類の業種情報のそれぞれの業種情報存在率を表 4.3 に示す．事業説明の存在率は 0.36 と低かった．それ以外の業種情報は 70%から 85%のウェブサイトで存在することが確認できた．また，何らかの業種情報が存在するウェブサイトの割合は 92%であった．したがって，業種情報を抽出した上で業種カテゴリを分類する本研究のアプローチは無効ではないことが確認された．

表 4.3: テストデータに対する業種情報存在率

	業種情報存在率
Keywords	0.8
Description	0.85
業種説明	0.7
事業説明	0.36
4 つの業種情報のいずれか	0.92

4.2 業種カテゴリの自動分類手法の評価

4.2.1 実験データ

実験データとして Open Directory Project (ODP) の日本語階層[1] から獲得した企業ウェブページの集合を用いる．ODP カテゴリを表 3.9 ならびに付録 A の対応表にしたがって本研究の 28 種類の業種カテゴリのいずれかに変換することで，業種カテゴリがタグ付けされたウェブサイトの集合を構築する．

ODP では，1 つのカテゴリが階層上の複数の位置に重複して配置されることがある．また，複数の位置に配置されているときでも，実体として存在するカテゴリは一つのみであり，それ以外の位置では実体として存在するカテゴリへのリンクが張られている．例えば,「ビジネス/建設・土木/システムとソフトウェア」と「ビジネス/IT/建築・土木」は同一のカテゴリであり,「ビジネス/IT/建築・土木」というカテゴリは「ビジネス/建設・土木/システムとソフトウェア」へのリンクとなっている．このようなカテゴリに属する企業ウェブページは，複数の業種カテゴリに属するとみなせる．今回の実験では，一つの企業ウェブページが複数の業種カテゴリに属する場合でも，その中で最も主要な業種カテゴリのみに属するとみなす．すなわち，一つの企業ウェブページは一つの業種カテゴリに属するものとする．最も主要な業種カテゴリは，ODP においてリンクではなく実体として存在する ODP カテゴリに対応する業種カテゴリとする．

このデータセットに含まれる企業ウェブページの合計は 29,364 であった．業種カテゴリ毎のウェブサイト数を表 4.4 に示す．このデータを訓練データ (90%) とテストデータ (10%) に分割した．訓練データは提案手法によって業種を分類するモデルを学習するために用いる．テストデータは，訓練データから学習したモデルの正解率を測るために用いる．

4.2.2 実験設定

この実験では，以下の 11 個の業種カテゴリの自動分類手法を比較する．

BL/NB
: 業種情報を抽出せず，企業ウェブサイトのトップページのみから素性を抽出する，学習アルゴリムとしてナイーブベイズを用いる．

BL/RF
: 業種情報を抽出せず，企業ウェブサイトのトップページのみから素性を抽出する，学習アルゴリムとしてランダムフォレストを用いる．

Pro-BT-W/NB
: 自動抽出した業種情報とトップページから学習素性を抽出する．業種情報による素

[1] http://dmoztools.net/World/Japanese/

表 4.4: 実験データにおける業種カテゴリの内訳

1 IT	519	15 環境・資源	645
2 食品	3898	16 投資	153
3 教育・受験	1395	17 建設・土木	2278
4 電機・エレクトロニクス	902	18 広告・マーケティング	668
5 雇用	275	19 小売	339
6 金融サービス	875	20 宿泊・飲食・接客	882
7 運輸・物流	2064	21 団体	729
8 農林・水産	478	22 印刷・出版	939
9 財務・会計	492	23 化学	468
10 製品・サービス（産業向け）	2933	24 企業向けサービス（法律など）	1169
11 アパレル・装飾品	836	25 不動産	278
12 薬品・バイオテクノロジー	374	26 医療・ヘルスケア	893
13 自動車	1191	27 ニュース・メディア	896
14 素材	585	28 アート・娯楽	2210

性の重みを式 (3.1) で定めたように 4 倍に設定する．学習アルゴリズムとしてナイーブベイズを用いる．

Pro-BT-W/RF
自動抽出した業種情報とトップページから学習素性を抽出する．業種情報による素性の重みを式 (3.1) で定めたように 4 倍に設定する．学習アルゴリズムとしてランダムフォレストを用いる．

Pro-BT/NB
自動抽出した業種情報とトップページから学習素性を抽出する．業種情報によって素性の重みを変更しない．学習アルゴリズムとしてナイーブベイズを用いる．

Pro-BT/RF
自動抽出した業種情報とトップページから学習素性を抽出する．業種情報によって素性の重みを変更しない．学習アルゴリズムとしてランダムフォレストを用いる．

Pro-B/RF
自動抽出した業種情報のみから学習素性を抽出する．業種情報の違いによって素性の重みを変更しない．学習アルゴリズムとしてランダムフォレストを用いる．

Pro-B-WD/RF
自動抽出した業種情報のみから学習素性を抽出する．素性の重みを決める際，式 (3.1) と同じように，Description と Keywords での出現頻度を 4 倍に設定する．学習アルゴリズムとしてランダムフォレストを用いる．

表 4.5: 業種カテゴリ分類手法の一覧

	素性の抽出元	素性の重み付け	学習アルゴリズム
BL/NB	トップページのみ	–	ナイーブベイズ
BL/RF	トップページのみ	–	ランダムフォレスト
Pro-BT-W/NB	業種情報とトップページ	全ての業種情報	ナイーブベイズ
Pro-BT-W/RF	業種情報とトップページ	全ての業種情報	ランダムフォレスト
Pro-BT/NB	業種情報とトップページ	なし	ナイーブベイズ
Pro-BT/RF	業種情報とトップページ	なし	ランダムフォレスト
Pro-B/RF	業種情報のみ	なし	ランダムフォレスト
Pro-B-WD/RF	業種情報のみ	Description,Keywords	ランダムフォレスト
Pro-B-WT/RF	業種情報のみ	業種説明	ランダムフォレスト
Pro-B-WB/RF	業種情報のみ	事業説明	ランダムフォレスト
H		(人手による判定)	

Pro-B-WT/RF

自動抽出した業種情報のみから学習素性を抽出する．素性の重みを決める際，式 (3.1) と同じように，3 種類の業種情報のうち業種説明での出現頻度を 4 倍に設定する．学習アルゴリズムとしてランダムフォレストを用いる．

Pro-B-WB/RF

自動抽出した業種情報のみから学習素性を抽出する．素性の重みを決める際，式 (3.1) と同じように，3 種類の業種情報のうち事業説明での出現頻度を 4 倍に設定する．学習アルゴリズムとしてランダムフォレストを用いる．

H

人手でウェブサイトの業種カテゴリを分類する．業種カテゴリの自動分類の正解率の上限とみなすことができる．300 件のウェブサイトについて調べた．

上記 11 種類の手法の違いを表 4.5 にまとめる．手法の略号に使われている記号の意味は以下の通りである．BL と Pro はそれぞれベースラインと提案手法を表わす．BT と B は，提案手法において，それぞれ業種情報とトップページの両方もしくは業種情報のみから素性を抽出することを表わす．W,WD,WT,WB は素性の重み付けの違いを表わす．NB と RF はそれぞれ学習アルゴリズムとしてナイーブベイズもしくはランダムフォレストを用いることを表わす．

評価基準は正解率を用いる．正解率は業種カテゴリ毎に算出する．その定義を式 (4.5) に示す．

$$\text{正解率} = \frac{\text{正解の業種カテゴリに分類されたウェブサイトの数}}{\text{業種カテゴリに属するウェブサイトの数}} \tag{4.5}$$

また，業種カテゴリの正解率のマイクロ平均も算出し，業種カテゴリの自動分類手法を比較する．

4.2.3 実験結果と考察

業種カテゴリ毎の各手法の正解率を表 4.6，表 4.7，表 4.8 に示す．スペースの都合により，ベースライン手法 (BL/NB, BL/RF) と人手による分類 (H) の結果を表 4.6 に，業種情報とトップページの両方から素性を抽出する提案手法 (Pro-BT*) の結果を表 4.7 に，業種情報のみから素性を抽出する提案手法 (Pro-B*) の結果を表 4.8 に分けて掲載した．一方，正解率のマイクロ平均を表 4.9 に示す．

ベースラインと提案手法を比較する．表 4.9 において，BL/NB と Pro-BT-W/NB，もしくは BL/RF と Pro-BT-W/RF を比較すると，ナイーブベイズ，ランダムフォレストともに提案手法はベースラインを上回った．その差は，ナイーブベイズのときは 1.8 ポイント，ランダムフォレストのときは 1.5 ポイントであった．

機械学習アルゴリズムを比較する．全体的に，ランダムフォレストの正解率はナイーブベイズの正解率を大きく 上回った．例えば，表 4.9 において，Pro-BT-W/RF と Pro-BT-W/NB を比較すると，ランダムフォレストはナイーブベイズを 23.8 ポイント上回った．

提案手法で，業種情報のみを使う場合と業種情報とトップページの両方を使う場合の比較する．表 4.9 において，Pro-B/RF と Pro-BT/RF を比較すると，業種情報とトップページの両方を使う場合が業種情報のみを使う場合を上回った．その差は，5.8 ポイントであった．

素性作成の際に重み付けをしたときとしなかった場合を比較する．まず，業種情報とトップページの両方を使う場合について述べる．表 4.9 において，Pro-BT/NB と Pro-BT-W/NB，もしくは Pro-BT/RF と Pro-BT-W/RF を比較すると，業種情報に重みをつけて素性を作成した場合が重みを用いない場合を上回った．その差は，ナイーブベイズのときは 0.3 ポイント，ランダムフォレストのときは 0.7 ポイントであった．次に，業種情報のみを使う場合について考察する．表 4.9 において，Pro-B/RF と Pro-B-WD/RF, Pro-B-WT/RF，Pro-B-WB/RF を比較すると，重み付けによって Pro-B/RF から最も正解率の上昇幅が大きかった業種情報は事業説明で，正解率が 0.8 ポイント向上した．次に上昇幅が大きかったのは業種説明で 0.3 ポイント向上した．逆に正解率が下がったのは Description と Keywords に重みを付けた場合で，0.8 ポイント下がった．この結果より，業種分類において業種説明と事業説明は重要な情報であるが、Description と Keywords はそれほど重要ではないことが推測される．

業種カテゴリごとの正解率の傾向を述べる．全体的に，訓練データ数の多い業種カテゴリは正解率が高く，少ないカテゴリは正解率が低い傾向がみられた．表 4.4 に示したように，一番データ量の多い業種カテゴリは「2 食品」であるが，ほとんどの手法で一番高い正解率が得られた．

佐々木と新納の手法 [5] では，本研究のように業種情報は抽出せず，ウェブページ中の全ての単語を素性とし，ナイーブベイスモデルで分類器を学習している．ただし，トップページだけではなく，それから長さ 5 で到達できるサイト内ページからも素性を抽出している点が本実験のベースラインと異なる．彼らの手法の正解率は 41.8%であった．ただし，実験データが異なるので，本実験との単純な比較はできない．

上記の考察は提案手法の有効性を示してはいるが，人による判定との差は 20.9 ポイントと大きく，改善の余地が大きい．人が業種カテゴリを判定する際には，カテゴリをすぐに決定できる特定の単語や特徴 (URL 内の「.ac」,「会計」,「税理」,「商工会」など) を見つけて判定することが多かった．このような特徴的な単語を自動的に特定できれば，業種判定の正解率が向上すると考えられる．

表 4.6: 業種カテゴリ分類の正解率 (その 1)

業種カテゴリ	BL/NB	BL/RF	H
1 IT	0.083	0.354	0.615
2 食品	0.652	0.905	0.930
3 教育・受験	0.390	0.794	0.941
4 電機・エレクトロニクス	0.225	0.247	0.500
5 雇用	0.192	0.538	0.750
6 金融サービス	0.442	0.837	1.000
7 運輸・物流	0.590	0.927	0.957
8 農林・水産	0.023	0.091	0.571
9 財務・会計	0.208	0.938	1.000
10 製品・サービス（産業向け）	0.353	0.620	0.521
11 アパレル・装飾品	0.098	0.183	0.909
12 薬品・バイオテクノロジー	0.278	0.528	1.000
13 自動車	0.339	0.636	0.833
14 素材	0.053	0.035	0.417
15 環境・資源	0.111	0.048	0.667
16 投資	0.214	0.286	0.750
17 建設・土木	0.500	0.884	0.793
18 広告・マーケティング	0.077	0.631	0.667
19 小売	0.063	0.031	0.500
20 宿泊・飲食・接客	0.058	0.174	0.733
21 団体	0.243	0.871	0.273
22 印刷・出版	0.478	0.867	1.000
23 化学	0.178	0.222	0.400
24 企業向けサービス（法律など）	0.348	0.357	0.500
25 不動産	0.000	0.154	0.833
26 医療・ヘルスケア	0.091	0.284	0.600
27 ニュース・メディア	0.425	1.000	0.700
28 アート・娯楽	0.333	0.352	0.727

表 4.7: 業種カテゴリ分類の正解率 (その 2)

業種カテゴリ	Pro-BT-W /NB	Pro-BT-W /RF	Pro-BT /NB	Pro-BT /RF
1 IT	0.122	0.490	0.082	0.490
2 食品	0.758	0.928	0.753	0.930
3 教育・受験	0.358	0.841	0.409	0.826
4 電機・エレクトロニクス	0.244	0.360	0.267	0.371
5 雇用	0.259	0.385	0.185	0.385
6 金融サービス	0.483	0.849	0.529	0.849
7 運輸・物流	0.602	0.961	0.621	0.956
8 農林・水産	0.000	0.130	0.000	0.130
9 財務・会計	0.184	0.729	0.204	0.729
10 製品・サービス（産業向け）	0.352	0.486	0.345	0.490
11 アパレル・装飾品	0.133	0.317	0.120	0.293
12 薬品・バイオテクノロジー	0.216	0.694	0.243	0.667
13 自動車	0.328	0.619	0.336	0.610
14 素材	0.034	0.053	0.034	0.035
15 環境・資源	0.141	0.270	0.125	0.286
16 投資	0.200	0.214	0.200	0.286
17 建設・土木	0.436	0.783	0.462	0.788
18 広告・マーケティング	0.136	0.646	0.121	0.662
19 小売	0.030	0.031	0.000	0.000
20 宿泊・飲食・接客	0.126	0.218	0.103	0.207
21 団体	0.268	0.803	0.211	0.817
22 印刷・出版	0.495	0.891	0.516	0.880
23 化学	0.348	0.489	0.304	0.400
24 企業向けサービス（法律など）	0.371	0.252	0.345	0.261
25 不動産	0.000	0.192	0.000	0.154
26 医療・ヘルスケア	0.191	0.250	0.202	0.261
27 ニュース・メディア	0.364	0.943	0.409	0.943
28 アート・娯楽	0.377	0.386	0.359	0.336

表 4.8: 業種カテゴリ分類の正解率 (その 3)

業種カテゴリ	Pro-B /RF	Pro-B-WD /RF	Pro-B-WT /RF	Pro-B-WB /RF
1 IT	0.163	0.163	0.184	0.265
2 食品	0.907	0.912	0.907	0.912
3 教育・受験	0.609	0.630	0.609	0.609
4 電機・エレクトロニクス	0.360	0.348	0.337	0.348
5 雇用	0.385	0.385	0.423	0.385
6 金融サービス	0.767	0.779	0.756	0.779
7 運輸・物流	0.698	0.698	0.688	0.702
8 農林・水産	0.087	0.087	0.087	0.109
9 財務・会計	0.729	0.729	0.729	0.750
10 製品・サービス（産業向け）	0.360	0.377	0.384	0.380
11 アパレル・装飾品	0.341	0.354	0.329	0.305
12 薬品・バイオテクノロジー	0.333	0.389	0.306	0.389
13 自動車	0.542	0.551	0.534	0.542
14 素材	0.088	0.053	0.053	0.070
15 環境・資源	0.381	0.365	0.381	0.365
16 投資	0.429	0.286	0.357	0.357
17 建設・土木	0.549	0.535	0.531	0.566
18 広告・マーケティング	0.646	0.677	0.677	0.723
19 小売	0.313	0.281	0.344	0.281
20 宿泊・飲食・接客	0.437	0.425	0.437	0.368
21 団体	0.394	0.338	0.423	0.408
22 印刷・出版	0.630	0.630	0.609	0.641
23 化学	0.111	0.044	0.111	0.133
24 企業向けサービス（法律など）	0.461	0.443	0.461	0.478
25 不動産	0.462	0.462	0.615	0.538
26 医療・ヘルスケア	0.284	0.295	0.273	0.318
27 ニュース・メディア	0.705	0.716	0.716	0.693
28 アート・娯楽	0.227	0.218	0.218	0.223

表 4.9: 業種カテゴリ分類の正解率のマイクロ平均

手法	マイクロ平均
BL/NB	0.252
Pro-BT-W/NB	0.270
Pro-BT/NB	0.267
H	0.717

手法	マイクロ平均
BL/RF	0.493
Pro-BT-W/RF	0.508
Pro-BT/RF	0.501
Pro-B/RF	0.443
Pro-B-WD/RF	0.435
Pro-B-WT/RF	0.446
Pro-B-WB/RF	0.451

第5章　おわりに

5.1　まとめ

本論文では，機械学習を用いて企業ウェブページを業種に基づいてカテゴリに自動分類する手法を提案した．また，業種に関連する情報として Description と Keywords, 業種説明，事業説明という 3 種類を設定し，それらを自動抽出し，それから学習素性を抽出することで分類精度の向上を図った．Description と Keywords は HTML のマークアップによって正確に抽出できることを確認した．業種説明の抽出の際には，それが記載されているウェブページをルールベースの手法で検出し，そのウェブページ内の業種説明を抽出するという，２段階の方法を用いることで F 値が 0.97 という高い精度で抽出できた．事業説明の抽出の際にも，それが記載されているウェブページをルールベースで検出することで，F 値が 0.95 という高い精度で抽出できた．

企業のウェブサイトを業種に分類するにあたり，まず業種カテゴリのセットを設計した．様々なウェブサイトがディレクトリ構造で整理されている ODP というウェブサイトを用いて構築した．ODP のディレクトリの「ビジネス」以下に属するカテゴリの集合に対し，不要なカテゴリを削除したり，カテゴリに属するウェブサイトの数ができるだけ均等になるようにカテゴリを統廃合することで，最終的に 28 個の業種カテゴリを定義した．

Bag-of-words を素性とした機械学習によって企業ウェブページを業種カテゴリに分類するモデルを学習した．機械学習アルゴリズムとしてランダムフォレストとナイーブベイズを用いた．学習素性とする自立語は，自動抽出した業種情報ならびに企業ウェブサイトトップページから抽出する．素性の重みは出現頻度によって決めた．その際，業種情報における出現回数に高い重みを与えた．

評価実験では，１１種類の手法を比較した．自動抽出した業種情報から得られた素性に高い重みを与える手法や与えない手法，業種情報とトップページの両方から素性を抽出した手法，業種情報のみから素性を抽出して業種情報の種類ごとに重みを付与した手法などを比較した．実験の結果，提案手法による業種カテゴリの分類の正解率は 50.8%となり，業種情報を抽出せずにトップページ内の単語を素性としたモデルよりも正解率が 1.5 ポイント向上したことを確認した．また，3 種類の業種情報のうち，事業説明が最も業種カテゴリの分類に重要な情報であることが確認された．

5.2 今後の課題

最後に今後の課題を述べる．提案手法では，素性の重みを決定する際，業種情報での素性の出現回数に 4 倍の重みを与えていた．つまり，式 (3.1) における α を 4 に設定した．しかし，4 という値は直観によって決めたものであり，必ずしも最適値ではない可能性が高い．したがって，開発データを用意して α を最適化することで正解率の向上が期待できる．具体的には，α を変動させて開発データにおける業種分類の正解率を測り，最も高い正解率が得られた α を求める．さらに，3 種類の業種情報に対して異なる重みを与えることも検討したい．また，ランダムフォレストの学習パラメタはデフォルトのものを用いたため，これも開発データで最適化する必要がある．最終的には，1 章で述べたように，業種の判定結果を検索エンジンの検索結果に表示できるようなシステムを開発したい．

参考文献

[1] Sahar Changuel, Nicolas Labroche, and Bernadette Bouchon-Meunier. Automatic web pages author extraction. In *FQAS*, pp. 300–311. Springer, 2009.

[2] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing and Management*, Vol. 53, No. 5, pp. 1043 – 1061, 2017.

[3] Yoshikiyo Kato, Daisuke Kawahara, Kentaro Inui, Sadao Kurohashi, and Tomohide Shibata. Extracting the author of web pages. In *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*, WICOW '08, pp. 35–42, New York, NY, USA, 2008. ACM.

[4] 堀達也, 白井清昭. ブログページからのウェブサイト情報・作成者情報の抽出. 言語処理学会第 21 回年次大会, pp. 349–352, 2015.

[5] 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, pp. 352–355, 2006.

[6] 福島隆寛, 内海彰. Web ページの信頼性の自動推定. 知能と情報 : 日本知能情報ファジィ学会誌 : journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 19, No. 3, pp. 239–249, jun 2007.

[7] 百瀬亮, 宮崎林太郎, 渋木英潔, 森辰則. Web ページからの情報発信者の抽出におけるレイアウト情報の利用. 言語処理学会第 16 回年次大会, pp. 94–97, 2010.

付 録A　設定した業種カテゴリとODPカテゴリの対応関係

以下，ODPにおける「ビジネス」「ニュース/メディア」「各種資料/教育」の下位のカテゴリと本研究における業種カテゴリの対応を示す．これら3つの最上位カテゴリ名の前には◎をつけ，他のカテゴリと区別している．

<◎ビジネス>
- 《IT》{1}　(596)
  - <通信>(165)　【←通信/】
  - <インターネット>(289)
  - <ベンチャーキャピタル>(5)
  - <資格・スキル>(34)
  - <雇用・スタッフ>(24)
- 《食品》{2}　(4,477)
  - <イベント>(4)
  - <コンサルティング>(6)
  - <ニュースとメディア>(7)
  - <パン・菓子類>(542)
  - <乾燥食品>(11)
  - <包装・容器>(47)
  - <卵・乳製品>(213)
  - <卸売・輸出入>(106)
  - <団体>(3)
  - <小売>(87)
  - <漬物>(22)
  - <肉類・畜産加工品>(195)
  - <設備・器具>(80)
  - <調味料>(521)
  - <調理済食品>(122)
  - <農産物・加工品>(805)
  - <関係先・リンク集>(1)
  - <食用油・油脂>(31)
  - <飲料>(1,126)
  - <香料・添加物>(62)
  - <魚介類・水産加工品>(480)
- 《教育・受験》{3}(1,587)【新設】
  - <教育>(1,553)　【←◎各種資料/】
  - <資格・スキル>(34)　【←資格・スキル/】
- 《電機・エレクトロニクス》{4}　(1,005)
  - <コンピュータ>(65)
  - <ディスプレイ>(23)
  - <トランス>(19)
  - <光エレクトロニクス>(15)
  - <制御>(38)
  - <卸売・輸出入>(42)
  - <回路基板>(128)

```
|   ├ <団体> (2)
|   ├ <基板部品> (174)
|   ├ <家電> (54)
|   ├ <検査・測定> (22)
|   ├ <機構部品> (112)
|   ├ <計装・計測> (55)
|   ├ <設備・器具> (18)
|   ├ <送電・変電> (17)
|   ├ <配線・アクセサリ> (87)
|   └ <電源> (68)
├ <関係先・リンク集> (0)【X】
├ 《雇用》{5} (305)
|   ├ <人材斡旋> (252)
|   └ <求人情報> (44)
├ <電子商取引> (24)【X】
|   ├ <コンサルティング> (1)【X】
|   └ <システムとソフトウェア> (20)【X】
├ 《金融サービス》{6} (1,028)
|   ├ <キャッシュフロー> (14)
|   ├ <クレジットカード> (115)
|   ├ <サービサー> (33)
|   ├ <システムとソフトウェア> (13)
|   ├ <ベンチャーキャピタル> (25)
|   ├ <ポータル> (4)
|   ├ <リース> (42)
|   ├ <企業向け融資> (19)
|   ├ <保険> (114)
|   ├ <信用保証> (6)
|   ├ <信用組合> (73)
|   ├ <信用金庫> (221)
|   ├ <労働金庫> (15)
|   ├ <団体> (2)
|   ├ <外国為替> (16)
|   ├ <消費者金融> (54)
|   ├ <銀行> (160)
|   ├ <雇用・スタッフ> (10)
|   └ <電子決済> (59)
├ 《運輸・物流》{7} (2,421)
|   ├ <コンサルティング> (0)
|   ├ <システムとソフトウェア> (2)
|   ├ <ニュースとメディア> (0)
|   ├ <交通インフラ> (9)
|   ├ <団体> (0)
|   ├ <技術・設計> (0)
|   ├ <旅客輸送> (9)
|   ├ <海運> (701)
|   ├ <航空> (215)
|   ├ <貨物輸送> (234)
|   ├ <資格・スキル> (0)
|   ├ <輸送機械> (10)
|   ├ <道路輸送> (981)
|   └ <鉄道> (260)
├ 《農林・水産》{8} (561)
```

```
|   ├ <ニュースとメディア> (2)
|   ├ <団体> (12)
|   ├ <林業> (56)
|   ├ <漁業> (73)
|   ├ <設備・器具> (11)
|   ├ <農業> (406)
|   └ <関係先・リンク集> (0)
├ <通信> (165)【X】
|   ├ <携帯電話・PHS> (73)      【→IT/】
|   └ <設備・器具> (48)【→IT/】
├ 《財務・会計》{9}  (598)
|   ├ <監査・公認会計士> (137)
|   └ <税務・記帳> (455)
├ <資格・スキル> (34)【X】
|   ├ <社会保険労務士> (3)      【→教育・受験/】
|   └ <関係先・リンク集> (5)    【→教育・受験/】
├ <起業・SOHO> (14)【X】
|   ├ <独立・起業> (2)【X】
|   ├ <資格・スキル> (8)【X】
|   └ <SOHO> (14)【X】
├ 《製品・サービス（産業向け)》{10}  (3,416)
|   ├ <光学機器> (33)
|   ├ <包装・容器> (356)
|   ├ <卸売・輸出入> (1)
|   ├ <成形・加工> (1,316)
|   ├ <技術・設計> (75)
|   ├ <機械・器具> (417)
|   └ <産業資材> (1,212)
├ 《アパレル・装飾品》{11}  (969)【新設】
|   ├ <服飾・アパレル> (666)      【←製品・サービス（一般消費者向け）/】
|   ├ <かばん・スーツケース> (96)【←製品・サービス（一般消費者向け）/】
|   ├ <宝飾・貴金属> (147)        【←製品・サービス（一般消費者向け）/】
|   ├ <時計> (60)                 【←製品・サービス（一般消費者向け）/】
├ 《薬品・バイオテクノロジー》{12}  (421)
|   ├ <ベンチャーキャピタル> (5)
|   ├ <団体> (5)
|   ├ <薬品> (421)
|   └ <雇用・スタッフ> (10)
├ <製品・サービス（一般消費者向け）> (3,202)【X】
|   ├ <おもちゃ・遊具> (91)【X】
|   ├ <かばん・スーツケース> (96)【→アパレル・装飾品/】
|   ├ <オフィス・文房具> (180)【X】
|   ├ <システムとソフトウェア> (0)
|   ├ <スポーツ用品> (339)       【→アート・娯楽/】
|   ├ <ニュースとメディア> (0)
|   ├ <ペット・動物> (134)【X】
|   ├ <健康・美容> (506)         【→医療・ヘルスケア/】
|   ├ <冠婚葬祭> (187)【X】
|   ├ <卸売・輸出入> (7)【X】
|   ├ <団体> (0)
|   ├ <宗教・儀典> (107)【X】
|   ├ <宝飾・貴金属> (147)       【→アパレル・装飾品/】
|   ├ <家庭・園芸> (545)         【→アート・娯楽/】
```

```
|   ├ <家電・カメラ> (28)          【→電機・エレクトロニクス/】
|   ├ <小売> (5)【→小売/】
|   ├ <探偵・調査> (5)【X】
|   ├ <教育産業> (60)              【→教育・受験/】
|   ├ <時計> (60)                  【→アパレル・装飾品/】
|   ├ <服飾・アパレル> (666)       【→アパレル・装飾品/】
|   ├ <育児・子供> (28)【X】
|   ├ <資格・スキル> (1)【X】
|   └ <雇用・スタッフ> (0)
├ 《自動車》{13} (1,387)
|   ├ <イベント> (1)
|   ├ <オートバイ> (43)
|   ├ <キャンピングカー> (18)
|   ├ <コンサルティング> (6)
|   ├ <システムとソフトウェア> (3)
|   ├ <トラックとバス> (56)
|   ├ <ニュースとメディア> (5)
|   ├ <パーツとアクセサリ> (745)
|   ├ <ビンテージカー> (28)
|   ├ <モータースポーツ> (44)
|   ├ <リース> (23)
|   ├ <一般乗用車> (126)
|   ├ <中古売買> (0)
|   ├ <作業車・特殊車両> (39)
|   ├ <卸売・輸出入> (1)
|   ├ <団体> (8)
|   ├ <塗装・板金・カーケア> (18)
|   ├ <小売> (0)
|   ├ <廃車・解体> (14)
|   ├ <整備・車検> (55)
|   ├ <消耗品> (10)
|   ├ <福祉車両> (21)
|   ├ <設備・機材> (2)
|   ├ <設計・デザイン> (12)
|   ├ <資格・スキル> (0)
|   ├ <運転免許> (89)
|   ├ <関係先・リンク集> (0)
|   ├ <雇用・スタッフ> (2)
|   ├ <電気自動車・エコカー> (10)
|   └ <霊柩車> (8)
├ <航空宇宙・防衛> (76)【X】
|   ├ <宇宙> (76)【X】
|   └ <航空機> (16)【X】
├ 《素材》{14} (440)
|   ├ <カーボン> (24)
|   ├ <ガラス・水晶> (29)
|   ├ <卸売・輸出入> (0)
|   ├ <団体> (0)
|   ├ <研磨材> (16)
|   ├ <紙・パルプ> (146)
|   ├ <金属> (183)
|   ├ <繊維・布> (241)             【←繊維・布/】
|   └ <陶器・セラミックス> (18)
```

```
├ 《経営・管理》（25）【X】
│  ├ <コンサルティング>（16）　【→企業向けサービス/】
│  └ <ニュースとメディア>（2）【X】
├ 《繊維・布》（241）【X】
│  ├ <ニュースとメディア>（1）　【→素材/】
│  ├ <不織布>（28）　【→素材/】
│  ├ <卸売・輸出入>（24）　【→素材/】
│  ├ <団体>（6）　【→素材/】
│  ├ <皮革>（18）　【→素材/】
│  ├ <糸・紡績>（32）　【→素材/】
│  └ <設備・器具>（45）　【→素材/】
├ 《環境・資源》{15}（713）
│  ├ <エネルギー>（298）
│  ├ <団体>（0）
│  ├ <廃棄物処理>（284）
│  ├ <環境>（713）
│  ├ <設備・器具>（0）
│  └ <資源>（713）
├ 《投資》{16}（175）
│  ├ <商品先物>（30）
│  ├ <対企業投資>（43）
│  └ <証券>（88）
├ 《建設・土木》{17}（2,764）
│  ├ <コンサルティング>（25）
│  ├ <システムとソフトウェア>（39）
│  ├ <デザイン・設計>（77）
│  ├ <ニュースとメディア>（24）
│  ├ <一般建築・施設>（1,299）
│  ├ <内外装・設備>（491）
│  ├ <団体>（5）
│  ├ <土木>（2,764）
│  ├ <工場・プラント>（13）
│  ├ <構造・基礎>（52）
│  ├ <機材・器具>（92）
│  ├ <測量・調査>（46）
│  ├ <社寺・宗教建築>（91）
│  ├ <空間・景観>（73）
│  ├ <立体駐車場>（19）
│  ├ <素材・消耗品>（208）
│  ├ <解体・移動>（29）
│  ├ <資格・スキル>（6）
│  ├ <関係先・リンク集>（5）
│  └ <雇用・スタッフ>（10）
├ 《広告・マーケティング》{18}（740）
│  ├ <インターネット>（66）
│  ├ <ダイレクトマーケティング>（60）
│  ├ <テレマーケティング>（52）
│  ├ <ニュースとメディア>（4）
│  ├ <マーケティングリサーチ>（135）
│  ├ <団体>（2）
│  ├ <広告>（299）
│  ├ <広報・IR>（41）
│  └ <雇用・スタッフ>（4）
```

```
├ 《小売》{19} (375)
│  ├ <小売> (5)                    【←製品・サービス（一般消費者向け）/】
│  ├ <コンサルティング> (11)
│  ├ <コンビニエンスストア> (23)
│  ├ <システムとソフトウェア> (17)
│  ├ <ショッピングセンター> (16)
│  ├ <ディスカウントストア> (16)
│  ├ <デパート> (74)
│  ├ <ドラッグストア> (78)
│  ├ <ニュースとメディア> (2)
│  ├ <団体> (5)
│  ├ <総合スーパー> (21)
│  ├ <設備・器具> (16)
│  ├ <資材・消耗品> (1)
│  ├ <通信販売> (25)
│  ├ <関係先・リンク集> (1)
│  ├ <雇用・スタッフ> (5)
│  └ <雑貨・ホームセンター> (46)
├ 《宿泊・飲食・接客》{20} (972)
│  ├ <イベント> (2)
│  ├ <カラオケ> (14)
│  ├ <コンサルティング> (4)
│  ├ <システムとソフトウェア> (2)
│  ├ <ホテル・旅館> (78)
│  ├ <団体> (2)
│  ├ <外食・レストラン> (464)
│  ├ <宴会場・結婚式場> (17)
│  ├ <旅行・観光> (21)
│  ├ <温浴・リラクセーション> (34)
│  ├ <給食・厨房サービス> (76)
│  ├ <複合カフェ> (18)
│  ├ <設備・器具> (41)
│  ├ <資材・消耗品> (54)
│  ├ <資格・スキル> (83)
│  ├ <雇用・スタッフ> (5)
│  ├ <食事の宅配> (32)
│  └ <麻雀> (12)
├ 《団体》{21} (870)
│  ├ <中小企業団体中央会> (48)
│  ├ <商工会> (299)
│  ├ <商工会議所> (474)
│  ├ <日本経済団体連合会> (4)
│  ├ <民主商工会> (34)
│  └ <経済同友会> (6)
├ <国際商取引> (78)【X】
│  ├ <コンサルティング> (4)【X】
│  ├ <システムとソフトウェア> (5)【X】
│  ├ <地域別> (34)【X】
│  └ <産業・商品別> (0)
├ 《印刷・出版》{22} (1,089)
│  ├ <出版> (1,089)
│  ├ <印刷> (1,089)
│  ├ <団体> (3)
```

```
|    ├ <製本> (27)
|    └ <設備・器具> (7)
├ <卸売> (0) 【X】
├ 《化学》{23} (519)
|    ├ <コーティング・接着> (122)
|    ├ <ニュースとメディア> (4)
|    ├ <ポリマー> (188)
|    ├ <卸売・輸出入> (26)
|    ├ <団体> (5)
|    ├ <産業用ガス> (12)
|    └ <顔料・着色材料> (44)
├ <企業と社会> (5) 【X】
├ 《企業向けサービス》{24} (1,449)
|    ├ <アウトソーシング> (21)
|    ├ <コミュニケーション> (200)
|    ├ <コンサルティング> (22)
|    ├ <デザイン> (196)
|    ├ <事務・営業> (1)
|    ├ <団体> (0)
|    ├ <文書・データ> (10)
|    ├ <施設・オフィス> (30)
|    ├ <法律・手続> (622)
|    ├ <規格・認証> (21)
|    └ <防災・セキュリティ> (313)
├ 《不動産》{25} (284)
|    ├ <システムとソフトウェア> (20)
|    ├ <不動産仲介> (22)
|    ├ <住宅> (83)
|    ├ <投資> (27)
|    ├ <賃貸ビルとオフィス> (15)
|    ├ <資格・スキル> (10)
|    ├ <鑑定> (42)
|    └ <駐車場> (14)
├ <ビジネスチャンス> (16) 【X】
├ 《医療・ヘルスケア》{26} (1,014)
|    ├ <健康・美容> (506)          【←製品・サービス（一般消費者向け）/】
|    ├ <システムとソフトウェア> (14)
|    ├ <介護> (40)
|    ├ <検査・診断> (18)
|    ├ <歯科> (63)
|    ├ <眼科> (23)
|    ├ <設備・器具> (108)
|    ├ <資格・スキル> (0)
|    ├ <障害> (159)
|    └ <雇用・スタッフ> (41)
├ <イベント> (4) 【X】
├ 《ニュースとメディア》{27} (949)
|    ├ <オンラインメディア, ラジオ, 新聞, 雑誌, テレビ, 放送, 通信社> (928)
|    |                                  【←◎ニュース/】
|    └ <雑誌> (7)
└ 《アート・娯楽》{28} (2,557)
     ├ <アミューズメント> (263)
     ├ <スポーツ> (49)
```

├ <ビジュアルアート>（185）
├ <映像・パフォーミングアート>（1,169）
├ <スポーツ用品>（339）　　【←製品・サービス（一般消費者向け）/】
├ <家庭・園芸>（545）　　【←製品・サービス（一般消費者向け）/】
├ <資格・スキル>（5）
└ <雇用・スタッフ>（2）