企業の公式Webサイトからの基本情報抽出

Corporate Profile Information Extraction from Web Sites

鶴田 雅信[*1] 関根 聡[*2] 増山 繁[*1]
Masanobu Tsuruta Satoshi Sekine Sigeru Masuyama

[*1]豊橋技術科学大学 Toyohashi University of Technology
[*2]ニューヨーク大学 New York University

We propose a method to extract corporate profile information of companies from their Web sites. Our method uses URLs of target companies' Web site, and the seed string "Corporate Profile (in Japanese)", and extracts information as the set of key-value attributes of corporate profile.

1. はじめに

現在，多くの企業が公式 Web サイトを開設し，情報を公開している．企業が公式 Web サイトに掲載している情報として，主として以下のようなものが存在する．

- 会社概要・沿革
- IR 情報
- 個別の製品・事業の詳細情報

これらの情報は，RDF など，そのデータの意味を機械が理解できるフォーマットで記述されていないことが多い．そのため，データマイニングなどの再利用のためには，人間が記述に対して意味のアノテーションを行うこと，もしくは，このような企業が提供する情報をアルゴリズムを用いて抽出し，再利用可能な形にすることが必要となる．これらの企業情報のうち，製品・事業の詳細情報などは一定のフォーマットで大量に記述されていることが多く，Web ラッパー [Kushmerick 00] などを用いた，アルゴリズムによる自動抽出が可能であると考えられる．しかしながら，会社概要・沿革，および，IR 情報などは 1 つの企業のサイトには 1 つしか用意されていないことが多いため，Web ラッパーなどの手法をそのまま適用することはできない．また，異なった企業サイト間において共通，かつ，機械可読なフォーマットが定義されていない．このような会社自体の基本的な情報を，本研究では「企業の基本情報」と呼ぶ．企業の基本情報を自動で抽出し，収集することができれば，投資家などの投資判断において有益だと考えられる．

本研究では，会社概要は，属性名と属性値の対で構成される，会社概要属性の集合であると定義する．例えば属性名としては,「代表者名」,「資本金」,「従業員数」などが存在し，その具体的なデータは属性値として扱われる．ここで，属性名は非常にバリエーションに富んでいることが特筆される．たとえば,「代表者名」と同義，もしくは非常に近い意味をもつ属性名として「代表取締役」,「代表取締役社長」,「代表取締役会長」などが存在する．さらに,「主要仕入先」などのように，一部の業界にのみ多く出現するような属性名も存在するため，これらを人手のみで全て収集することは困難である．したがって，Web サイト集合から統計的な手法を用いて会社概要属性の属性名，および，属性値を抽出する手法は，非常に有効であると考えられる．そこで，本研究では，企業の公式 Web サイト集合，および「会社概要」という文字列を用いて，企業の基本情報のうち，会社概要を会社概要属性の集合の形で自動的に抽出する手法を提案する．

連絡先: 鶴田雅信，豊橋技術科学大学電子・情報工学専攻，愛知県豊橋市天伯町雲雀ヶ丘 1-1，TEL: 0532-44-6867，FAX: 0532-44-6873，tsuruta@smlab.tutkie.tut.ac.jp

2. 手法

提案手法は,「抽出対象となる企業の公式 Web サイトのドメイン名集合」，および「『会社概要』という文字列」を入力とし，対象となる企業ごとの会社概要属性の集合を出力とする手法である．属性名のリストや，会社概要ページへの URL リストなどは必要としない．

提案手法は，大きく以下の 3 つの段階に分けられる．

Step 1 (2.1 節) 抽出対象となる企業の公式 Web サイト集合 S の部分集合である学習用 Web ページ集合 S_l から，会社概要が含まれるページに偏って出現する語を抽出する．

Step 2 (2.2 節) S から，属性名を抽出する．

Step 3 (2.3 節) S から，属性名を用いて，会社概要属性を抽出する．

また，Step 2，および，Step 3 において，クロールのコストを下げるために，リンク遷移先をナイーブベイズによって識別する手法についても 2.4 節で述べる．

2.1 会社概要が含まれるページに偏って出現する語の抽出

まず，抽出対象となる企業サイトのうち，少数のものの全体をクロールすることで会社概要が含まれるページに偏って出現する語の集合を求める．このような語は，すなわち会社概要属性にもよく出現する語であるという直観に基づいた手法である．

Step A-1 複数の企業の公式 Web サイトに含まれる全てのページをクロールし，収集したものを学習用 Web ページ集合 S_l とする．

Step A-2 S_l に含まれるすべてのページから,「会社概要」という文字列が含まれるリンクを抽出し，学習用リンク集合 $L_{positive}$ とする．また，$L_{positive}$ に含まれるリンクの遷移先の Web ページ集合を $D_{positive}$ とする．

Step A-3 $D_{positive}$ に含まれるページ d が属するサイトのトップページから，d への最短路を辿る．そのとき，$D_{positive}$ に含まれるどのページの場合においても経由することのない Web ページの集合を $D_{negative}$ とする．また， $D_{negative}$ に含まれる Web ページを遷移先とするリンク集合を $L_{negative}$ とする．

Step A-4 $D_{positive}$ に含まれる Web ページのうち，語 t を含むものの数を $df_{positive}(t)$，$D_{negative}$ に含まれる Web ページのうち，t を含むものの数を $df_{negative}(t)$ とする．[ここで語とは，HTML をパーズした木構造 におけるテキストノード，および alt，title 属性の値に含まれる形態素 *1 の 1-gram および 2-gram のことを指す．以下，「あるノードに含まれる語」は，この語のことを指す．]*2

Step A-5 $D_{positive}$ および $D_{negative}$，2 つの Web ページ集合における df 値の偏りをもとにした語 t のスコア $w(t) = df_{positive}(t) \times (1 - df_{negative}(t))$ を，$D_{positive}$ に含まれるすべての語において求める．ここで，$w(t) < \alpha$ である場合，$w(t) = 0$ とする．ここで α は定数.

Step A-5 で求めた語 t のスコアは，会社概要ページでのみ多く出現する語に高い重みを割り当てる．一方，会社概要ページ以外でも頻出する語については，低い重みを割り当てる．$w(t)$ の大きな語は，多くが属性名，および属性値の一部に含まれていると考えられる．

2.2 会社概要属性における属性名の抽出

会社概要ページの HTML をパーズした木構造において，子を 2 つのみ持つ部分木が，会社概要属性を含んでいる（条件 1）ことが多いという直観に基づき，会社概要属性における属性名を抽出する手法について解説する．例えば，図 1 のように，表，または適切な構造化がなされた HTML 文書において，属性名，属性値の両方を含む部分木は条件 1 に適合する．しかし，図 2 のように，適切な構造化がなされていない HTML 文書においては，条件 1 は適合しない．そのため，条件 1 を最終的な会社概要属性抽出にそのまま用いることはできない．しかしながら，条件 1，および 2.1 節の手法で求めた語の分布の知識を用いることで，属性名のみであれば高精度で抽出できると考えられる．

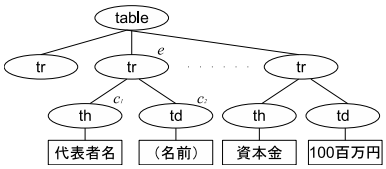

図 1: 子を 2 つのみ持ち，会社概要属性の属性名と属性値がそれぞれ格納された部分木

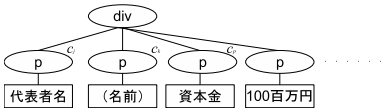

図 2: 適切な構造化がなされていない会社概要属性

Step B-1 KL_{cand} を属性名らしい場所に出現した文字列の集合， VL_{cand} を属性値らしい場所に出現した文字列の集合と定義する．両方の初期値は ϕ となる．

Step B-2 抽出対象となる企業サイト $S_i \in S$ に含まれる Web ページ d の HTML を木構造にパーズし，その木構造に含まれるノードの集合を D とする．

Step B-3 ノード $e \in D$ ，および，その子孫ノードから構成される部分木に含まれる語 $t_{i,i=1...n}$ のスコア $w(t_i)$ の平均値を e の属性要素スコア $W_{kv}(e)$ とし，これをすべてのノードに対して求める．この際，部分木における $t_{i,i=1...m}$ の出現頻度はすべて 1 とする．[この Step は，会社概要ページにのみよく出現する語を含むノードは，属性名，もしくは属性値を含んでいる可能性が高いという直観に基づく．]

Step B-4 すべてのノードに対して，e が直接の子として持つノード $c_{i,i=1...o}$ の会社概要属性要素スコア $W_{kv}(c_i)$ の平均値を e の会社概要属性スコア $W_a(e)$ として求める．[属性名，属性値を多く含んでいるノードは，会社概要属性そのものを含んでいる可能性が高いという直観に基づく．]

Step B-5 会社概要属性のリストである可能性の高い部分木のみを属性名抽出の対象とするため，すべての部分木から，会社概要属性リスト候補となる部分木を抽出する．直接の子 $c_{i,i=1...o}$ のうち，$W_a(c_i) > 0$ であったものの割合が 0.5 以上であったノード e の集合を会社概要属性リスト候補部分木の根の集合 $AL_{cand}(d)$ とする．[会社概要属性そのものを含んでいる子ノードを多く持つノードは，会社概要属性のリストである可能性が高く，属性名，さらに属性値を抽出する対象として重要であるという直観に基づく．その場合，会社概要属性を含んでいない子ノードを多く含むノードは，おそらくナビゲーションリンクなど，この後の処理においてノイズになるノードであると考えられる．]

Step B-6 $e \in AL_{cand}$ のうち，直接の子を c_1, c_2 の 2 つのみ持つノードすべてに対して，属性名の有無を推定するためのスコア $score_k(e) = W_{kv}(c_1) - W_{kv}(c_2)$ を求める．[図 1 のような形の部分木において，c_1 には属性名がよく含まれているという直観に基づく．] そして，$score_k(e) > 0$ であったノード e の集合を， 属性名抽出対象部分木の根の集合 $KL(d)$ とする．

Step B-7 Web ページ d の会社概要ページらしさを推定するためのスコア $score_{KL}(d) = \sum_{e \in KL(d)} score_k(e)$ を求める．

Step B-8 Step B-2 から Step B-7 までをサイト S_i に含まれるすべてのページに対して行う．その結果 $score_{KL}(d)$ が最大となった，もっとも会社概要ページらしい Web ページ d において，属性値，属性名らしい文字列を抽出する．$KL(d)$ に含まれるすべてのノード e において，その 1 番目の直接の子であるノード c_1 をテキストとして見た場合の文字列を $string(c_1)$ *3 とする．そして，S_i において $string(c_1)$ が属性名らしい場所に出現したことを示すフラグ $sf_k(string(c_1), S_i)$ を 1 とする．KL_{cand} に $string(c_1)$ が含まれていない場合，追加する．また， S_i において $string(c_2)$ が属性値らしい場所に出現したことを示すフラグ $sf_v(string(c_2), S_i)$ を 1 とする．

*1 形態素解析には MeCab 0.97 を用いた.

*2 以後，[ ] 内はその Step についてのコメントとする.

*3 ノード c_i の子孫ノードに含まれる形態素を HTML 文書における出現順に結合する．また，記号，時刻表記，数値のみであった場合削除した.

Step B-9 Step B-2 から B-8 をすべてのサイトに対して繰り返し行う.

Step B-10 KL_{cand} に含まれる全ての文字列 str に対して，属性名らしさのスコア を以下の式によって求め，スコアが β より大きなものの集合を，抽出対象 Web サイト集合 S における属性名集合 $Label(S)$ とする.

$$score_{label}(str) = \frac{\sum_{S_i \in S} sf_k(str, S_i)}{1 + \sum_{S_i \in S} sf_v(str, S_i)}. \quad (1)$$

2.3 会社概要属性の抽出

2.2 節で抽出した会社概要の属性名集合 $Label(S)$ を用いて，Web サイト S_i から，会社概要属性を属性名：属性値の対の形で抽出する.

Step C-1 出力する会社概要属性集合を $Attr(S_i) = \phi$ とする.

Step C-2 Step B-2 から B-5 を行い，会社概要属性リスト候補部分木の根の集合を $AL_{cand}(d)$ とする.

Step C-3 $AL_{cand}(d)$ に含まれるノード e が直接の子として持つノード $c_{i,i=1...p}$ に対して， $string(c_i) \in Label(S)$ であった場合属性名ノードとし，そうでなければ属性値候補ノードとする.

Step C-4 図 2 のような構造をしたページに対応するために，ある属性名ノードの右隣に位置するノードから，次の属性名ノードまでの間に位置するノードについて，それらをすべて属性値であると考え，属性値の抽出を行う．子ノード集合 $\{c_1, ..., c_p\}$ から，ある属性名ノード $c_{j,1 \leq j \leq p}$ から，次の属性名ノード $c_{k,j<k \leq p}$ までの間にあるすべての属性値候補ノード $c_{l,j<l<k}$ を抽出する．ここで抽出された属性値候補ノード部分集合 $\{c_{j+1}, ..., c_{k-1}\}$ が，属性名ノード c_j に対応する属性値となる．抽出された属性名:属性値の対の集合 $\{(c_j, \{c_{j+1}, ..., c_{k-1}\}), ..., (c_q, \{c_{q+1}, ..., c_p\})\}$ を出力となる会社概要属性集合 $Attr(S_i)$ に追加する.

Step C-5 Web サイト S_i に含まれるすべてのページに対して Step C-1 から Step C-5 を行い，S_i の会社概要属性集合 $Attr(S_i)$ を出力する.

2.4 リンク文字列を用いたナイーブベイズによる抽出対象ページの探索

遷移前にリンク先のページの内容を知ることができれば，クロールするページ数を飛躍的に減少させることができると考え，リンク文字列，およびリンク先の URL に含まれる文字列を特徴として用い，ナイーブベイズを利用することで抽出対象ページを探索する手法を提案する．2.2，2.3 節で述べた手法は，サイト内の全ての Web ページを抽出対象とする．しかし，会社概要が含まれるページは少なく，Web サイト全体をクロールするのに必要なコストは大きい．そのため，今回のような抽出対象とするデータが絞られているタスクの場合，クロールするページ数を減少させることで，抽出全体のコストを下げ，スピードを向上させることができると考えられる.

ナイーブベイズ識別器の学習について述べる．Step A-2 で求めた，「会社概要」という文字列が含まれるリンクの集合 $L_{positive}$ に含まれるリンク全てをそれぞれ以下の素性からなるベクトルとし，訓練データの正例とする.

- 子孫のノードに含まれる語
- リンク先の URL を Perl の正規表現によって \W で分割したもの.

素性の値には文書中における出現確率 tf を用い，またそれぞれのベクトルにおいて最大値を 1 とする正規化を行う．また，$L_{negative}$ に含まれるリンクをそれぞれ同様にベクトルとし，訓練データの負例とする．抽出対象ページの探索は次のように行う.

Step D-1 手法の対象となる Web ページ d を，対象となる企業サイト $S_i \in S$ のトップページとする．ここで，S は抽出対象となる企業サイト集合である．また，探索経路中に存在した Web ページのスコア $score_T$ の最大値 max_{score_T} を 0 に初期化する．さらに，探索打ち切りまでの回数 $rest$ を定数 γ に初期化する.

Step D-2 対象となる Web ページ d に対して，スコア $score_T(d)$ を算出する．スコアについては後述する.

Step D-3 $max_{score_T} \leq score_T(d)$ であれば，$max_{score_T} := score_T(d)$ とし，$rest := \beta$ とする．$max_{score_T} > score_T(d)$ であれば，$rest := rest - 1$.

Step D-4 $rest > 0$ であれば，Web ページ d に含まれるリンク全てに対して，学習済みの識別器を用いて識別を行う．その結果，正例である確率が最も高いリンクの遷移先である Web ページを d とし，Step D-2 に戻る.

この手法を用いることで，Web サイト全体を対象とする多くの手法を，適当な Web ページの探索と同時に適用する手法へと変更することができる．2.2 節で述べた会社概要属性における属性名の抽出手法においては，Step D-2 における $score_T(d) = score_{KL}(d)$ とする．2.3 節で述べた会社概要属性の抽出手法においては，$score_T(d)$ は Web ページ d において抽出された会社概要属性の個数とする.

3. 評価実験・考察

提案手法の評価のために，実験を行った．評価実験に使用する Web サイトは，2008 年 1 月の段階で株式市場に上場していた企業のうち，2500 社のものとする。実験では,「会社概要属性の抽出」において対象となった Web ページ集合のなかで，会社概要属性が存在したページのみに対し，抽出された会社概要属性の精度，また，属性名抽出の精度を人手で調査し，評価を行う．さらに，2.4 節で述べた探索手法の性能を評価するため，会社概要属性が存在したページを探索経路上で発見できたかどうかについて，成功率を人手で調査する．評価者は，工学系の学生 1 名とした．また，予備実験より，Step A-1 で使用する Web サイトの数を 100，Step A-5 における定数 $\alpha = 0.2$，Step B-11 における定数 $\beta = 3$， Step D-1 における定数 $\gamma = 2$ とした.

3.1 会社概要属性の抽出結果

会社概要属性が存在するページを探索経路上で発見できたサイトの中から，ランダムで 30 サイトを抽出し，それらのサイトから抽出された会社概要属性を属性名ごとに人手で調査した．その結果，抽出された属性値がすべて正解であった属性名は 352 個，ひとつでも不正解があった属性名は 64 個となり，精度は 0.846 となった．不正解であった会社概要属性の内訳を表 1 に示す．大半が属性名の誤取得によるミスであり，属性名抽出の精度，再現率を向上させることで，会社概要属性の抽出性能も向上すると考えられる.

表 1: 不正解であった会社概要属性の内訳

属性名の誤取得によるミス	41
過剰に広い範囲を属性値として取得	10
会社概要属性ではない部分に属性名が出現	9
取得対象ページを誤り，対象企業以外の情報を取得	6
属性名と属性値の分割に失敗	3
属性値と属性名，両方に出現しうる文字列	2

3.2 属性名の抽出結果

抽出された属性名は 235 個，そのうち誤って抽出された属性名は 22 個であり，精度は 0.906 となった．正しく抽出できた属性名を図 3 に示す．会社概要において自然な属性名が抽出できていると考えられる．一方，誤って抽出された属性名の全てを図 4 に示す．ホームページ，トップ，FAQ など，ページの構造における見出しによく使われる語，および属性名を誤って取得してきたものなどがあることがわかる．

創立 取締役 代表者 取締役常務執行役員 事業内容
社名 従業員 上場 発行済株式総数 取締役専務執行役員
連結従業員数 主要取引先銀行

図 3: 正しく抽出できた属性名（ランダムな抜粋）

注 応募方法 mmLoadMenus ホームページ ニュース
ISO14001 トップ FAQ HOME "HOME 会社情報会社概要"

図 4: 誤って抽出された属性名（$score_{label}$ の上位から抜粋）

3.3 会社概要ページ探索手法の評価結果

ランダムに抽出した 100 社に対して調査を行ったところ，会社概要を含むページを正しく探索できた企業サイトの数は 66 サイトであった．探索を誤ったサイトの中で，もっとも多く誤った会社概要属性が抽出されたページの内訳を表 2 に示す．会社概要を含むページへの経路上に存在するナビゲーションページ，事業所・グループ企業のリスト，および製品・独自サービス紹介ページには，属性名に合致するノードが非常に多く出現する．そのため，このようなページを探索の初期で発見した場合，多くの場合，会社概要ページに到達する前に探索が打ち切られていた．

4. 関連研究

南野は，繰り返し構造に着目し，人間が理解する構造に近い形で HTML 文書の構造を解析する手法を提案している [南野 03]．提案手法は情報抽出のための研究であり，また，類似した構成の Web サイトを大量に利用することで，HTML 文書の構造だけからは得られない語彙的情報を使用することが可能になっている．表形式データの構造認識については，[増田 03] が詳しい．提案手法は表以外の情報も対象としているが，表形式データの正規化については改善が必要であり，その参考としたい．また，[中根 08] では，ブートストラップ的手法を用いて生成したテンプレートを用い，Web 全体からデータベースのスキーマを抽出するための手法を提案しようとしている．提案手法は企業の公式 Web サイトのみを対象とし，企業それぞれの会社概要属性の抽出を目的としているため，直接の比較はできない．[板井 02] は，授業内容のシラバスである HTML 文書を収集し，SVM によって付与したラベル系列に対して，オートマトンを用いて属性名，および属性値を抽出する研究を行っている．提案手法は人手で訓練データを作成する必要はないが，属性値の抽出手法は類似している．

表 2: 会社概要ページであると誤って判断されたページの内容

ナビゲーションページ	11
事業所・グループ企業のリスト	7
製品・独自サービス紹介	4
採用情報	4
IR 情報	3
英文の会社概要ページ	2
その他	3

抽出対象ページの探索については，フォーカストクローリングの分野に先行研究がいくつか ([Ester 03] など) 存在する．会社概要ページの探索について限定的に行った研究はないが，現状の提案手法の正解率は低く，これらの知見をさらに導入する必要があると考えられる．

5. 結論

企業の公式サイトから，会社概要属性の抽出を行う手法を考案した．実験では 100 個の Web サイト全体，抽出対象となる 2500 社の Web サイト，および「会社概要」という文字列のみから，会社概要ページの探索に成功したサイトに限り，0.846 の精度で会社概要属性を抽出できた．しかし，会社概要ページの探索についての正解率は 0.66 と低く，性能向上が課題となる．また，直観に基づく仮定が手法の重要な部分にいくつか用いられている．これらの理論的背景について考察する必要があると考えられる．

参考文献

[Kushmerick 00] Nicholas Kushmerick, Wrapper Induction: Efficiency and Expressiveness, Artificial Intelligence, Vol. 118, 2000.

[南野 03] 南野 朋之, 齋藤 豪, 奥村 学, 繰り返し構造を用いた Web ページの構造化に関する研究, 情報処理学会研究報告. 自然言語処理研究会報告, pp.185–192, 2003.

[中根 08] 中根 史敬, 大坪 正典, 土方 嘉徳, 西田 正吾, Web からのスキーマ抽出に関する基礎検討, DEWS 2008, 2008.

[板井 02] 板井 久美, 高須 淳宏, 安達 淳, HTML からの情報抽出と統合, NII journal, Vol. 6, pp.9–19, 2002.

[Ester 03] Martin Ester, Hans-Peter Kriegel, Matthias Schubert, Accurate and Efficient Crawling for Relevant Websites, Proc. VLDB 2004, pp.396–407, 2004.

[増田 03] 増田 英孝, 塚本 修一, 安富 大輔, 中川 裕志, HTML の表形式データの構造認識と携帯端末表示への応用, 情報処理学会論文誌. データベース, Vol. 44, No. 12, pp.23–32, 2003.