企業 Web ページを用いた関連企業の抽出

本間 友実子, 酒井 浩之, 坂地 泰紀

成蹊大学 理工学部 情報科学科

us122109@cc.seikei.ac.jp, h-sakai@st.seikei.ac.jp, hiroki_sakaji@st.seikei.ac.jp

概要 本稿では, 企業のWebページを用いて関連企業の抽出を行う手法を提案する. ある企業の株価が上昇したということは, その企業と関連のある企業の株価も上昇する可能性があると考えられる. そのため, 関連のある企業を検索する技術の需要がある. しかし, 例えば「キヤノン」は多くの事業を行っているため, 個人投資家にとって, キヤノンに関連する企業を多く見つけることは困難である. そこで, 本研究では, 企業 Web ページからその企業にとって重要な語を抽出し, それを用いて企業の業務内容に基づき, 関連企業を抽出する.

1 はじめに

近年, 証券市場における個人投資家の比重が増大しており, 個人投資家に対して投資判断の支援を行う技術の必要性が高まっている. ここで, ある企業の株価が上昇したということは, その企業と関連している企業の株価も上昇する可能性があると考えられる. そのため, 関連している企業を検索する技術の需要がある. 例えば「キヤノン」と関連している企業として「セイコーエプソン」が考えられる. しかし, キヤノンは多くの事業を行っているため, 個人投資家にとって, キヤノンと関連している企業を多く見つけることは困難である. そこで, 本稿では, 企業の Web ページを用いて関連企業を抽出する手法を提案する. 関連研究として, 酒井らは企業 Web ページから重要語を抽出し, それを検索対象とした企業検索システムを提案している[1]. 文献[1]では入力したキーワードと関連している企業を検索できるが, 本研究による結果を使用することで, 入力を企業名として, その企業と関連している企業を検索できる. 金らは, ニュースサイトなどの Web 情報を用いて資本提携や係争段階といった企業間関係を抽出する手法を提案している[2]. それに対し, 本手法では, 企業 Web ページを用いて関連企業を抽出するため, より多くの関連企業を抽出することができる.

2 提案手法

本手法では, 企業 Web ページから重要な語を抽出し, それを用いてベクトル空間法によって企業間の類似度を求め, 関連企業を認定する. 本研究では, 非上場企業を含めた 16,461 社の企業 Web ページを収集した. この HTML ファイル集合から, 企業ごとに重要語を抽出する.

2.1 重要語の抽出

収集した企業Webページから企業ごとに重要語を抽出する. 重要語は文献[1]の手法を用いて抽出した. 具体的には, 企業 t の Web サイトにおける名詞 n_i に対して, 以下の式 1 で重み $W(n_i,S(t))$を計算し, 重みが大きい語をその企業にとっての重要語とする.

$$W(n_i,S(t)) = \frac{Tf(n_i,S(t)) \times H(n_i,S(t)) \times idf(n_i)}{R(n_i,S(t))} \quad (1)$$

ここで, $S(t)$は企業 t の Web サイトを構成する HTML ファイルの集合, $Tf(n_i,S(t))$は $S(t)$において名詞 n_i が出現する頻度, $H(n_i,S(t))$は S(t)の各 HTML ファイルに名詞 n_i が出現する確率に基づくエントロピー, $R(n_i,S(t))$は企業 t の Web サイトにおいて名詞 n_iの出現する階層, $idf(n_i)$は名詞 n_i の idf 値である. そして, ある企業 t の Web サイトに含まれる名詞 n_i の重み $W(n_i,S(t))$の平均を企業毎に求め, 平均より大きい名詞をその企業 t の重要語とする. 表 1 に,「キヤノン」の Web ページから上記の手法によって抽出された重要語とその重みの値を示す.

表 1 キヤノンから抽出された重要語の例

重要語	$W(n_i,S(t))$
プリンター	68,391
カメラ	42,301
複合機	41,706
デジタル	23,080

2.2 企業間類似度

非上場企業も含めた 16,461 社に対し, ベクトル空間法によって企業間の類似度を求める. 具体的には, 企業の Web ページから抽出した重要語の重みを要素としたベクトルを用いて, 企業間のコサイン類似度を算出した. そして, 類似度が 0.01 以上の企業を関連している企業と認定した. ここで,「キヤノン」の関連企業として抽出された企業の例を表 2 に示す.

表 2 キヤノンの関連企業の例

関連企業名	類似度
ブラザー工業	0.048
セイコーエプソン	0.045
ニコン	0.016
富士フイルム	0.016

2.3 クラスタリング

2.2 節の手法により抽出された，ある企業の関連企業をクラスタリングする．例えば，表 2 のキヤノンの関連企業では「ブラザー工業」と「セイコーエプソン」を 1 つのクラスタとし，「富士フイルム」と「ニコン」を 1 つのクラスタとすれば，関連企業の抽出結果をより分かりやすく提示できる．そこで，階層クラスタリングの完全連結法を用いて，ある企業の関連企業のクラスタリングを行う．以下にクラスタリング手法を示す．

Step 1: 1 つのクラスタに 1 つの企業を割り当てる．

Step 2: 2 つのクラスタ間の類似度を求め，最も類似度が高いクラスタ対を 1 つのクラスタとして融合する．

Step 3: Step 2 を指定したクラスタ数になるまで繰り返す。

Step 2 の 2 つのクラスタ(C_i, C_j)間の類似度 $sim(C_i, C_j)$ を以下の式 2 で求める．

$$sim(C_i, C_j) = \min_{x_k \in C_i, x_l \in C_j} sim(x_k, x_l) \quad (2)$$

$sim(x_k, x_l)$は 2.2 節で求めた2つの企業の類似度である．キヤノンの関連企業に対して，クラスタ数を 10 としてクラスタリングを行った結果の一部を表 3 に示す．

表 3 キヤノンの関連企業のクラスタリング結果の一部

C_i	企業名
i = 1	キヤノン電子，キヤノン ITS，…
i = 2	ニコン，オリンパス，…
i = 3	リコー，東芝テック，…

3 評価

本手法を実装し，評価を行った．関連企業抽出の評価は，ある企業の関連企業を抽出し，企業間類似度の上位 30 社に対し正解，不正解を人手で判定した．5 社に対して評価した精度を表 4 に示す．

表 4 関連企業抽出の精度

キヤノン	87%	富士通	100%
積水ハウス	93%	天下一品	63%
成田国際空港	90%		

4 考察

関連企業抽出の精度は 5 社で平均 86.6%であり，比較的良好な精度を達成した．しかし，「天下一品」の精度は 63%であり，コンビニを展開している企業が関連企業として抽出された．これは例えば「店舗」や「宅配」のような語が「天下一品」とコンビニを展開している企業からも抽出されており，それにより類似度が高くなってしまった．そのため，業務と関連の低い語を重要語から除去できれば，精度の向上が期待できると考える．

5 最適クラスタ数の調査

抽出された関連企業のクラスタリングについて，最適なクラスタ数の調査を行った．ここで，関連企業数が 245 社であった「成田国際空港」においてクラスタ数を変更し，最適クラスタ数を調査した結果，20 が適切であると判断した．表 5 にクラスタ数を 20 とした場合の結果の一部を示す．

表 5 成田国際空港のクラスタ数 20 の場合

C_i	クラスタ数【20】		
1	中国国際航空，新日本航空・・・	11	芝パークホテル，ロコモ・・・
2	日本旅行，沖縄ツーリスト・・・	12	青山ブックセンター，森ビルシティエアサービス・・・
3	富士運輸，第一貨物・・・	13	出雲空港，空港施設・・・
4	水谷建設，鹿島建設・・・	14	三菱総研DCS，大興電子通信・・・
5	アジア航測，日本工営・・・	15	オリックス，日産レンタカー・・・
6	ジャムコ，帝国繊維・・・	16	JALスカイ，全日本空輸・・・
7	関東バス，成田空港交通・・・	17	日本アルキルアルミ
8	沖縄三越，よーじや・・・	18	いわき大王製紙
9	成田珈琲，成田ケーブルテレビ・・・	19	鴻池運輸，共和電業・・・
10	東鉄タクシー，日野交通・・・	20	本庄ケーブルテレビ，シーエス日本・・

同様にいくつかの企業で調査した結果，関連企業数 156 社のキヤノンの最適クラスタ数は 15，関連企業数 1018 社の積水ハウスは 100，関連企業数 1504 社の富士通は 150 であった．このことから，関連企業数が異なることにより最適クラスタ数は企業毎に異なる．

6 むすび

本研究では，企業 Web ページを用いてある企業の関連企業を抽出する手法を提案した．また，関連企業のクラスタリングを，階層クラスタリングを用いて行った．評価の結果，関連企業抽出は高い精度を達成した．今後の課題として，企業ごとの最適なクラスタ数の自動推定を行う手法を検討する．

参考文献

[1] 酒井浩之，坂地泰紀：企業 Web ページを対象とした企業検索システムのための検索クエリに関連するタグの推定，第5回 テキストマイニング・シンポジウム, pp.41-45, 2014.

[2] 金 英子，松尾 豊，石塚 満：Web 上の情報を用いた企業間関係の抽出，人工知能学会論文誌，Vol. 22, No. 1, pp.48-57, 2007