6W-01

企業 Web サイトからの事業情報の抽出と業種別の類似事業推定

松浦 遼† 藤井 章博‡

法政大学大学院理工学研究科† 法政大学理工学部‡

1. はじめに

近年， Web サイトの数は年々増加しており，企業の多くは Web サイトで自社の情報を発信している．また，現在の検索エンジンは一定数の単語を用いて検索を行うのが一般的である．そのため単語レベルの検索は行えるが，文書全体の特徴量を用いた検索を行うことはできない．本研究では文書全体の特徴量を用いた文書検索の一例として，企業が Web サイトで公開している事業情報についてのテキストを用いた．

本研究は企業の Web サイトより事業情報文書を抽出し，その文書と類似した他企業の事業情報文書を推定することによる，文書全体の特徴量を用いた類似事業文書の検索を目的とする．

2. 事業情報の抽出

本研究において事業情報とは企業が Web サイト上で事業を示唆するワードを用いて紹介しているテキストや，そのリンク先のテキストとする．

事業情報の抽出は以下の手順で行う．

① トップページから事業情報ページへのリンク抽出
② 事業情報ページに記載のあるリンクをトップページと比較，抽出
③ 事業情報ページ及び，抽出したリンク先ページのテキストをトップページと比較，抽出

2-1. 事業情報リンクの抽出

事業情報ページへのリンク抽出は 3 つの手法を用いて行う．会社情報ページへのリンク抽出も同様にして行う．また<a>タグとは html において記述される href 属性にハイパーリンクを作成できるアンカー要素である．

(i). 事業情報を示唆する単語を含む<a>タグ抽出

<a>タグ内に事業を示唆する単語を含んだ場合，その<a>タグの href 属性を抽出した．

例：”事業情報”

(ii). 誘導を示唆する単語を含み，事業情報を示唆する単語をリンク先 URL に含む<a>タグ抽出

<a>タグ内に誘導を示唆する単語を含み，href 属性に事業を示唆する単語を含んだ場合，href 属性を抽出した．

例：”こちら”

Extraction of business information from company websites and estimation of similar businesses by industry

† Ryo Matsuura, Hosei University

‡ Akihiro Fujii, Hosei University

(iii). 事業情報を示唆する単語を含んだ alt 属性の直前の<a>タグ抽出

alt 属性が事業を示唆する単語が含まれた場合，その直前の href 属性を抽出した．alt 属性とは画像に対して付与できる代替テキストである．

例：<img src=....jpg” alt=”事業情報”>

2-2. トップページとの比較

企業 Web サイトの特徴として同一サイト内の Web ページは同じレイアウトで記載されており，全てのページで記載がある箇所とそのページにのみ記載がある箇所がある．本研究では事業情報ページにおいて，トップページと比較しそのページにのみ記載がある文書を抽出対象とした．リンク抽出では<a>タグ全体を，テキスト抽出ではタグで分割されたテキストを一つの文章とした．

3. 類似事業情報の推定

本研究では類似事業推定を 2 つの工程を経て行う．まず業種を分類する学習モデルを作成し，テストデータに対して業種分類を行う．次に業種分類したテストデータの一部を用いて業種ごとの分散表現モデルを獲得し，分散表現モデルを用いて分散表現モデルの学習に用いていない業種分類後テストデータを対象としたコサイン類似度より文書類似度を求める．

3-1. 業種分類モデル

業種を分類する学習モデルの作成には，BERT[1]を用いた．BERT は膨大なテキストを用いて pre-training し，各タスクに合わせて fine-tuning を行う学習モデルである．また，本研究では東北大学乾研究室の事前学習モデルを用いた．

3-1-1. 業種と教師データ

本研究では財務省が指定している日本標準産業分類の大分類を業種分類ラベルとして用いた．

また教師データのラベルとして科学技術・学術政策研究所の NISTEP 企業名辞書 2020 を用いた．

3-1-2. 入力 token 数の拡張

Transformer モデルは文章の長さに対して二次関数的に必要メモリ量が変化するため，長い文章に対して必要メモリ量が急激に増加してしまう．そのため入力 token 数を制限することで必要メモリ量を削減できるが，情報量が欠損してしまいタスクの精度が低下してしまう．

抽出したテキストにおいて， 1099token までのテキス

トが全体のおよそ 80%を占めていた． 1000token の学習は難しいため文書を複数のウィンドウに分け学習，分類を行なった．一つのウィンドウサイズを 254token とし，200token ずつスライドした．ウィンドウ数は 5 とし，一つの文書につき最大で 1054token の入力を行なった．学習では各ウィンドウをその業種のテキストとみなして学習し，テストでは各ウィンドウの分類スコアを集約し業種の推測を行った．

テキストDのi番目のウィンドウP_iの分類スコアp_i^{cls}とすると，テキストdの分類スコアは

$$\{p_1^{cls}, \cdots, p_n^{cls}\}$$

と表せる．本研究では，先頭ウィンドウを文書全体のスコアとする手法(first)と各スコアp_i^{cls}を全スコアの合計値(sum)，事業jにおける各スコアの最大値(max)で集約を行った． D^{cls}をテキストDの集約した分類スコアとした各式を下記に示す．

first：$D^{cls} = p_1^{cls}$

sum：$D^{cls} = \sum_{i=1}^{n} p_i^{cls}$

max：$D^{cls}[j] = max(p_1^{cls}[j], \ldots, p_n^{cls}[j])$

3-2. 分散表現モデル

分散表現モデルの作成に fastText[2]を用いた．テスト文書に対して分散表現モデルを用いて文書ベクトルを取得し，他文書とのコサイン類似度より文書類似度を算出した．本研究では業種ごとの分散表現モデルと業種全体の分散表現モデルにおいて精度の差を検証するため，情報通信業モデルと業種全体モデルで比較を行った．

4. 実験

企業 Web サイトより抽出した全 16 業種 1698 社 22910 文書を対象に提案手法の実験を行った．

業種分類モデル作成において学習用データは各業種最大 1000 文書合計 9008 文書，テスト用データは各業種最大 200 文書合計 1943 文書で実験を行った．

単語分散表現モデル作成においては単独業種モデルの学習用データに情報通信業に関する 1172 文書，業種全体モデルの学習用データに 16 業種 10020 文書を用いた．各業種ごとの tf-idf 値を求めた結果，情報通信業において tf-idf 値の大きかった単語である，「アプリケーション」，「セキュリティ」，「クラウド」，「回線」の 4 種類をカテゴリーとして設定した．テスト用データにはそれぞれの単語が含まれ，それらに関する事業文書を各 10 文書，計 40 文書を対象とした．1 つの文書に対して，残りの 39 文書との類似度を算出し上位 9 文書に同じカテゴリーの文書がどれほど含まれているかで精度とした．

5. 結果・考察

5-1. 業種分類の結果

max を用いた手法が最も高い精度であった． 業種別の再現率適合率を計測した結果，製造業と卸売業，学術研究専門・技術サービス業が他の業種と比べて精度が低く，およそ 70%前後であった．販売と製造，研究など行なっている工程は違うが，扱っている製品，技術などが似ている企業があるためこのような結果になったと考えられる．

表 1 業種分類用データ(左：学習 右：テスト)

業種	文書数	業種	文書数
鉱業	95	不動産業	643
建設業	1000	学術研究	1000
製造業	1000	宿泊業	72
電気ガス	236	娯楽業	124
情報通信業	1000	教育	84
運輸業	716	医療福祉	202
卸売業	1000	サービス業	836
金融業	1000	合計	9008

業種	文書数	業種	文書数
鉱業	24	不動産業	161
建設業	200	学術研究	200
製造業	200	宿泊業	18
電気ガス	59	娯楽業	31
情報通信業	200	教育	21
運輸業	178	医療福祉	51
卸売業	200	サービス業	200
金融業	200	合計	1943

表 2 業種分類精度

	平均値	最大値
first	0.7890	0.8037
sum	0.7971	0.8153
max	**0.8091**	**0.8224**

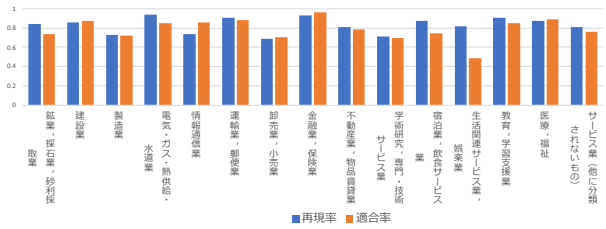

図 1 業種別の再現率と適合率

表 3 類似文書精度

	アプリケーション	セキュリティ	クラウド	回線	合計
情報通信業	0.5222	0.6556	0.7556	0.6333	**0.6417**
全業種	0.4556	0.6111	0.6222	0.6111	0.5750

5-2. 単語分散表現の結果

業種全体よりも業種を分けたほうがより良い結果となった．「アプリケーション」と「クラウド」においてより大きな精度差が得られた．これらは他 2 つと比べて tf-idf 値が大きかった．他の業種ではあまり出現する事業ではなかったため，より大きな精度の差が出たと考えられる．

6. まとめ

文書全体の特徴量を用いた検索の一例として，企業 Web ページより抽出した事業情報文書を元に，業種分類を経て類似事業を推定する手法の提案を行った．情報通信業に分類された文書において，全業種で作成したモデルより情報通信業のみで作成したモデルの方が精度の高い結果が得られた．今後は他業種での精度比較や，グラフデータベースを用いた類似事業検索システムの構築を行う．

7. 参考文献

[1]. Devlin, J., Chang, M., Lee, K. et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018).

[2]. Bojanowski, P., Grave, E., Joulin, A. et al.: Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics , Vol.5, pp.135-146 (2017).