複数の顧客企業からの共通要素と新規関連企業の抽出

田中 瑞竜　　酒井 浩之　　坂地 泰紀
成蹊大学 理工学部 情報科学科

1 はじめに

企業は新規顧客を開拓するため，その可能性のある企業を探す必要がある．しかし，その作業は，企業の事業に関する専門的な知識と多大な労力が必要とされるため新規顧客を探す作業を支援する技術が求められている．

そこで本研究では複数の顧客を保有している状態から新規の顧客を探すことを想定し，複数の顧客企業の共通要素を推定，それを使用して新規顧客となる可能性のある関連企業を自動的に推定する手法を提案する．例えば「キヤノン」「エプソン」「ブラザー工業」といった複数の顧客企業の共通要素（「プリンタ」「複合機」など）を推定し，それを使用して新規関連企業（「リコー」など）を自動的に推定するなどである．本研究によって，顧客リストに記載された企業群の共通要素を推定し，それらから顧客となる見込みのある企業を自動的に抽出することができる．これによって，顧客リストにある企業と似た事業を手がけている企業を容易に探し出すことができ，新規顧客を獲得する助けになる．

本研究では複数の顧客企業の共通要素を推定する情報源として企業の決算短信 PDF を使用し，それに含まれる業績の要因に関連のある文（以降，業績要因文とする）から複数企業の共通要素を推定する．ここで，酒井らは決算短信 PDF から業績要因文を抽出する手法を提案しており [1]，本研究では，それらによって抽出された業績要因文を使用する．業績要因文の一例として，ソニーの場合,「この大幅な減収は、主に、液晶テレビの販売台数が大幅に減少したことによるものです」というような文が該当する．

決算短信を用いる関連研究として，坂地らは決算短信 PDF から原因・結果表現を抽出する手法 [2] を提案している．坂地らの研究では，例えば原因として「猛暑」，結果として「エアコンの売上が増加」という原因と結果の対を決算短信 PDF から抽出する．しかし，必ず対で抽出する必要があるため，抽出できない場合も多く存在する．従って，本研究において複数企業の共通要素を推定する情報源としては適切ではない．企業間関係を抽出する関連研究としては，金らはニュースサイトなどの Web 情報から企業間関係を抽出する手法 [3] を，本間らは企業 WEB ページから関連企業を抽出する手法 [4] をそれぞれ提案している．これらの研究は 2 つの企業間の関係をニュースサイトや WEB ページを使用して推定しているが，本研究では複数企業の共通要素を推定し，それを使用して新規関連企業を抽出している点がこれらの研究とは異なる．

2 本手法の概要

本手法の概要を以下に示す．

Step 1: 決算短信 PDF から抽出した業績要因文 [1] から企業の事業に関連する語（以降，Business Keyword と定義）を企業ごとに抽出する．

Step 2: Business Keyword を使用して，複数の顧客企業（以降，関連企業群と定義）の共通要素を推定する．

Step 3: Business Keyword と共通要素を使用して，入力とする関連企業群を除いた全上場企業の中から関連企業群と関連のある新規関連企業を抽出する．

Step 4: 共通要素と業績要因文を使用して，新規関連企業を事業内容別に分類する．　□

3 業績要因文からの Business Keyword の抽出

ある企業の業績要因文の中に多く出現し，かつ，他の企業の業績要因文にはあまり出現しない語を Business Keyword として抽出する．具体的には業績要因文における名詞 N グラム w_i に対して，以下の式 (1) で tf・idf 値 $Tf \cdot idf(w_i, S(t))$ を計算することで行う．

$$Tf \cdot idf(w_i, S(t)) = Tf(w_i, S(t)) \times Idf(w_i), \qquad (1)$$

ここで，

$S(t)$**:** ある企業 t の業績要因文の集合．

$Tf(w_i, S(t))$**:** $S(t)$ において，名詞 N グラム w_i が出現する頻度．

$Idf(w_i)$**:** 以下の式 (2) で計算される名詞 N グラム w_i の idf 値．

$$Idf(w_i) = \log_2 \frac{|N|}{df(w_i)}, \qquad (2)$$

ここで，$df(w_i)$ は名詞 N グラム w_i が業績要因文に出現する企業の数，N は業績要因文を抽出した上場企業の集合である．表 1 に，上記の手法によって業績要因文から抽出した Business Keyword をいくつか示す．

表 1: 業績要因文から抽出した Business Keyword の例

企業名称	Business Keyword
東芝	パソコン事業，家庭電器
大日本印刷	印刷，産業部門包装，清涼飲料
カゴメ	野菜飲料，トマトジュース，乳酸菌
エーザイ	アリセプト，ビタミンE
三菱商事	資源関連，石化事業，原料炭
キヤノン	プリンター，複合機，カメラ

4 共通要素の推定

新規関連企業を抽出するため共通要素を推定する（詳細は文献 [5] を参照）．その手法の概要を以下に示す．

Step 1: 業績要因文から「液晶パネル製造装置」の「装置」など（以降，Suffix Term と定義）を抽出する．

Step 2: Suffix Term を使用して，業績要因文から「液晶パネル製造装置」の「製造」など（以降，Mid Term と定義）を抽出する．

Step 3: Mid Term と Business Keyword を使用して，関連企業群の共通要素を推定する． □

文献 [5] の手法を用いて Suffix Term は 48 個，Mid Term は 43 個抽出した．

抽出した Suffix Term の一部を以下に示す．

> 費用，用途，等，製品，利用，事業，採用，
> 用品，機器，需要，部品，運用，販売，剤

抽出した Mid Term の一部を以下に示す．

> 用，向け，関連，機器，事業，等，メーカー，
> 業界向け，メーカー向け，市場，製品，業界

4.1 Mid Term と Account Term を用いた共通要素の推定

Mid Term と Business Keyword を使用して，関連企業群の共通要素を推定する．さらに，Mid Term と会計用語キーワード辞典 [1] に登録された会計用語（「償却」「剰余金」など）926 語に，会計用語の名詞 N グラム 1708 語を加えた汎用会計用語 2634 語（以降，Account Term と定義）を利用して，間違った共通要素を除去する．その手法を以下に示す．

Step 1: 関連企業群の各企業の過半数で共通する Business Keyword を共通要素として抽出する．

Step 2: 抽出した複数の共通要素の中で，Mid Term が含まれていて，かつ，Mid Term の前に単語がない共通要素を除去する．

Step 3: 残った共通要素を形態素解析し，先頭の形態素が「名詞」または「接頭詞」でない共通要素を除去する．

Step 4: 残った共通要素の中で，先頭の形態素が「名詞」で，かつ,「名詞・一般」または「名詞・固有名詞」でない共通要素を除去する．

Step 5: 残った共通要素の中で，先頭の形態素が「接頭詞」，その次の形態素が「名詞」で，かつ，その「名詞」の形態素が「名詞・一般」または「名詞・固有名詞」または「名詞・サ変接続」でない共通要素を除去する．

Step 6: 残った共通要素の中で，末尾の形態素が「名詞」でないものを除去する．

Step 7: 残った共通要素の中で，末尾の形態素が「名詞」で，かつ,「名詞・一般」または「名詞・固有名詞」または「名詞・サ変接続」でない共通要素を除去する．

Step 8: 更に残った共通要素の中で，Account Term が含まれていて，かつ，その Account Term が「名詞・サ変接続」でない共通要素を除去する．

Step 9: 最後まで残った共通要素の tf・idf 値を 2 倍し，他の共通要素を含んでいる共通要素には，含まれている共通要素の tf・idf 値の半分をさらに加算する．

Step 10: Step 9 で残っている共通要素の tf・idf 値の降順に，上位 3 個まで共通要素を保持している語（「半導体製造」など）が共通要素に含まれている（「半導体製造装置」など）場合，またはその逆の場合のいずれか以外の共通要素を除去する． □

例えば,「医薬品」に関する関連企業群（ヤクルト本社，日本曹達，日医工）に本手法を適用すると，共通要素として「化学」「医薬品」「医薬品業界」「ジン」「医薬」「薬品」が推定される．

5 新規関連企業の抽出

入力とする関連企業群を除いた全上場企業に対して，以下の式 (3) を用いて類似度 $dist(\boldsymbol{V_{ce}}, \boldsymbol{V_t})$ を計算することで新規関連企業を抽出する．

$$dist(\boldsymbol{V_{ce}}, \boldsymbol{V_t}) = \frac{\boldsymbol{V_{ce}} \cdot \boldsymbol{V_t}}{\|\boldsymbol{V_{ce}}\| \times \|\boldsymbol{V_t}\|}$$

[1] http://kaikei-yougo.sigyo.net/

$$= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}, \quad (3)$$

$$\boldsymbol{V_{ce}} = \{x_1, x_2, ..., x_n\}, \boldsymbol{V_t} = \{y_1, y_2, ..., y_n\},$$

ここで,

V_{ce}: 4 章の手法で，入力とする関連企業群から共通要素として推定された Business Keyword のベクトル．$x_1, x_2, ..., x_n$ は式 (1) で求めた Business Keyword の tf・idf 値．

V_t: 3 章の手法で，ある企業 t の業績要因文の集合 $S(t)$ から抽出した Business Keyword のベクトル．

6 抽出した新規関連企業の分類

新規関連企業を「共通要素と直接的に関連のある企業」と「共通要素と間接的に関連のある企業」のどちらかに分類する（詳細は文献 [5] を参照）．その手法を以下に示す．

Step 1: 新規関連企業の業績要因文の集合から共通要素を含む最長の名詞 N グラムを抽出する．

Step 2: 抽出した中から共通要素の後に続く語を抽出する．

Step 3: さらに抽出した中に Mid Term が含まれている場合（パターン A），含まれていない場合（パターン B）の頻度をそれぞれ集計する．

Step 4: 集計結果を基に，パターン A の頻度がパターン B の頻度以上の場合は新規関連企業を「共通要素と間接的に関連のある企業」に，そうでない場合は「共通要素と直接的に関連のある企業」に分類する．□

例えば,「半導体製造装置」に関する関連企業群（東京応化工業，東芝機械，石井工作研究所）に本手法を適用すると，新規関連企業として「芝浦メカトロニクス」「大日本スクリーン製造」「日本トムソン」「朝日工業社」「安川電機」が抽出され,「芝浦メカトロニクス」「大日本スクリーン製造」が直接的に,「日本トムソン」「朝日工業社」「安川電機」が間接的にそれぞれ分類される．

7 実装

酒井らの手法 [1] によって，3,821 社の企業 Web ページから 106,885 個の決算短信 PDF ファイルを取得し，それらから業績要因文を抽出した．その中から「Business Keyword」「Suffix Term」「Mid Term」をそれぞれ抽出し，抽出した各語を使用して入力とする関連企業群の共通要素と，それに関連する新規関連企業を推定した．さらに推定された新規関連企業は事業内容別に分類した．実装にあたり，形態素解析器として MeCab[2]，係り受け解析器として CaboCha[6] を使用した．

[2]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

8 評価

業績要因文が存在する企業の中からランダムに選んだ企業 300 社において,「太陽電池」「液晶パネル」「セキュリティシステム」「半導体製造装置」「無停電電源装置」の各関連企業群（各 3 社）を対象に新規関連企業を抽出し，本手法の評価を行った．その評価では，抽出した新規関連企業の Web ページを確認し，実際に関連がある（共通要素または共通要素の関連製品を製造，販売している）場合は正解とし，精度，再現率を算出した．新規関連企業の分類結果の評価では，分類した新規関連企業の Web ページを確認し，実際に関連があり，なおかつそれが分類通り（共通要素を製造，販売する企業は直接的のクラス，共通要素の関連製品を製造，販売する企業は間接的のクラス）の場合に正解とし，精度，再現率を算出した．ただし，例外的に Mid Term が「事業」の場合は，共通要素を製造，販売する企業を正解とした．抽出の比較手法として，Business Keyword の代わりに，決算短信 PDF 全体から tf・idf 値を計算して抽出した語を用いた場合の精度，再現率を算出した．分類の比較手法として，Mid Term の代わりに,「用」と「向け」のみを用いた場合の精度，再現率を算出した．本手法，比較手法ともに，抽出できる新規関連企業がなくなるまで類似度 $dist(\boldsymbol{V_{ce}}, \boldsymbol{V_t})$ の順に 1 社ずつ抽出，分類し，その精度と再現率を基にグラフ（◯：本手法，△：比較手法）を作成した.「液晶パネル」,「セキュリティシステム」の評価結果をまとめたグラフを図 1，図 2 と図 3，図 4 にそれぞれ示す．その他のグラフについては「液晶パネル」と似たような結果であるため，ここでは割愛する．

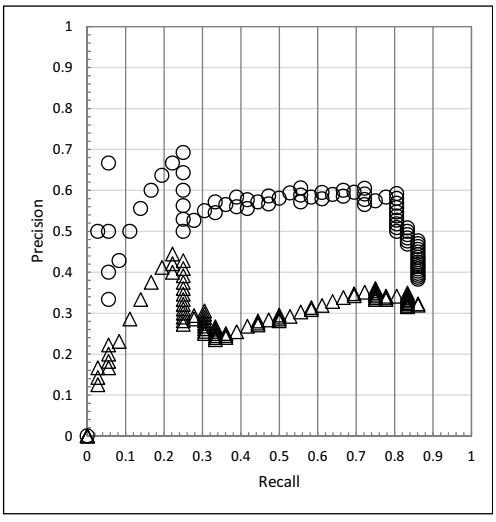

図 1: 液晶パネル・抽出

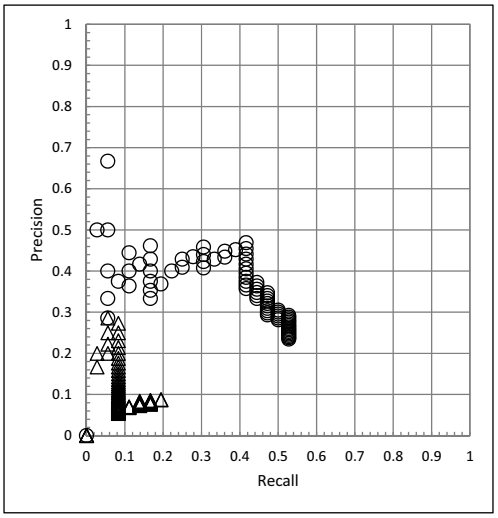

図 2: 液晶パネル・分類

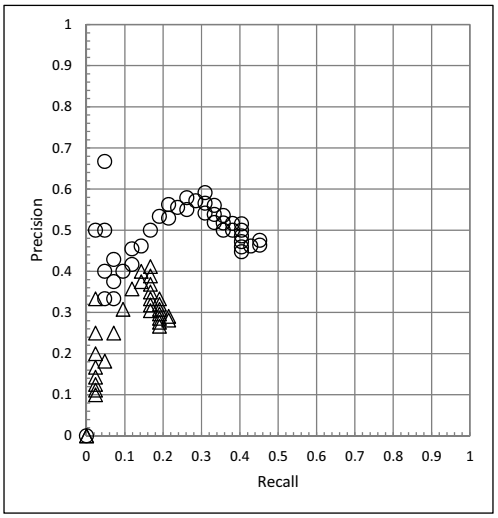

図 3: セキュリティシステム・抽出

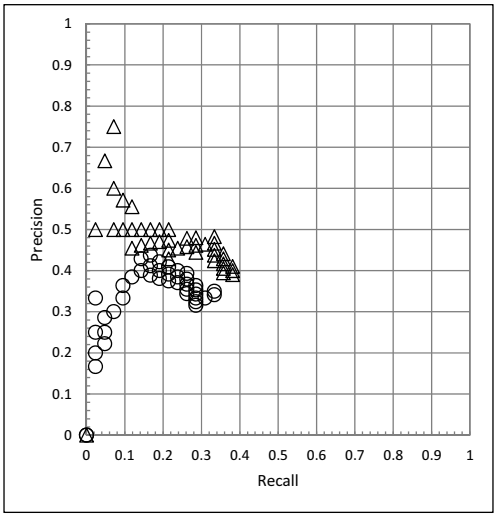

図 4: セキュリティシステム・分類

9 考察

図 1，図 2 と図 3，図 4 から，本手法は比較手法と比べて，抽出と分類，どちらにおいても平均して高い精度と再現率を達成できた．唯一，セキュリティシステムの分類では比較手法の方が良い結果となった．これは企業 300 社のうち，セキュリティシステムに直接関連する企業の数が間接的に関連する企業に比べて 3 倍以上多いことと，セキュリティシステムが意味の広い語であることが関係していると考える．

10 おわりに

本論文では，例えば「キヤノン」「エプソン」「ブラザー工業」といった関連企業群における共通要素（「プリンタ」「複合機」など）を推定し，推定した共通要素を使用して新規関連企業（「リコー」など）を自動的に抽出する手法を提案した．さらに，抽出した新規関連企業は「共通要素と直接的に関連のある企業」と「共通要素と間接的に関連のある企業」に分類した．本手法を評価した結果，抽出と分類，どちらにおいても比較手法より平均して高い精度と再現率を達成できた．

参考文献

[1] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀. 企業の決算短信 PDF からの業績要因の抽出. 人工知能学会論文誌, Vol. 30, No. 1, pp. 172–182, 2015.

[2] 坂地泰紀, 酒井浩之, 増山繁. 決算短信 PDF からの原因・結果表現の抽出. 電子情報通信学会論文誌 D, Vol. J98-D, No. 5, pp. 811–822, 2015.

[3] 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.

[4] 本間友実子, 酒井浩之, 坂地泰紀. 企業 web ページを用いた関連企業の抽出. 第 7 回 Web インテリジェンスとインタラクション研究会, pp. 13–14, 2015.

[5] 田中瑞竜, 酒井浩之, 坂地泰紀. 関連企業集合からの共通要素と新規関連企業の抽出. 第 9 回 テキストマイニング・シンポジウム, pp. 19–24, 2016.

[6] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.