ベンチャー企業の会社概要ページの自動抽出

Automatic Extraction of the Company Overview from Startup Company's Website

石垣 航大[*2] Kodai Ishigaki　柴田 有基[*1] Naoki Shibata　澤井 千春[*2] Chiharu Sawai　篠田 広人[*1] Hiroto Shinoda　石野 亜耶[*3] Aya Ishino　竹澤 寿幸[*1] Toshiyuki Takezawa

[*1] 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

[*2] 広島市立大学情報科学部
School of Information Sciences, Hiroshima City University

[*3] 広島経済大学メディアビジネス学部
Faculty of Media Business, Hiroshima University of Economics

Recently, many startup companies using new technologies such as AI have been founded. Investing in these startup companies from venture capital and angel investors is increasing. However, unlike listed companies, it is difficult for us to collect corporate information of startup companies because such information is often unorganized and undisclosed. To solve this problem, we are conducting research to collect information on startup companies from the Web automatically and comprehensively. As a first step, we propose a method for extraction of the company overview from startup company's website automatically. To confirm the effectiveness of our method, we conducted several experiments on extraction of the company overview. From the experimental results, we obtained precision of 0.794 and recall of 0.452.

1. はじめに

近年, AI などの新しい技術を使用したベンチャー企業が数多く創業されている. このような先進性や行動力に富むベンチャー企業に対する, ベンチャーキャピタルやエンジェル投資家からの投資も増加している. 投資を行う上で, どのようなベンチャー企業があり, どのような活動をしているのかといった企業情報は, 重要な判断材料である.

しかし, ベンチャー企業は, 上場企業とは異なり, 情報が整理され公開されていないことが多いため, 企業情報を収集するのは困難である. ベンチャー企業のデータベースとして, INITIAL (https://initial.inc/) などのデータベースがある. このようなデータベースには, 会社名, サービス説明概要文, 資金調達額などの情報が登録されているが, 公開範囲が限定的であることや, 人手で作成されているためコストがかかるといった問題点がある.

この問題を解決するために, 我々は, Web からベンチャー企業の情報を網羅的に自動で収集するための研究を行っている. その第一歩として, AI 関連のベンチャー企業のホームページから, 代表者名や資本金などの重要な情報がまとめて記載されている会社概要が掲載されているページ (以下, 会社概要ページと呼ぶ) を, 機械学習を利用して自動抽出する手法を提案する.

2. 関連研究

本研究と同様に, ベンチャー企業を対象にした研究を説明する. 上野山ら [上野山 2014] は, Web 上のベンチャー企業情報を用いて, ベンチャー企業の上場または事業売却 (Exit) を, 機械学習を用いて予測する手法を提案している. 機械学習には SVM を使用し, 特徴量には, Web で公開されている Crunch base という人材データベースから抽出した人材の転職履歴の情報を利用する. 上野山らの手法では, 人材の転職履歴情報を活用することで, 資金調達額や従業員数など, 社内の資源のみを特徴量として使用した手法と比較し, 高い精度で Exit 予測を行うことに成功している.

今井ら [今井 2015] は, 日本で今後有望なベンチャー企業の事業を予測する手法を提案している. 具体的には, 米国の企業群をクラスタリングし, その各クラスタを 1 つの事業分野ととらえ, それが今後日本で成功する事業かどうかを, 機械学習を用いて予測している. 機械学習には SVM を使用し, 特徴量には, 各クラスタ内の資金調達総額, 会社数の増加率などを使用している.

ベンチャー企業の情報は整理され公開されていないため, ベンチャー企業を対象にした研究は限定的であった. しかし, Web にベンチャー企業の情報が公開されるにつれ上野山らや今井らのような研究が行われるようになっている. 本研究の目的であるベンチャー企業の情報を網羅的に収集できれば, このような研究を支援することが可能になる.

安道ら [安道 2018] は, 企業のホームページから, 会社概要ページを抽出するための手法を提案している. 安道らは, ホームページのトップページにリンクされているページから, HTML タグを抽出し, 会社概要などの単語があれば, そのページを会社概要ページと判定するというルールベースの手法を用いている. 本研究では, 会社概要ページを漏れなく抽出するため, 各ページに含まれる単語と, ページの URL 文字列を特徴量として機械学習を用いる手法を提案する.

3. ベンチャー企業の会社概要ページの自動抽出手法

本研究では, ベンチャー企業の会社概要ページを自動抽出する手法を提案する. 提案手法の流れは, 以下のステップに分かれる. Step1 については 3.1 節, Step2 については 3.2 節で説明する.

Step1. ベンチャー企業リストの作成
Step2. 会社概要ページの自動抽出

連絡先:石野 亜耶, 広島経済大学メディアビジネス学部, ay-ishino@hue.ac.jp

3.1 ベンチャー企業リストの作成

本節では，ベンチャー企業リストの作成方法について説明する．ベンチャー企業のリストを作成するための情報源として，安価でプレスリリースを配信できるニュースリリースサイトに着目する．本研究では，ニュースリリースサイトの一例として，PR TIMES (https://prtimes.jp/) を利用する．

まず，PR TIMES に掲載されている，キーワードに「ベンチャー」と「AI」が設定されているプレスリリースから，そのプレスリリースを配信している企業と，その企業のホームページの URL を収集する．その中から，業種が情報通信であり，未上場である企業を，AI 関連のベンチャー企業としてリストに登録する．次に，作成したリストの企業ホームページの URL を利用して，企業のホームページの全ページのデータを収集する．

3.2 会社概要ページの自動抽出

本節では，3.1 節で収集した AI 関連のベンチャー企業リストに登録されている企業のホームページから，会社概要ページを自動で抽出する手法について説明する．まず，企業のホームページには，大量のページが登録されている場合があるため，「会社」という単語が含まれているページを会社概要ページの候補として抽出する．会社概要ページには「会社」という単語が含まれることは，開発用データセットを用いて確認した．

次に，会社概要ページの候補となるページから，機械学習を用いて会社概要ページを自動で抽出する．機械学習の特徴量には，下記のページに含まれる名詞と，URL に含まれる文字列を利用する．機械学習にはサポートベクターマシン (SVM) を採用する．

- ページに含まれる名詞

会社概要ページには，「所在地」や「資本金」などの名詞が多く頻出する．そのため，各ページに含まれる名詞の頻度を特徴量として利用する．なお，名詞の抽出には MeCab を用いる．

- URL に含まれる文字列

会社概要ページの URL は，「company」や「about」など，会社概要ページであることを示す文字列が含まれている場合が多い．そのため，URL に含まれる文字列を特徴量として使用する．

4. 実験

3節で述べた提案手法の有効性を確認するため，実験を行った．

4.1 実験設定

実験に使用するデータ

PRTIMES から収集した AI 関連のベンチャー企業 730 件のホームページに対し，人手で会社概要ページの判定を行った結果を実験に使用する．

実験条件

表 1: 提案手法と比較手法の条件

条件		対象ページ	品詞	URL
提案手法		「会社」を含むページ	名詞	○
比較手法	(1)	全ページ	全品詞	×
	(2)	「会社」を含むページ	全品詞	×
	(3)	全ページ	名詞	×
	(4)	「会社」を含むページ	名詞	×
	(5)	－	－	○

提案手法と比較手法の条件を表 1 に示す．本研究では，全てのページを対象にした場合と，「会社」という単語を含むページのみを対象にした場合，全単語を用いた場合と名詞のみの単語を用いた場合，また，URL 文字列の特徴量のみを用いた場合をそれぞれ組み合わせ比較手法として実験する．また，正例に対して負例の数が非常に多いため，入力データに対してアンダーサンプリングを行い，学習に使用する正例と負例の割合を調節する．機械学習には SVM を用い，カーネル関数には RBF カーネルを採用する．評価尺度には，精度，再現率，F 値を用い，5 分割交差検定を行う．

4.2 実験結果

実験の結果を表 2 に示す．比較手法の中で最も F 値が高かったのは，「会社」を含むページを対象に，名詞を特徴量として使用した比較手法(4)であった．比較手法(4)と提案手法を比較すると，わずかではあるが F 値を 0.004 ポイント改善することができた．提案手法では，比較手法(4)の特徴量に，URL に含まれる文字列を追加した特徴量を使用している．よって，URL に含まれる文字列を使用する提案手法の有効性を確認できたといえる．

表 2: 実験結果

条件		精度	再現率	F 値
提案手法		0.794	0.452	0.576
比較手法	(1)	0.597	0.471	0.524
	(2)	0.62	0.459	0.524
	(3)	0.748	0.459	0.568
	(4)	0.795	0.445	0.572
	(5)	0.023	0.62	0.045

5. おわりに

本研究では，テキストと URL に含まれる文字列を用いて，ベンチャー企業の会社概要ページを自動抽出する手法を提案した．今後は，Word Embedding やニューラルネットワークなど深層学習ベースの手法を用いてさらなる精度の向上を目指す．

今後は，鶴田ら [鶴田 2009] の研究成果を利用して，会社概要ページから，「資本金」という属性名と，「100 万円」などの属性値の抽出や，平前ら [平前 2018] の研究成果を利用して，企業間の投資や提携関係を明らかにすることで，ベンチャー企業の情報を網羅的に収集していく予定である．

参考文献

[上野山 2014] 上野山勝也, 大澤昇平, 松尾豊: 人材の転職履歴情報を素性としたベンチャー企業の Exit 予測, 情報処理学会論文誌, Vol. 55, No. 10, pp. 2309-2317, 2014.

[今井 2015] 今井響, 大知正直, 松尾豊: 日米スタートアップのキーワードによるクラスタリングを用いた事業トレンド予測, 第 29 回人工知能学会全国大会, 2015.

[安道 2018] 安道健一郎, 白井清昭: 企業ウェブページからの業種情報の抽出と分類, 言語処理学会 第 24 回年次大会, 2018.

[鶴田 2019] 鶴田雅信, 関根聡, 増山繁: 企業の公式 Web サイトからの基本情報抽出, 第 23 回人工知能学会全国大会, 2019.

[平前 2018] 平前歩, 難波英嗣, 竹澤寿幸: 技術関連記事の分析に基づいた経営判断支援システムの構築, 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2018), 2018.