人工知能学会研究会資料
SIG-FIN-015-02

企業の決算短信ＰＤＦから抽出した業績要因への極性付与

Assigning Polarity to Causal Information Extracted from PDF Files of the Summary of Financial Statements of Companies

酒井浩之[1*]　小林義和[1]　坂地泰紀[1]
Hiroyuki Sakai [1]　Yoshitaka Kobayashi　Hiroki Sakaji [1]

[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

Abstract: In this paper, we propose a method of assigning polarity to causal information extracted from PDF files of the summary of financial statements of companies. Our method assigns polarity (positive or negative) to causal information in accordance with business performance, e.g. "Orders of semiconductor manufacturing equipments were strong". (The polarity positive is assigned in this example.) First, we assigns polarity to clue expressions to be used to extract causal information. Using them, our method assigns polarity (positive or negative) to causal information. We evaluated our method and confirmed that it attained 88.4% precision and 84.2% recall of assigning polarity positive, and 91.4% precision and 74.5% recall of assigning polarity negative, respectively.

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おこなう技術の必要性が高まっている．そのため，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行う技術が注目されている．その一例として，日本銀行が毎月発行している「金融経済月報」や経済新聞記事をテキストマイニングの技術を用いて，経済市場を分析する研究などが盛んに行われている [1][3][6].

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 [7][4][5][6]．なぜなら，業績回復の要因が，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東京証券取引所の上場企業だけでも２０１５年９月４日現在，3490 社と多いうえに [1]，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，人手によって多く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．そのために酒井らは，企業が業績発表を行った直後に，企業の Web サイト等で一般に公開されている企業の決算短信 PDF に着目し，その中から業績要因を含む文（例えば「半導体製造装置の受注が好調でした。」）を抽出する手法を提案した [7].

決算短信 PDF より抽出された業績要因は，例えば，証券アナリストへの支援材料として利用できる．しかし，より有効な情報として利用するためには，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」）を付与する必要がある．例えば，業績要因「半導体製造装置の受注が好調でした。」に対しては「ポジティブ」,「世界的な太陽電池市況の低迷により太陽電池製造装置の販売が減少しました。」に対しては「ネガティブ」の極性を付与する．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．そのため，本稿では，決算短信 PDF より抽出された業績要因に対して極性（「ポジティブ」,「ネガティブ」）を自動的に付与する手法を提案する．

*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
[1]http://www.jpx.co.jp/listing/stocks/co/index.html

2　関連研究

酒井らは，経済新聞記事の企業の業績発表記事から抽出した業績要因（例えば「自動車の売り上げが好調」）に対して極性を付与する手法を提案している [5]．ここで，例えば「が好調」の手がかり表現が含まれる業績要因には「ポジティブ」のラベルを付与できるが，手がかり表現「が増加」のように,「売り上げが増加」はポジティブであるが「リストラ費用が増加」はネガティブであるため，手がかり表現だけでは極性を判定できない．そこで，まず，業績発表記事を機械学習手法の１つであるＳＶＭ [8] を使用して業績に対する極性で分類し（すなわち，業績が向上したなら「ポジティブ」，業績が悪化したなら「ネガティブ」に分類），その情報を利用して,「売り上げ」と「が増加」の組合せで構成される業績要因に対しては「ポジティブ」,「リストラ費用」と「が増加」の組合せで構成される業績要因に対しては「ネガティブ」の極性が付与する．しかし，この酒井らの提案した手法は，業績発表記事がＳＶＭで高精度で極性付与できることが前提条件となっている．そのため，酒井らの手法をそのまま決算短信 PDF に適用するのであれば，ＳＶＭで決算短信 PDF に対して極性を付与する必要があるが，ＳＶＭでは決算短信 PDF に対して高い精度で極性を付与できない．そのため，酒井らの手法を決算短信 PDF から抽出した業績要因への極性付与には適用できない．

3　決算短信ＰＤＦから抽出した業績要因への極性付与

3.1　手がかり表現への極性付与

酒井らの手法 [7] では，業績要因抽出のための手がかり的な形態素列（以降，手がかり表現と定義）を決算短信 PDF から自動的に獲得し，その手がかり表現を使用することで業績要因を抽出している．例えば「半導体製造装置の受注が好調でした。」では，手がかり表現が「が好調でした」である．ここで，業績要因に極性を付与する場合,「手がかり表現」に注目して極性付与を行う．

酒井らの手法 [7] では，企業 Web ページから取得した 106,885 個の決算短信 PDF から 121 種類の手がかり表現を獲得し，それを使用して業績要因を抽出している．本手法では，この手がかり表現に極性（ポジティブ，ネガティブ）を人手で付与し，その手がかり表現の極性を使用して業績要因への極性付与を行う．表 1 に，人手で極性付与した手がかり表現の一部を示す．しかしながら，例えば「推移しました」のような手がかり表現は極性を付与できない．なぜなら「堅調に推移しました」はポジティブとなるが,「厳しい状況で推移しました」はネガティブとなる．このように，手がかり表現に係る文節列（上記の例では「堅調に」「厳しい状況で」）によって極性が変わる手がかり表現がある．以降，手がかり表現に係る文節列によって極性が変わる手がかり表現を「極性変化手がかり表現」と定義する．121 種類の手がかり表現を分析したところ，17 種類の極性変化手がかり表現が存在した．以下に，極性変化手がかり表現の例を示す．

表 1: 極性付与された手がかり表現の例

手がかり表現	極性
堅調でした	positive
好調でした	positive
伸び悩んだ	negative
寄与しました	positive
低調でありました	negative
低迷した	negative

推移する，転じる，推移し，推移いたしました，継続し

ここで，極性変化手がかり表現に係っている文節列と極性変化手がかり表現の組み合わせを決算短信 PDF 集合から求め，それぞれの組み合わせに対して人手で極性を付与すればよい．しかし，例えば「推移しました」の場合でも,「堅調に推移しました」のように直近に係っている 1 文節のみをみれば極性が確定できる場合もあるが,「厳しい状況で推移しました」「上回る状況で推移しました」のように 2 文節を考慮しなくては極性が判定できない場合もある．そのため，極性変化手がかり表現と係っている文節列との組み合わせは膨大な数となり，全ての組み合わせに対して人手で極性を付与することは多大な労力を要する．そのため，次節にて，極性変化手がかり表現と係っている文節列との組み合わせを絞り込み，少ない労力で極性変化手がかり表現と係っている節との組み合わせに対して極性を付与する手法について述べる．

3.2　極性変化手がかり表現への極性付与

極性変化手がかり表現への極性付与を行うにあたり，106,885 個の決算短信 PDF において極性変化手がかり表現に係っている文節列を取得する．例として，図 1 に「依然として厳しい状況で推移し」の文において，極性変化手がかり表現「推移し」に係っている文節列を取得する手法を示す．ここで，手がかり表現の「推移し」のみでは極性は付与できないが,「厳しい状況で推移し」では極性を付与できる.（この場合は「ネガティブ」を付与する.）しかし，例えば「推移し」に係って

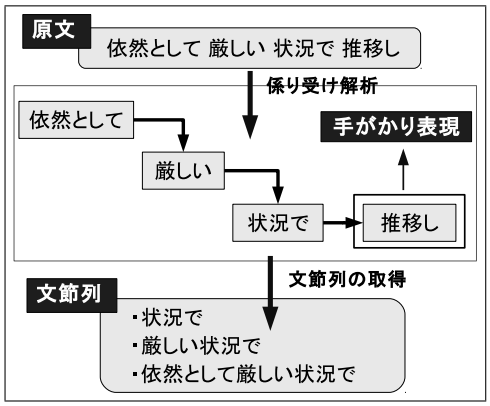

図 1: 文節列の取得例

いる「状況で」という 1 文節のみでは「状況で推移し」となり，これだけでは極性が確定できない．なぜなら「厳しい状況で推移し」であればネガティブであるが，「堅調な状況で推移し」であればポジティブとなるからである．(ただし，「厳しい状況で推移し」で極性は確定できるため，「依然として厳しい状況で」までの情報は必要ない．) そのため，決算短信 PDF 集合において極性変化手がかり表現に係る文節列を全て取得し，その文節列と極性変化手がかり表現との組み合わせにおいて極性を判定する．

ここで，106,885 個の決算短信 PDF において，極性変化手がかり表現「推移し」に係る文節列を取得したところ，11077 個の文節列を取得した．以下に，「推移し」に係る文節列の例を示す．

> 不透明な状況で，厳しい経営環境で，計画を上回って，堅調に，好調に，低調なまま，前年同期を上回って，計画を上回る水準で，燃料価格が高水準で，好調な売上で

このように，「推移し」だけでも非常に多くの文節列が取得され，さらに他の極性変化手がかり表現に係る文節列も含めれば，極性変化手がかり表現とそれに係る文節列との組み合わせ数は膨大なものになる．そこで，極性変化手がかり表現とそれに係る文節列との組み合わせを絞り込む．具体的には，極性変化手がかり表現 c に係る文節列 p に対して以下の式 1 でスコアを求め，このスコアがある閾値を上回る文節列のみを抽出する．

$$Score(p,c) = -f(p,c)\sqrt{fp(p)}\log_2 P(p,c) \quad (1)$$

$$P(p,c) = \frac{f(p,c)}{N(c)} \quad (2)$$

ただし，決算短信 PDF から取得した業績要因を含む文の集合において，

$P(p,c)$: 極性変化手がかり表現 c から取得される文節列 p の出現確率，

$f(p,c)$: 極性変化手がかり表現 c から取得される文節列 p の取得回数．

$N(c)$: 極性変化手がかり表現 c から取得される文節列の総数

$fp(p)$: 文節列 p に含まれる文節の数，

例えば，決算短信 A に「依然として厳しい状況で推移し」という文が存在していたとすれば，手がかり表現「推移し」から「状況で」,「厳しい状況で」,「依然として厳しい状況で」という 3 つの文節列を取得する．また，決算短信 B に「引き続き厳しい状況で推移し」という文が存在していたとすれば，この文から「状況で」「厳しい状況で」「引き続き厳しい状況で」という 3 つの文節列を取得する．そして，決算短信 A と決算短信 B からは「状況で」,「厳しい状況で」が 2 回,「依然として厳しい状況で」,「引き続き厳しい状況で」が 1 回，取得したことになる．そのため,「厳しい状況で」の $f(p,c)$ の値は 2 であり，c から取得する文節列の総数 $N(c)$ は 6 であるため，$P(p,c)$ の値は 2/6 となる．

ここで，c から取得される文節列の中で，$f(p,c)$ の値が 5 以上であり，スコアが 100 以上の文節列を抽出する．これにより，例えば，極性変化手がかり表現「推移し」に係る 11077 個の文節列を 222 個に絞り込むことができた．そして，絞り込まれた文節列と極性変化手がかり表現との組み合わせに対して，人手にて極性を付与する．

3.3 「増加」と「減少」を含む手がかり表現への極性付与

極性付与された手がかり表現のなかで,「増加」と「減少」を含む手がかり表現は，その手がかり表現に係っている文節列によって極性が反転する場合がある．例えば，手がかり表現の「増加しました」はポジティブの極性を付与するが,「費用が」や「コストが」が係っている場合は，ネガティブの極性を付与する必要がある．手がかり表現の「減少しました」の場合も同様である．このような極性を反転させる文節列を取得するために，極性付与した手がかり表現のなかで「増加」と「減少」を含む手がかり表現に係っている文節列を取得する．そして，3.2 節で示した式 1 により各文節列のスコアを求め，$f(p,c)$ の値が 5 以上であり，スコアが 100 以上の文節列を抽出した．それにより，620 個の文節列を取得し，その中から極性を反転させる文節列を選択した．以下に極性を反転させる文節列をいくつか示す．

一般管理費が，費用が，減価償却費が，短期借入金が，人件費が，有利子負債が，営業損失が，コストが

3.4 極性付与された手がかり表現を使用した業績要因への極性付与

極性付与された手がかり表現を使用して，業績要因への極性付与を行う．例えば，以下の業績要因に対して極性を付与することを考える．

海外売上高は、中国向け昇降機事業が堅調に推移した社会・産業システム部門等が増加したものの、ハードディスクドライブ事業を売却したことや、電子装置・システム部門等が前年同期を下回ったことから、前年同期に比べ１１％減少し、８，６７７億円となりました

この業績要因には，ポジティブが付与された手がかり表現として「増加し」「堅調に推移し」が含まれている．また，ネガティブが付与された手がかり表現として「減少し」「下回った」が含まれている．(「増加し」「減少し」に係っている文節は極性反転を伴わない.）この業績要因は、２つのポジティブな手がかり表現、ネガティブな手がかり表現があるが，業績要因の極性は最後の手がかり表現（この例では「減少し」）の極性に従う．そのため、この業績要因には「ネガティブ」が付与される．

4 決算短信 PDF への極性付与

３章の手法により極性付与された業績要因を使用して，決算短信 PDF への極性付与を試みる．ここで，酒井らは，ある企業の決算短信 PDF から業績要因を抽出する際に，その企業にとって重要なキーワード（企業キーワードと定義）を抽出し，手がかり表現と企業キーワードを使用して業績要因を抽出している [7]．例えば，セイコーエプソンの決算短信 PDF から「デバイス精密機器事業」，「センサー産業機器事業」，「インクジェットプリンター」といった語を企業キーワードとして抽出している．この企業キーワードは，企業 t の決算短信 PDF 集合における名詞 n に対して，以下の式 3 で重み $W(n, S(t))$ を計算し，高いスコアが割り当てられる名詞 n を抽出することで行う．

$$W(n, S(t)) = TF(n, S(t))H(n, S(t)) \log_2 \frac{N}{df(n)} \quad (3)$$

ここで，

$S(t)$: ある企業 t の決算短信 PDF の集合．

$TF(n, S(t))$: $S(t)$ において，名詞 n が出現する頻度．

$H(n, S(t))$: $S(t)$ の各決算短信 PDF である d に名詞 n が出現する確率 $P(n, d)$ に基づくエントロピー．以下の式 4 によって求める．

$$H(n, S(t)) = -\sum_{d \in S(t)} P(n, d) \log_2 P(n, d) \quad (4)$$

$df(n)$: 名詞 n を含む決算短信 PDF をもつ企業の数．

N: 決算短信 PDF を収集した企業の数．

決算短信 PDF への極性付与は，その決算短信 PDF から抽出した業績要因に付与した極性と，それに含まれる企業キーワードのスコア $W(n, S(t))$ を使用して行う．まず，企業 t の業績要因 $ce(t)$ の重要度 $W(ce(t))$ を式 5 で求める．

$$W(ce(t)) = \sum_{n_i \in T(ce(t))} W(n_i, S(t)) \quad (5)$$

ここで，

$W(n_i, S(t))$: 企業キーワード n_i のスコア

$T(ce(t))$: 企業 t の業績要因 $ce(t)$ に含まれる，t の企業キーワードの集合

次に，ある決算短信 PDF に含まれる「ポジティブ」が付与された業績要因の重要度 $W(ce(t))$ の和と，「ネガティブ」が付与された業績要因の重要度 $W(ce(t))$ の和を求め，和が大きいほうの極性を，その決算短信 PDF の極性とする．

5 実装

本手法を実装し，106,885 個の決算短信 PDF ファイルから，酒井らの手法 [4] で業績要因を含む文を抽出し，それに対して本手法にて極性を付与した．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[2] を使用した．また，決算短信 PDF から抽出した業績要因を含む文を検索対象とした，決算短信 PDF の検索システムを実装し [3]，その検索システムにて表示される業績要因に対して，本手法にて付与された極性が表示されるようにした．図 2 に，実装したシステムにて「半導体」で検索した場合の検索結果を示す．図 2 にて「半導体」を含む業績要因が表示され，その業績要因を含む決算短信 PDF と企業名が表示される．企業名の下に表示されているキーワードは，その企業の企業キーワードである．業績要因の左

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[3] http://hawk.ci.seikei.ac.jp/cees/

図 2: 業績要因を対象とした検索システム（「半導体」での検索例）

図 3: 業績要因を対象とした検索システム（「三菱電機」での検索例）

に表示されている矢印が極性を表しており,「↑」でポジティブ,「↓」でネガティブを表す．また，決算短信 PDF のタイトルの左の矢印がその決算短信 PDF の極性であり，業績要因と同じく,「↑」でポジティブ,「↓」でネガティブを表す．また，企業名で検索することでその企業の決算短信 PDF を検索することができ，検索された決算短信 PDF に含まれる業績要因が表示される．図 3 に「三菱電機」で検索した場合の検索結果を示す．

6 評価

6.1 業績要因への極性付与の評価

業績要因への極性付与手法の評価を行った．ここで，正解データは 30 個の決算短信 PDF から抽出された 411 個の業績要因を含む文に対して人手にて極性を付与して作成した．なお，ポジティブの業績要因は 254 文，ネガティブの業績要因は 157 文であった．そして，同じ 411 個の業績要因文に対して本手法にて極性を付与し，その精度，再現率を求めた．結果を表 2 に示す．比較手法として，人手によって極性付与された手がかり表現のみを使用した場合（極性変化手がかり表現，および，極性反転の文節を使用しない）の精度，再現率を表 3 に示す．

表 2: 業績要因への極性付与の評価結果（本手法）

	positive	negative
精度（%）	88.4	91.4
再現率 (%)	84.2	74.5

表 3: 業績要因への極性付与の評価結果（比較手法）

	positive	negative
精度（%）	88.0	86.5
再現率 (%)	69.2	57.3

6.2 決算短信 PDF への極性付与の評価

極性付与された業績要因を使用した決算短信 PDF への極性付与手法の評価を行った．正解データとして，55 個の決算短信 PDF を選択し，人手にて極性を付与した [4]．決算短信 PDF への極性付与は，その決算短信 PDF を精読し，例えば，黒字であったとしても業務環境が悪化して前年同期よりも減益であった場合や，当初の予想よりも業績が悪化した場合はネガティブの極性を付与する．逆に，たとえ赤字であったとしても，当初の予想よりも赤字幅が縮小した場合はポジティブとなる．それにより，36 個の決算短信 PDF にポジティブ，19 個の決算短信 PDF にネガティブの極性が付与された．

本手法にて評価対象とした 55 個の決算短信 PDF に対して極性を付与し，正解データと比較して，ポジティブの精度，ネガティブの精度，全体の正解率を求めた．ここで，比較手法として，SVM を使用して極性を付与した手法による正解率を示す．SVM の学習用データは 100 個とし，素性として名詞を使用した．評価結果を表 4 に示す．

[4]投資の経験が豊富で決算短信をよく読んでいる，株式投資歴が約１５年の評価者が極性を付与した．

表 4: 決算短信 PDF への極性付与の評価結果

手法	精度 (positive)	精度 (negative)	正解率
本手法	80.5%	78.9%	80.0%
SVM	90.0%	46.6%	54.5%

7 考察

本手法による業績要因への極性付与手法の精度はポジティブで 88.4%，ネガティブで 91.4% であり，高い精度を達成した．また，再現率はポジティブで 84.2%，ネガティブで 74.5% であり，極性変化手がかり表現や極性反転文節列の取得手法において，手がかり表現に係っている文節列を大幅に絞り込んでいるが，比較的良好な結果を得ることができた．人手によって極性付与された手がかり表現のみを使用する比較手法では再現率がポジティブで 69.2%，ネガティブで 57.3% であることから，本手法は極性変化手がかり表現と文節列の組み合わせを使用して極性を付与したため，再現率が向上していることが分かる．また，ネガティブの精度が比較手法では 86.5% であることから，本手法は極性反転文節列によって適切に極性を付与できたため，精度が向上している．例えば，本手法では，以下の業績要因に対してネガティブの極性を付与したが，比較手法ではポジティブの極性が付与された．

> 人材不足の状況を改善するため人員を採用し人件費が増加した結果、営業利益は８百万円（前年同期比５８．０％減）となりました

これは，本手法では「人件費が」と手がかり表現「増加した」の組み合わせで，ネガティブの極性を付与しているからである．

本手法による決算短信 PDF への極性付与の正解率は 80% であり，SVM による極性付与より高い正解率を達成した．SVM による決算短信 PDF への極性付与の正解率は 54.5% であり，これにより，経済新聞記事の業績発表記事から抽出した業績要因に対して極性を付与する酒井らの手法 [5] を，そのまま本タスクへは適用できないことが分かる．

8 まとめ

本稿では，決算短信 PDF より抽出された業績要因に対して極性（「ポジティブ」,「ネガティブ」）を自動的に付与する手法を提案した．また，極性付与された業績要因を使用して決算短信 PDF に極性を付与する手法を提案した．業績要因への極性付与手法では，人手によって極性付与された手がかり表現を使用して極性を付与し，例えば「推移しました」のような，それだけでは極性が確定できない手がかり表現の場合は，手がかり表現とそれに係っている文節列との組み合わせを使用して極性を付与した．決算短信 PDF への極性付与は，その決算短信 PDF から抽出した業績要因に付与した極性と，それに含まれる企業キーワードのスコアの和を使用し，スコアの和の大きいほうの極性を，その決算短信 PDF へ付与する極性とした．評価の結果，業績要因への極性付与手法の精度はポジティブで 88.4%，ネガティブで 91.4%，再現率はポジティブで 84.2%，ネガティブで 74.5% であり，良好な精度，再現率を達成した．

参考文献

[1] 和泉潔, 後藤卓, 松井藤五郎: 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011).

[2] 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).

[3] 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志: 新聞記事のテキストマイニングによる長期市場動向の分析, 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296 (2013).

[4] 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).

[5] 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).

[6] 酒井浩之, 増山繁: 企業の業績発表記事からの重要業績要因の抽出, 電子情報通信学会論文誌 D, Vol. J96-D, No. 11, pp. 2866–2870 (2013).

[7] 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. J98-D, No. 5, pp. 172–182 (2015).

[8] Vapnik, V.: *Statistical Learning Theory*, Wiley (1999).