企業の業績発表記事からの業績要因抽出と最も重要な業績要因の判定

Determination of the Most Important Causal Expression Extracted from Articles Concerning Business Performance of Companies

酒井浩之 [1*]　増山繁 [2]
Hiroyuki Sakai [1]　Shigeru Masuyama [2]

[1] 成蹊大学 理工学部 情報科学科
[1] Department of Computer and Information Science, Faculty of Science and Technology, Seikei University
[2] 豊橋技術科学大学 大学院 工学研究科 情報・知能工学専攻
[2] Department of Computer Science and Engineering, Toyohashi University of Technology

Abstract: We introduce a method of extracting causal information (e.g., Demand for semi conductor manufacturing equipments is good) from Japanese financial articles concerning business performance of companies. Causal information is useful for investors in selecting companies to invest. Our method automatically extracts causal information as a form of causal expression by using statistical information and initial clue expressions. We extract keywords from Web sites of companies and improve the previous method by using them. Moreover, our method automatically determines the most important causal expression by using these keywords extracted from Web sites of companies.

1　はじめに

近年，人工知能分野の手法や技術を，金融市場における様々な場面に応用することが期待されており，例えば，膨大な金融情報を分析して投資判断の支援をする技術が注目されている．さらに，最近では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おこなう技術の必要性が高まっている．投資家にとって，企業の業績に関する情報を収集することは重要であるが，実際の業績に関する情報だけでなく，その業績要因が重要である．なぜなら，業績拡大の要因が，その企業の主力事業が好調であることであったならば株価への影響は大きいが，株式売却益の計上などの特別利益の計上が要因であるならば株価への影響は軽微であるからである．しかしながら，証券市場の上場企業数は約 3,500 社と多いうえに，近年では年に 4 回，決算発表がある．さらに，大幅な業績の修正を行う場合にも業績修正発表を行う必要があるため，人手によって全ての企業の業績要因を取得するのには多大な労力を要する．そのため，我々は，経済新聞記事から企業の業績発表記事を抽出し，その中から業績要因（例えば,「主力の半導体製造装置の受注が好調」）を抽出する手法を提案した [6].

我々の業績要因の抽出手法では，抽出すべき業績要因を「共通頻出表現」と「手がかり表現」の 2 つの表現で構成される形態素列と定義した．ここで，手がかり表現を業績要因獲得のための手がかり的な形態素列と定義し，共通頻出表現を異なった業績要因に対して共通して頻出する形態素列と定義した．例えば,「ソフト販売の収益が寄与する」では，手がかり表現が「が寄与する」であり，共通頻出表現は「ソフト販売」,「収益」である．業績要因の抽出手法では，数多くの「手がかり表現」と「共通頻出表現」を自動的に獲得し，それらを使用することで業績要因を抽出する．それに加え，ある企業の業績発表記事から業績要因を抽出する際に，その企業の Web サイトから当該企業にとって重要なキーワードを抽出し，抽出したキーワードを共通頻出表現として使用することで，重要であるにもかかわらず，文献 [6] では抽出できなかった業績要因を抽出できるような手法を追加した．本稿では，まず，文献 [6] の手法と，追加した手法について述べる．

我々の手法では，一つの業績発表記事からは複数の

*連絡先：成蹊大学 理工学部 情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: h-sakai@st.seikei.ac.jp

業績要因が抽出されることが多い．ここで，抽出した複数の業績要因の中で，その企業にとって特に重要な業績要因を提示できれば，高度な専門知識がない個人投資家に対する投資判断支援を行うための有用な情報源になることが期待できる．しかし，大きな企業になると複数の事業を行っており，何が主力の事業であるかは，その企業に対する知識を要し，そして，個人投資家が必ずしも多くの企業の主力事業を熟知しているわけではない．例えば,「三菱電機」は多くの事業を行っているが，会社四季報 [9] によれば，三菱電機の特色欄に「ＦＡが収益柱」という記述がある．そのため，三菱電機の業績要因として「ＦＡ（ファクトリーオートメーション）が好調（もしくは不振）」であれば，投資判断を行ううえで重要な情報となる．そこで，企業の Web サイトから当該企業にとって重要なキーワードを抽出し，それを使用することで，最も重要な業績要因を自動的に判定する手法の開発を行った．

2 関連研究

関連研究として，我々は，抽出した業績要因に対して業績に対する極性（「ポジティブ」,「ネガティブ」）を付与する手法を提案した [7]．例えば，業績要因「ソフト販売の収益が寄与する」に対しては「ポジティブ」,「繊維部門の不振が響く」に対しては「ネガティブ」のラベルを付与する．業績要因に対して極性を付与することで，業績要因を使用した景気動向予測，および，業績要因に基づいて株取り引きを行うコンピュータトレーディングにも応用できることが期待できる．藤村らは，業績要因が当該企業の本業と関連があるか否かで分類する手法を提案した [1]．分類には SVM を使用し，素性として形態素のユニグラム，バイグラムを使用した．また，イベントスタディ法に基づく分析によって，業績要因を含む業績発表記事が株式市場に対し影響を与えている可能性があることが示された．西崎らは，業績要因文に含まれる製品や部門情報の抽出を行う手法を提案した [5]．具体的には，業績要因文を対象とした調査の結果得られた規則性を用いてパタンを作成する．そして，作成されたパタンを適用するために，抽出対象を含む業績要因文を単語に分割し，取得した形態素列に対してパタンを適用して製品や部門情報の抽出を行った．しかし，抽出される業績要因の中には，例えば「前期から取り組む販売体制の見直しで費用が発生した」のような重要ではない業績要因も存在している．より精度の高い，業績要因を使用した景気動向予測，および，業績要因に基づいて株取り引きを行うコンピュータトレーディングには，特に重要な業績要因のみを使用する必要があると考える．

Koppel らは，企業に関する記事がその企業の株価に影響を与えるかどうかを判別する手法を提案している [2]．Lavrenko らは，企業に関する記事内容によって引き起こされる株価の動きを予測するための手法を提案している [4]．和泉らは，日本銀行の金融経済月報を題材としたテキストマイニングによる債券市場の動向分析を行っている [8]．しかしながら，これらの研究では，株価や債券市場に影響を与える原因は分析できない．それに対して，本研究では，業績発表記事から業績要因を抽出し，さらに最も重要な業績要因を判定することで，株価に影響を与える原因を業績発表記事の中から取得することができる．

3 企業の業績発表記事からの業績要因抽出

本節では，文献 [6] の企業の業績発表記事から業績要因を自動的に抽出する手法について簡単に述べる．文献 [6] の手法では，抽出すべき業績要因を「共通頻出表現」と「手がかり表現」の 2 つの表現で構成される形態素列と定義し，共通頻出表現と手がかり表現を自動的に獲得することで抽出を行う．

Step 1: 少数の手がかり表現（「が好調」,「が不振」）を人手で与え，それに係る節を取得する．

Step 2: 取得した節の集合から，その中で共通して頻繁に出現する表現を共通頻出表現として抽出する．

Step 3: 共通頻出表現が係る節を，新たな手がかり表現として獲得する．

Step 4: 獲得した手がかり表現から，それに係る節を取得する．

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す（図 1 を参照）．□

3.1 共通頻出表現の抽出

本節では，共通頻出表現の自動獲得について述べる．まず，手がかり表現に直接係っている文節から助詞を除去した形態素列を c とおく．そして，c に対して，それに係る文節を追加することで派生する表現を取得し，既に得られている表現に係る文節を次々に追加することで派生する表現を全て取得する．例えば,「新型の自動車の売り上げが好調」という文において，手がかり表現「が好調」を使用する場合,「売り上げ」が c となり,「売り上げ」,「自動車の売り上げ」,「新型の自動車

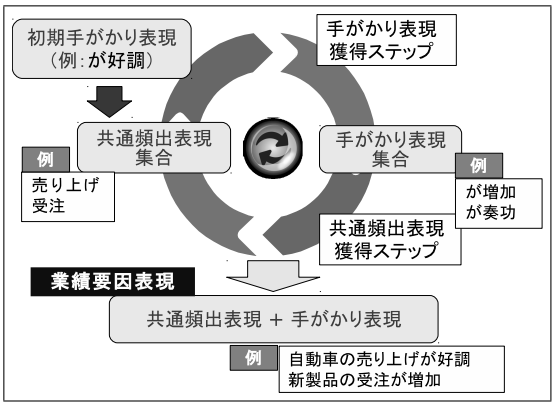

図 1: 共通頻出表現・手がかり表現自動獲得手法の概要

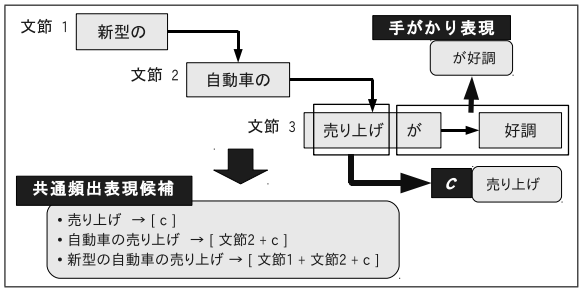

図 2: 共通頻出表現候補の獲得

の売り上げ」の 3 つの表現を取得する（図 2 を参照）．次に，c から派生した各表現 e に対して，以下の式 1 で表されるスコアを計算する．

$$Score(e,c) \quad = \quad -f_e(e,c)\sqrt{f_p(e)}\log_2 P(e,c) \quad (1)$$

ただし，

$f_p(e)$: 表現 e に含まれる文節の数，

$P(e,c)$: c から派生する表現 e の派生確率，

$f_e(e,c)$: c から派生する表現 e の派生回数．

例えば，文書 A に「新店の紳士服の売り上げが好調」という文が存在していたとすれば，「売り上げ」，「紳士服の売り上げ」，「新店の紳士服の売り上げ」という 3 つの表現を取得する．また，文書 B に「主力のカードゲームの売り上げが好調」という文が存在していたとすれば，この文から「売り上げ」「カードゲームの売り上げ」「主力のカードゲームの売り上げ」という 3 つの表現を取得する．そして，文書 A と文書 B からは「売り上げ」が 2 回，「カードゲームの売り上げ」，「主力のカードゲームの売り上げ」，「紳士服の売り上げ」，「新店の紳士服の売り上げ」が 1 回，派生したことになる．そのため，「売り上げ」の $f_e(e,c)$ の値は 2 であり，c から派生する表現の総数は 6 であるため，$P(e,c)$ の値は 2/6 となる．ここで，c から派生する表現の中で，$f_e(e,c)$ の値が 2 以上である表現のうちスコアが最大の表現を共通頻出表現候補として抽出する．

次に，抽出された共通頻出表現候補の中から適切な共通頻出表現を選別する．具体的には，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 2 で求め，その値が閾値 T_e 以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s\in S(e)} P(e,s)\log_2 P(e,s) \quad (2)$$

ただし，業績発表記事集合において，

$P(e,s)$: 共通頻出表現 e が手がかり表現 s に係る確率，

$S(e)$: 共通頻出表現 e が係る手がかり表現の集合．

閾値 T_e は，以下の式 3 によって設定する．

$$T_e = \alpha\log_2|Ns| \quad (3)$$

ただし，Ns は共通頻出表現を取得するのに使用した手がかり表現の集合，α は定数 $(0<\alpha<1)$ である．

3.2 新たな手がかり表現の獲得

共通頻出表現の選別を行った後，その選別した共通頻出表現から新たな手がかり表現を獲得する．まず，抽出した共通頻出表現を含む文を抽出し，その中で共通頻出表現を含む節 P_a が係っている文節 P_b を獲得する．次に，P_a に含まれる助詞を P_b に追加し，それを手がかり表現候補とする．ここで，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定に基づき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを式 4 で求め，閾値以上の候補を手がかり表現として抽出する．

$$H(s) = -\sum_{e\in E(s)} P(s,e)\log_2 P(s,e) \quad (4)$$

ただし，業績発表記事集合において，

$P(s,e)$: 手がかり表現 s に対して共通頻出表現 e が係る確率，

$E(s)$: 手がかり表現 s に係る共通頻出表現の集合．

閾値は，共通頻出表現と同様に式 3 によって設定するが，Ns は新たな手がかり表現を獲得するのに使用した共通頻出表現の集合である．

表 1 に，定数 α を 0.5 とした場合に獲得した手がかり表現と共通頻出表現の例をいくつか示す．

表 1: 獲得された共通頻出表現・手がかり表現の例

共通頻出表現	売り上げ，リストラ費用，受注量，自社製品 電子材料，原材料費，採算，人件費，開発費
手がかり表現	が順調。，が苦戦した。，が堅調だった。，で補う。，が低迷した。 が回復する，で落ち込んだ。，が伸び悩む。，が貢献する。

3.3 共通頻出表現，手がかり表現を使用した業績要因表現の抽出

獲得した手がかり表現と共通頻出表現を使用して業績要因を抽出する．ここで，獲得された手がかり表現を業績要因を抽出するために使用する場合，手がかり表現に追加した格助詞を除去した文字列を使用する．例えば，手がかり表現「が好調」の場合,「好調」を使用する．そして，手がかり表現を含む文節に係る文節から派生する表現のうち，最長の表現に共通頻出表現が含まれている場合，その表現と手がかり表現を連結して業績要因とする．

表 2 に，定数 α を 0.5 とした場合に獲得した手がかり表現と共通頻出表現を使用して抽出された業績要因をいくつか示す．

4 共通頻出表現の追加による改良

文献 [6] による共通頻出表現の獲得手法では，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを求め，エントロピーが閾値より高い値が割り当てられる共通頻出表現候補を共通頻出表現として選別する．従って，多くの種類の手がかり表現に係っている共通頻出表現候補を共通頻出表現として選別しているが，そのためには，業績発表記事中に高い頻度で共通頻出表現が出現している必要がある．しかしながら，この手法では，商品名や部門名等の，その企業にとって重要なキーワードでありながら，特定の企業の業績発表記事にのみ出現しているため出現頻度が低い表現は，共通頻出表現として選別されない．例えば,「靴向け人工皮革も伸び悩んだ。」を業績要因として抽出するには,「人工皮革」が共通頻出表現として獲得される必要があるが，この語は人工皮革を製造している企業の業績発表記事中にのみ出現しているため，出現頻度が低く，共通頻出表現として獲得されなかった．そのため，ある企業の業績発表記事から業績要因を抽出する際に，その企業の Web サイトから重要なキーワードを抽出し，そのキーワードも共通頻出表現に追加することで，抽出される業績要因を増やす．

収集した企業 Web サイトからのキーワードの抽出は，企業 t の Web サイトにおける名詞 n_i に対して，以下の式 5 で重み $W(n_i, S(t))$ を計算することで行う．

$$W(n_i, S(t)) = (0.5 + 0.5\frac{Tf(n_i, S(t))}{max_{i=1,..,m}Tf(n_i, S(t))}) \times H(n_i, S(t))/R(n_i, S(t)) \quad (5)$$

ここで，

$S(t)$: ある企業 t の Web サイトを構成する HTML ファイルの集合．

$Tf(n_i, S(t))$: $S(t)$ において，名詞 n_i が出現する頻度．なお，0.5 を足している理由は，名詞 n_i の頻度による重みへの影響を，0.5 から 1 の間にするためである．

$H(n_i, S(t))$: $S(t)$ の各 HTML ファイル d に名詞 n_i が出現する確率 $P(n_i, d)$ に基づくエントロピー．以下の式 6 によって求める．

$$H(n_i, S(t)) = -\sum_{d \in S(t)} P(n_i, d) \log_2 P(n_i, d) \quad (6)$$

$R(n_i, S(t))$: 企業 t の Web サイトにおいて名詞 n_i の出現する階層．例えば，Web サイトのトップページに出現していれば，$R(n_i, S(t)) = 1$．/products/list/index.html に出現していれば，$R(n_i, S(t)) = 3$ とする．

$W(n_i, S(t))$ は，企業 t の Web サイトを構成する HTML ファイルの集合中に多く，まんべんなく出現し，かつ，トップページのような Web サイトの上位階層に出現している名詞に対して，大きな値が割り当てられる．

さらに，企業 t の Web サイトに出現しており，他の企業の Web サイトには出現していない名詞を判定するために，以下の式 7 で $idf(n_i)$ を計算する．

$$idf(n_i) = \log_2 \frac{N}{df(n_i)} \quad (7)$$

$df(n_i)$: 名詞 n_i を含む HTML ファイルで構成される Web サイトを持つ企業の数．

N: Web サイトを収集した企業の数（後述する評価では，1200 社の Web サイトを使用）．

表 2: 抽出された業績要因の例

業績要因 1 ：	主力の液晶テレビが海外で価格下落が強まり、販売苦戦も響く。
共通頻出表現：	下落，苦戦，価格下落
手がかり表現：	が響く
業績要因 2 ：	子会社が手掛ける情報サービス処理事業の受注増などが寄与した。
共通頻出表現：	受注，受注増
手がかり表現：	が寄与した。
業績要因 3 ：	鉄鋼商品の価格上昇が追い風となり、主力の鉄鋼事業の利益が拡大。
共通頻出表現：	利益，上昇，価格上昇
手がかり表現：	が拡大。

そして，$Tf(n_i, S(t))$ が 2 より大きく，$H(n_i, S(t))$ が 1 より大きく，かつ，$idf(n_i)$ が 1 より大きい名詞を企業 Web サイトからキーワードとして抽出し，それらを共通頻出表現として追加する．すなわち，名詞 n_i がキーワードとして抽出される条件は，企業 t の Web サイトに少なくとも 3 回以上，かつ，別々の HTML ファイルに出現し，さらに，名詞 n_i が半分以上の他の企業 Web サイトに出現していない必要がある．例えば，「トップページ」や「お問い合わせ」といった語は企業 Web サイトには頻出する語であるため，$Tf(n, S(t))$ と $H(n, S(t))$ が大きくなるが，そのような語の $idf(n_i)$ は 1 より大きくならないため，キーワードとして抽出されない．また，「人工皮革」のように「人工皮革」を製造している企業の Web サイトには頻出するような語は，$Tf(n, S(t))$ と $H(n, S(t))$ に加え，$idf(n_i)$ も大きくなるため，キーワードとして抽出されやすくなる．

表 3 に，文献 [6] の手法では獲得できなかったが，企業の Web サイトから抽出したキーワードを共通頻出表現として追加したことで抽出できるようになった業績要因をいくつか示す．

5 最も重要な業績要因の自動判定

本節では，企業の Web サイトから上記の手法によって抽出した，当該企業にとって重要なキーワードを利用し，最も重要な業績要因を自動的に判定する手法について述べる．具体的には，企業 t の Web サイトから抽出したキーワードの重み $W(n_i, S(t))$ を使用して業績要因に対して重要度を付与し，最も大きい重要度が付与された業績要因を最も重要な業績要因として判定した．企業 t の業績要因 $ce(t)$ への重要度 $W(ce(t))$ は，以下の式 8 で求めた．

$$W(ce(t)) = \frac{\sum_{n_i \in T(ce(t))} W(n_i, S(t))}{x(ce(t))} \tag{8}$$

ここで，

$W(n_i, S(t))$: 式 (5) で求めた名詞 n_i の重み

$T(ce(t))$: 企業 t の業績要因 $ce(t)$ に含まれる，t の企業 Web サイトから抽出したキーワードの集合

$x(ce(t))$: 業績発表記事において，業績要因 $ce(t)$ が出現している文番号．例えば，$ce(t)$ が記事の最初から 2 番目の文に出現していれば，$x(ce(t)) = 2$ となる．

6 実装

1990 年から 2008 年の日経新聞記事集合から 71070 個の業績発表記事を取得し，その記事集合から，手がかり表現，および，共通頻出表現を獲得し，それらと，企業の Web サイトから取得したキーワードを使用して業績要因の抽出を行った．また，抽出した業績要因を検索対象にした検索システムを作成した．図 3 は，作成した検索システムにおいて「太陽電池」で検索した結果である．検索結果では，業績発表記事の表題と記述されている企業名が表示されている．実装にあたり，形態素解析器として MeCab[1]，係り受け解析器として CaboCha[3] を使用した．検索された業績発表記事を選択すると，その記事に含まれる業績要因を抽出し，重要度を付与する．また，文献 [7] の手法を使用することで，業績要因に極性（ポジティブ，ネガティブ）を付与する（図 4 を参照.）ポジティブと判定された業績要因は上矢印，ネガティブと判定された業績要因は下矢印で表現される．重要度は星印で表現し，最も大きい重要度が付与された業績要因は星印 3 つで表す．図 4 では，三菱電機の業績発表記事から業績要因を抽出し，極性，および，重要度を付与している．本手法により，「設備投資関連のファクトリーオートメーション機器、エアコンの好調が寄与した。」に対して最も大きい重要度を付与しており，それに対して，「独禁法関連費用四

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

図 3: 業績要因を対象とした検索システム（「太陽電池」での検索例）

図 4: 業績要因の抽出と極性，重要度付与

表 3: 抽出された業績要因の例（企業 Web サイトから抽出したキーワードを追加）

業績要因 1 ：	インクジェットプリンターの販売が順調だった
企業名： 追加共通頻出表現： 手がかり表現：	セイコーエプソン インクジェット，プリンター，インクジェットプリンタ が順調
業績要因 2 ：	金融向けシステム開発が好調だった。
企業名： 追加共通頻出表現： 手がかり表現：	ＮＴＴデータ システム開発，金融向け が好調だった
業績要因 3 ：	主力の有料老人ホーム事業で入居者が予想以上に増えた。
企業名： 追加共通頻出表現： 手がかり表現：	ツクイ 有料老人ホーム，老人，入居，有料老人ホーム事業 が増えた。

百二十一億円を営業外費用に計上したが、」には重要度を付与していない．

7 評価

本手法の評価を行った．本タスクにおける正解データの作成には，企業の主力事業といった知識を必要とするため，多く，かつ，精度の高い正解データを作成することは多大な労力と専門知識を要する．そこで，以下のような手法で正解データを作成した．

Step 1: 2007 年，2008 年の日経新聞記事集合から取得した業績発表記事において，「主力の」というフレーズを 1 つ含む記事を取得する．この結果，410 個の業績発表記事を取得した．

Step 2: 取得した 410 個の業績発表記事から，人手により一番重要な業績要因を含む文を判定する．

「主力の」というフレーズを採用した理由は，正解データの精度向上と作成の負担を軽減できるためである．そして，正解データとして用意した業績発表記事から抽出した業績要因に対して，本手法により重要度を付与した．正解データ中の，人手により一番重要と判断した業績要因を含む文と，本手法により最も高い重要度を付与された業績要因を含む文が一致すれば正解，異なっていれば不正解とした．図 5 に評価結果を示す．ここで，

Lead(ce) 法: 業績発表記事において最初に抽出される業績要因を最重要の業績要因として判定する手法，

Lead(s2) 法: 業績発表記事において，2 番目に出現する文を最重要の業績要因として判定する手法（正解データより，最重要の業績要因は 2 番目の文に出現することが多いことに基づく.），

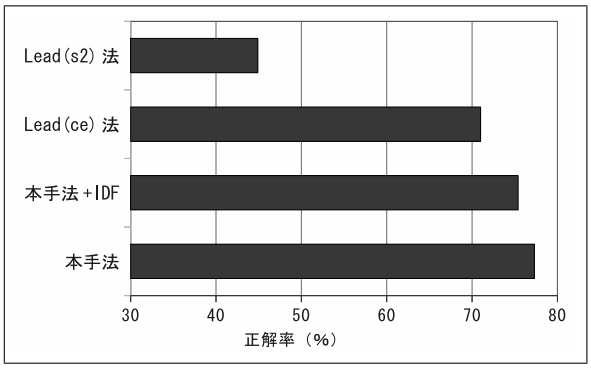

図 5: 最重要な業績要因判定の評価結果

本手法＋ IDF: 企業から抽出したキーワードの重みとして，以下の式のように，idf 値を導入したもの．

$$W(n_i, S(t))' = W(n_i, S(t)) \times idf(n_i) \qquad (9)$$

本手法では，410 記事中 317 件が正解であり，正解率は 77.3% であった．

8 考察

図 5 から，本手法は業績発表記事において，2 番目に出現する文を最重要の業績要因として判定する Lead(s2) 手法や業績発表記事において最初に抽出される業績要因を最重要の業績要因として判定する Lead(ce) 手法より高い正解率を達成している．例えば，「リコー」の業績発表記事から，「主力の複写機販売は景況感悪化を受け、日米で落ち込んだ。」，「研究開発費の増加や原材料高も響いた。」，「前期から取り組む販売体制の見直しで

費用が発生した」,「モノクロ機の販売が減り」といった業績要因が抽出された．この中では,「リコー」の Web サイトから抽出したキーワードである「複写機」や「複写」に高い重みが付与されていたため,「主力の複写機販売は景況感悪化を受け、日米で落ち込んだ。」という業績要因に最も高い重要度を付与することができた．

本手法に idf 値を導入した「本手法＋ IDF」が，本手法よりも正解率が低下した．その理由は以下のとおりである．本手法に idf 値を導入することによって，例えば,「東京楽天地」の業績発表記事から抽出された「新規に開店したフィットネスクラブが堅調だったが、ビルメンテナンスが苦戦した。」と「主力の不動産賃貸事業が不振。」では,「新規に開店したフィットネスクラブが堅調だったが、ビルメンテナンスが苦戦した。」が最重要な業績要因と判定された．これは,「開店」といった一般的な名詞であっても，企業 Web サイトにはあまり出現しない語に対して，高い idf 値が付与されることと,「不動産」のような名詞は多くの企業 Web サイトに出現している名詞であるため，idf の値が低くなるためである．そのため，重要度付与のための重みには，idf 値を導入しないほうが，高い正解率を達成できた．

9 まとめ

本稿では，経済新聞記事における企業の業績発表記事を対象としたテキストマイニングの一環として，業績発表記事から業績要因に関する情報（例えば,「半導体製造装置の受注が好調」）を抽出する手法について報告した．文献 [6] では，抽出すべき業績要因を「共通頻出表現」と「手がかり表現」の 2 つの表現で構成される形態素列と定義し，それらを企業の業績発表記事集合から自動的に獲得し，使用することで業績要因を抽出した．それに加えて，ある企業の業績発表記事から業績要因を抽出する際に，その企業の Web サイトから当該企業にとって重要なキーワードを抽出し，抽出したキーワードを共通頻出表現として使用することで，文献 [6] では抽出できなかった業績要因を抽出できるような手法を追加した．また，企業の Web サイトから抽出したキーワードを使用して業績要因に重要度を付与することで，最重要な業績要因の自動判定を行った．評価の結果，本手法は業績発表記事において，２番目に出現する文を最重要の業績要因として判定する Lead(s2) 手法や業績発表記事において最初に抽出される業績要因を最重要の業績要因として判定する Lead(ce) 手法より高い正解率を達成した．

謝辞

本研究は，科研費 若手研究 B(21700158) の助成を受けたものである．また，言語データとして，日経新聞 CD-ROM の使用を許可して頂いた日本経済新聞社に深謝する．

参考文献

[1] 藤村真太郎, 酒井浩之, 増山繁: 企業業績要因文の経常的か否かに基づく分類とイベントスタディ法に基づく分析, 第 23 回人工知能学会全国大会 (2009).

[2] Koppel, M. and Shtrimberg, I.: Good News or Bad News? Let the Market Decide, *In Proceedings of the AAAI Spring Symposium on Exploring At titude and Affect in Text*, pp. 86–88 (2004).

[3] 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).

[4] Lavrenko, V., Schmill, M., Lawrie, D. and ilvie, P. O.: Mining of Concurrent Text and Time Series, *In Proceedings of the KDD 2000 Conference Text Mining Works hop*, pp. 37–44 (2000).

[5] 西崎海人, 酒井浩之, 増山繁: 製品・部門情報の企業業績要因表現からの抽出, 第 24 回人工知能学会全国大会 (2010).

[6] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008).

[7] Sakai, H. and Masuyama, S.: Assigning Polarity to Causal Information in Financial Articles on Business Performance of Companies, *IEICE Trans. Information and Systems*, Vol. E92-D, No. 12, pp. 2341–2350 (2009).

[8] 和泉潔, 後藤卓, 松井藤五郎: テキスト情報を用いた金融市場分析の試み, 第 22 回人工知能学会全国大会 (2008).

[9] 東洋経済新報社: 会社四季報 2012 年 1 集 新春号, 東洋経済新報社 (2011).