企業ウェブページからの業種情報の抽出と分類

安道 健一郎　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s1610224,kshirai}@jaist.ac.jp

1 はじめに

ウェブ上には様々な情報が存在するが，中には誤った情報も存在する．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．この判断の助けになるのがウェブサイトの作成者情報である．例えば，法律に関することを調べる際，法律家や法律事務所のサイトで発信されている情報の方が，他の一般のサイトに比べて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 [1]．業種情報とは，企業が展開する事業の内容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．図 1 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，小売よりも電機の企業のウェブサイトを調べた方がよさそうといえる．

2 関連研究

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情報発信者名を抽出する手法を提案した [2]．この手法は，多様なウェブページから情報発信者名を抽出できる．しかし，人名以外は抽出できないことと，

図 1: 本研究が想定する検索エンジンの出力

事前に辞書を用意しておく必要があるという問題点がある．加藤らは，ウェブページの情報発信者情報を抽出するためのサブタスクとして情報発信者名を抽出した [5]．この手法は事前に辞書を用意しなくても情報発信者名が抽出できるという利点がある．しかし，情報発信者名以外で情報の信頼性の判定に有用な情報，例えば発信者のプロフィールなどは抽出されない．堀らは，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 [3]．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．

ウェブページの信頼性を評価する研究も行われている．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 [4]．また，企業のウェブページを業種別に分類する試みとして，企業のトップページならびにそこから深さ 5 までの下位ページに含まれる名詞，動詞，形容詞を素性とし，ナイーブベイズモデルを学習する研究がある [6]．これに対し，本研究では，ウェブページ全体から素性を抽出するのではなく，まず企業の業種情報を抽出し，それから機械学習の素性を抽出する点に特徴がある．

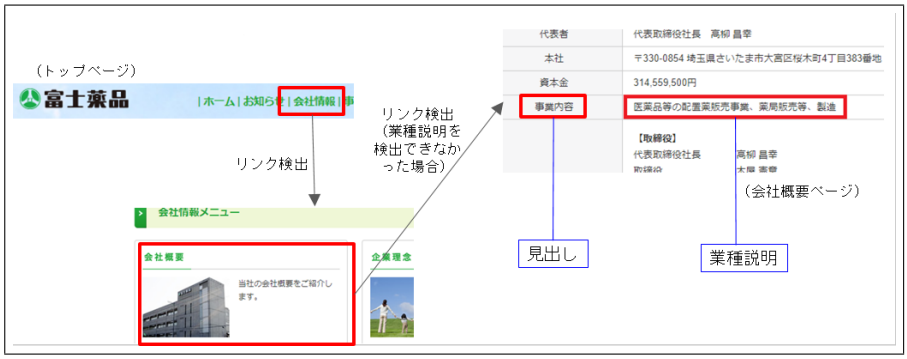

図 2: 業種説明の抽出例

図 3: 事業説明の例

3 提案手法

1 節で述べたように，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分類する．本研究では，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．

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 2 に示す[1]．

事業説明 企業の事業内容を説明したテキスト．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 に示す．

図 4 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「事業紹介ページ」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仮定する．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．

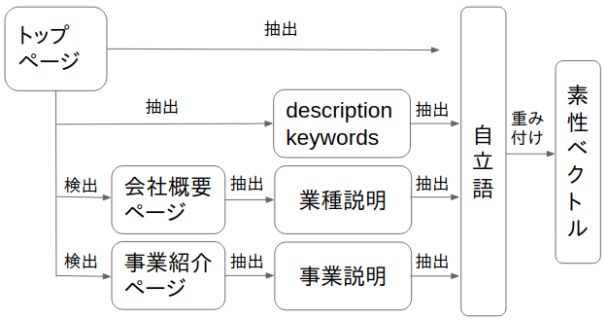

図 4: 提案手法の概要 (素性の抽出)

[1] この図は業種説明だけでなくこれを抽出する処理の流れを示している．詳細は 3.1.2 で述べる．

3.1 業種情報の抽出

3.1.1 Description と 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．

3.1.2 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 (Kw-Ta) 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．例を以下に示す．

  Kw-TA: 会社概要, 会社案内, 会社情報

- **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．それぞれの例を以下に示す．

  Kw-Tb: について, こちら, 詳細, 特色

  Kw-Tc: about, company, profile, corporate

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．その例を以下に示す．

Kw-Td: 事業内容, 業務内容, 営業内容, 主な事業

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 1 に示す条件にしたがって抽出する．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 1: 業種説明を含む HTML タグの抽出条件

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

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繰り返す．

図 2 の例では，富士薬品のトップページから「会社情報」というリンクを検出し，リンク先のページからが業種説明が取り出せなかったため,「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...」というテキストを業種説明として抽出する．

3.1.3 事業説明の抽出

まず，事業の内容を紹介するページを事業紹介ページと定義し，トップページから事業紹介ページへのリンクを検出する．リンクの検出は以下の 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．

- **Rule-B1**　事業紹介を示唆するキーワード (Kw-Ba) を含む ⟨a⟩ タグを検出する．用意した Kw-Ba の数は 34 である．例を以下に示す．

  Kw-Ba: 事業内容, 業務内容, 営業案内

- **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．それぞれの例を以下に示す．

  Kw-Bb: 事業, 業務

  Kw-Bc: business, project

次に，リンク先の事業紹介ページの HTML ファイルを取得する．広告やメニューなど，事業紹介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 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 [5] を用いて検出する．

3.2 業種カテゴリによる分類

業種カテゴリをあらかじめ設定する．本研究では，ウェブディレクトリーサービスの一つである Open Directory Project(ODP) の日本語サイト[2]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．その一覧を表 2 に示す．

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．訓練データから前項に示した手法で学習素性 (自立語) を抽出し，素性ベクトルを作成する．素性ベクトルの値は単語頻度とする．ただし，自立語が業種情報 (Description,Keywords, 業種説明, 事業説明) に出現したときの出現頻度は 4 回とカウントする．業種情報から抽出した素性は業種の種類をよく表すものと考えられるため，その出現頻度に 4 倍の重みを与える．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．

表 2: 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

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

4 評価実験

実験データとしてODPから獲得した企業ウェブページの集合を用いた.「ビジネス」「ニュース/メディア」「各種資料/教育」の3つのODPカテゴリおよびその下位カテゴリに登録されている企業ウェブページを収集した．各企業ウェブページの業種カテゴリ (表2) はそれが属するODPカテゴリから決定する．ODPでは，ひとつのウェブページが複数のカテゴリに属することがあるが，最も主要なカテゴリ以外はそれへのリンクという形でカテゴリに登録されている．今回の実験では，一つの企業ページは最も主要なカテゴリのみに属するものとし，正解として与える業種カテゴリは常に1つとした．取得した企業ウェブページの合計は29,364であった．このデータを訓練データ (90%) とテストデータ (10%) に分割し，業種カテゴリ分類の正解率を測った.

実験結果を表3に示す.「ベースライン」は企業のトップページから素性を抽出する手法,「提案手法」は業種情報から抽出した素性も使用する手法である．後者は業種情報での出現頻度に高い重みを与える手法 (重み付けあり) と与えない手法 (重み付けなし) を比較する.「人による判定」は，データセットからランダムに300件の企業ウェブページを抽出し，その業種カテゴリを人手で判定したときの結果である.

表 3: 業種カテゴリ分類の正解率

	NB	RF
ベースライン	0.252	0.493
提案手法 (重み付けなし)	0.381	0.513
提案手法 (重み付けあり)	0.380	0.517
人による判定	0.694	

(NB=ナイーブベイズ，RF=ランダムフォレスト)

機械学習アルゴリズムを比較すると，ランダムフォレストはナイーブベイズを大きく上回った．ランダムフォレストを用いたとき，提案手法はベースラインをわずかに上回った．また，業種情報から素性を抽出したときにその頻度に高い重みを与えることで正解率が高くなることが確認された．佐々木らの手法 [6] 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て同様に扱い，ナイーブベイズモデルで分類器を学習している．ただし，トップページから長さ5で到達できるページから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0.418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.

5 おわりに

本論文では，企業のウェブページから企業の業種情報を自動抽出し，またその結果を基に企業を28種の業種カテゴリに自動分類する手法を提案した．業種情報はルールベースの手法によって抽出し，業種カテゴリは Bag-of-words を素性とした機械学習によって分類した．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51.7%となり，業種情報を抽出せずにページ内の単語を素性としたモデルよりも正解率が2.4%向上したことを確認した.

最後に今後の課題を述べる．素性ベクトルの値を設定する際に業種情報における出現頻度に与える重み (4倍) は人手で設定していたが，開発データを用意して最適化することで正解率の向上が期待できる．3種類の業種情報に対して異なる重みを設定す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節で述べたように，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アドオンを開発したい.

参考文献

[1] 安道健一郎. 企業ウェブページからの業種情報の抽出と分類. 修士論文, 北陸先端科学技術大学院大学, 3 2018.

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

[3] 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.

[4] 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.

[5] Yoshikiyo Kato, Daisuke Kawahara, Kentaro Inui, Sadao Kurohashi, and Tomohide Shibata. Extracting the author of web pages. In *Proceedings of WICOW*, pp. 35–42, 2008.

[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, pp. 352–355, 2006.