成蹊大学理工学研究報告
J. Fac. Sci.Tech., Seikei Univ.
Vol.51 No.2 (2014)pp.53-60
（特別研究費に係る論文）

株価に影響を与える重要な出来事が記載された記事の自動抽出

中山　大*1，坂地　泰紀*2，勝田　研一郎*3，酒井　浩之*4

Extraction of Important Articles that Influence the Stock Price of Companies from Financial Articles

Masaru Nakayama*1, Hiroki Sakaji*2, Kenichiro Katuda*3, Hiroyuki Sakai*4

ABSTRACT：This paper proposes a method of extracting important articles that influence the stock price of companies from financial articles. In the first step, our method obtains dates that have large difference from previous day at stock price of a selected company. In the second step, our method acquires articles that are published around the obtained dates and concern the selected company from financial articles. In the third step, our method extracts articles that influence the stock price of the selected company by using SVM as a machine learning method from the acquired articles. Finally, we evaluated our method. As a result, our method achieved 69.8% precision and 53.1% recall.

Keywords：Extraction of Important Articles, Text Mining, Natural Language Processing

(Received September 21, 2012)

１．はじめに

近年，企業の株価の上昇とともに証券市場における個人投資家の比重が増大している．しかし，全ての個人投資家が投資や投資対象の企業に関して深い知識を持ち合わせているとは言い難い．そこで，投資家が参考にする物として，「有価証券報告書」や「会社四季報」が挙げられる．しかし，前者では保有している会社の数だけ目を通さなければならない．後者では，記載されている情報に限りがある．それに加え，コンピュータ技術の進展により，企業のホームページにおける決算資料や証券会社の分析レポートなど，経済市場を分析したコンテンツは日々増加しており，個人投資家が全ての情報に目を通し取捨選択するのは不可能に近い．そこで，近年では金融テキストマイニングが注目されている．具体的には，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．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」や新聞記事を，テキストマイニング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)8)]．

図１　日本触媒の株価推移（2012/6/29～2012/12/28）

投資家にとって，経済新聞に掲載されるような企業に関する情報（出来事）を閲覧することは重要である．なぜなら，経済新聞に掲載されるような企業に関する出来事と，その企業の株価とは関連性が顕著に出る場合があるからである．図１に示したのは日本経済新聞のWEBサイトに掲載されている「日本触媒」の株価推移である[1]チャートの中で，９月末に株価が大きく下落している事が見てとれる．この時，以下のような事故が起きた．

*1：情報科学科　学部４年
*2：情報科学科　助教
*3：情報科学科　学部４年
*4：情報科学科　准教授（h-sakai@st.seikei.ac.jp）

[1] http://www.nikkei.com/markets/company/index.aspx?scode=4114

姫路市の日本触媒工場で爆発　負傷者約 30 人か
2012/9/29 15:30 兵庫県警によると、29 日午後２時半ごろ、兵庫県姫路市網干区の日本触媒の敷地内で爆発があった。黒煙が立ち上がっており、約 30 人のけが人が出ているもよう。

上の記事は，日経新聞電子版のニュースである．この事故は，自社の姫路工場で主力のアクリル酸のプラントで爆発事故が発生し，1 人が死亡したというものである．この事故をきっかけに，週明けの株価が前週末 終値比 135 円安の 738 円となった．

日本触媒が大幅安、１年半ぶり安値　工場爆発事故を嫌気
2012/10/1 10:42　1 日の東京株式市場で日本触媒株が大幅安となり、一時は前週末比 135 円（15 ％）安の 738 円と、約１年半ぶりの安値を付けた。紙おむつ用の高吸水性樹脂（ＳＡＰ）や、原料のアクリル酸を生産する姫路製造所（兵庫県 姫路市）で爆発事故が発生。工場全体の生産再開見通しが立たず、収益への悪影響を懸念する売りが膨らんだ。

この他にも，株価が大きく変動する出来事としては，企業の決算発表，他企業との業務提携，他企業との大型契約の締結，企業買収，画期的な新製品の発表，粉飾決算の発覚などが考えれる．

ここで，個人投資家にとって，ある企業の現在の株価になった理由を，過去の出来事から検索したいという要求が存在している．例えば，ある企業の株価が企業価値から考慮しても割安で推移しているとし，この企業の株を購入することを考えるとする．しかし，過去に何かの事件（例えば，業績の下方修正，粉飾決算の発覚，工場の爆発など）があったために，割安の株価で推移していることが懸念される．そのようなときに，その企業の株価に影響を与えた出来事が記載された記事を素早く検索することができれば，株価が割安で推移している理由を容易に調べることができ，一般の個人投資家に対する投資判断の有益な支援になると考える．

そこで，本研究では，ある企業の株価に影響を与えるような，その企業に関連する経済新聞記事を自動的に抽出することを目的とする．例えば，「ボーイング７Ｅ７炭素素材、東レ、３３００億円独占受注，１８年契約.」といったポジティブな出来事や，「日立、２６００億円の赤字．今期最終損失、中間配当見送り.」といったネガティブな出来事のどちらに関しても抽出することを目指す．また，ある企業の株価が大きく変動した日付の記事を検索したとしても，その内容が株価に影響を与えるような重要な内容であるとは限らない．例えば「世界自動車戦争の勝者と敗者は誰か（社説）」のようなコラムが，ホンダの株価が大きく変動した日に掲載されていた．本研究では，このような株価に影響を与えない出来事が記載された記事は自動的に排除する．本研究により，ある企業の現在の株価（例えば，割安な株価で推移している等）となった主な出来事を，個人投資家に対して素早く提示できると考える．

2. 関連研究

本研究の関連研究として，和泉らは，日銀が毎月発行している「金融経済月報」を用い，株式市場（日経平均株価），外国為替市場（円ドルレート），国債市場（金利）の分析を行っている [1]．文献 [1]の研究では，共起解析（co-occurrence analysis），主成分分析（principal component analysis），回帰分析（regression analysis）からなるCPR法を提案し，テキストを解析することで 経済動向の分析を行っている．さらに，蔵本らは文献 [1]の研究をさらに発展させ，入力テキストを新聞記事として，長期的な株式市場の分析を行っている [3]．文献 [3]の研究では，金融経済月報より形式が定まっていない文章である新聞記事を入力としたため，前述のCPR法を拡張することで 60%以上の株式市場の騰落正答率を収めた．Sakajiらは，景気動向記事を対象に，景気が上がるか下がるかの根拠表現を抽出し，抽出した根拠表現に対して極性（ポジティブ，ネガティブ）を付与する手法を提案している [9]．Mileaらは，欧州中央銀行が発行している報告書からFuzzy Grammar Fragmentを抽出し，それに基づき，MSCIユーロ・インデックスを予測（上向き，もしくは，下向きに推移するかどうか）している [5]．これらの研究では，企業の株価データなどの基礎情報を用いず，日銀の金融経済月報や景気動向記事のような新聞記事といったテキスト情報を用い，将来の市場予測を行っている．しかし，過去にその企業に関する出来事を投資家が求めたときに提示することを目的としているわけではない．個人投資家は今後の株価予測も重要視するが，その企業で過去にどのような事が起きて現在の株価になったのかを知り，その上で投資判断をすると考える．

経済記事の内容をポジティブかネガティブかに判定する研究がいくつか行われている．Koppelらは，企業に関する記事に対して，株価が上昇する内容であるか下落する内容であるかを分類する手法を提案している [2]．

Lavrenkoらは，企業に関する記事が発表された後 の株価動向を推定する手法を提案している [4]．これらの研究では，入力として与えられる記事が株価に影響を与えるほどの重要な記事であることを前提としている．しかし，実際には，株価に変化を与えるほどの影響がない記事も多く含まれており，さらに，ある企業の株価が大きく変動した日付の記事を検索したとしても，その内容が株価に影響を与えるような重要な内容であるとは限らないため，そのような記事を判別する技術が必要である．

酒井らは，日本経済新聞記事における企業の業績発 表記事より，例えば「半導体製造装置の受注が好調」のような業績要因表現を抽出している [6]．さらに，抽出された業績要因表現の中から最も重要な業績要因の判定や，業績要因表現への極性（ポジティブ，ネガティブ）の付与を行っている [7][8]．しかし，これらの研究では，企業の業績発表記事のみを対象としている．実際は，企業の業績発表記事以外にも，株価に変動を与えるような出来事が記載された記事が存在しており，それらの記事を分析の対象にできることが望ましい．そこで，本研究では日本経済新聞記事の中で，業績発表 記事だけでなく，「経営統合」などの一般の記事をも対象とし，過去にその企業で株価に影響を与えるような出来事が記載された記事を自動的に抽出することを目指す．

3．株価に影響を与える出来事が記載された 記事の自動抽出

本節では，日本経済新聞記事より株価に影響を与える出来事が記載された記事を取得する手法の説明を行う．

3．1　手法概要

手法の概要を以下に示す．

Step 1：株価変動の前日比が± 8%以上変動している日付を取得する．

Step 2：取得した前日比のリストを元に，日付の周辺の新聞記事を検索し，調べたい企業に関連している記事を取得する．

Step 3：取得した新聞記事を調べ，株価に影響を与えるような記事であればポジティブ，そうでなけれ ばネガティブとする訓練データを作成する．

Step 4：作成した訓練データからサポートベクトルマシン(以下，SVMと記す)10)により分類器を生成し，訓練データ以外の記事に対して，株価に影響を与える記事とそうでない記事に分類する．

以上のステップで解析を行う．

3．2　使用するデータ

株価のデータは 1983 年 1 月 4 日から 2013 年 1 月 15 日までの約 30 年のデータを使用する．株価データとしては，東京証券取引所，大阪証券取引所，ジャスダックデータの株価データがあるが，複数の証券取引所に上場している企業の場合，以下の優先順位で株価データを使用する．

東証 ＞ 大証 ＞ ジャスダック

なお，東証と大証は 2013 年 7 月 16 日に統合し，改編などが行われたが，本研究では統合前のデータを使用している．経済新聞記事は，1990 年 1 月 1 日から 2008 年 12 月 31 日までの日本経済新聞を使用する．

3．3　新聞記事と企業との関連付け

日本経済新聞記事の記事が，ある企業と関連のある記事であるかを調べる．まず，上場企業の証券コードと企業名が記載されたリストを用いる．次に，日本経済新聞記事の記事を 1 記事ずつ調べて記事内の複合名詞を取得し，複合名詞と企業名のリストを照らし合わせ，完全に一致した際に，その企業と関連のある記事として扱う．なお，企業名が複数出現した場合には，最初に出現した企業と関連のある記事として扱うこととする．以下に例を示す．

> 「日立・伊藤忠、水力発電設備を受注。環境円借款、中国向け第一号。」日立製作所と伊藤忠商事は共同で、中国湖南省の電力会社から環境配慮型の水力発電設備を受注した。

上記の記事では，日立製作所と伊藤忠商事が企業名として取得できるが，日立製作所が最初に完全一致した企業名なので，日立製作所と関連のある記事として扱う．なお，1990 年 1 月 1 日から 2008 年 12 月 31 日までの日本経済新聞（3,360,881 記事）において，上記の手法で取得した，企業に関連のある記事の数は519,579 記事あった．

3．4　株価データと新聞記事データの関連付け

ある企業の株価の前日比±8%を規準とし，ある企業の株価が±8%以上変動したときの日付を取得する．なお，前日比 8%とした理由は，ストップ高，ストップ安となる制限値幅の 3 割から 5 割程度の変動となり，経済新聞に掲載されるような出来事が起きている可能性が高いと考えたからである．株価の前日比の計算には以下の式を用いる．

$$\text{前日比} = \frac{\text{当日の終値} - \text{前日の終値}}{\text{前日の終値}} \times 100 \qquad (1)$$

得られた株価の前日比のデータを用い，ある企業の株価が±8%以上変動した当該日付の前日と当該日付から 2 日後の合計 4 日間における，その企業に関する新聞記事を取得する．図 2 の例では，13 日に+8%の変動があるので，前日の 12 日から 2 日後の 15 日までの計 4 日間の記事を取得する．

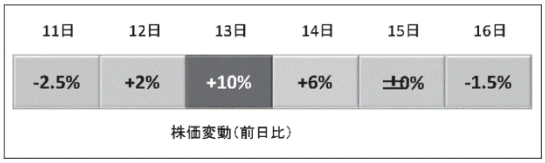

図２　前日比の例

3．5　訓練データの作成

SVMに用いる訓練データを人手で作成する．訓練データに用いる企業は，時価総額が高く，かつ，多くの新聞記事がある企業を 10 企業選択した．選択した 10 企業を表 1 に示す．なお，以降は各企業を表 1 で示した証券コードで示す．3.3 節の手法で取得した 10 企業に関連する新聞記事の全てに目を通し，株価の変動に関連しうる記事であればポジティブ，そうでなければネガティブとする訓練データを作成する [2]．その結果，ポジティブな記事は 292 記事あり，その他のネガティブな記事（1000 以上）より無作為に 292 記事取得し，合計で 584 記事を訓練データとした．

表１　訓練データとして選択した企業

企業名	証券コード
日立製作所	6501
東芝	6502
ソニー	6758
日産自動車	7201
トヨタ自動車	7203
ホンダ	7267
キヤノン	7751
伊藤忠商事	8001
東海旅客鉄道	9022
日本電信電話	9432

3．6　素性選択とSVMによる記事分類

前節で取得した訓練データより，SVMによる分類に使用する素性を選択する．素性は，訓練データの記事を形態素解析したもので，かつ，以下の条件をすべて満たしたものとした．

・単語の文字数が 2 文字以上
・記事に出現する回数が 5 回以上
・品詞が名詞

形態素解析器にはMeCab [3] を使用した．訓練データと素性を使用し，SVMにより分類器を生成し，生成した分類器により株価に影響のある記事，そうでない記事に分類する．なお，SVMの実装として，*SVMlight* [4] を用いた．

４．予備実験と評価

予備実験として，SVMにより生成した分類器で分類した記事の評価を行う．評価方法は交差検定を用いた．今回は 10 企業の訓練データを用意したので，9 企業のデータを訓練データとし，残りの 1 企業をテストデータとしてSVMにより分類器を生成する．そして，その結果より精度と再現率を出す．これを 1 回ずつずらして 10 回行い，全ての精度と再現率の平均をとることで，推定精度と推定再現率を算出する．評価結果を以下の表 2 に示す．なお，表 2 の抽出記事数は，SVMによって株価に影響がある記事と分類された記事の数である．再現率は 49.7%，精度は 42.6%であり，十分な結果が得られなかった．精度が低い理由を次節で考察し，それに基づく手法の改良を示す．

4．1　考察

手法を改良する上で，なぜ十分な結果が得られなかったのかについて考察する．３章で述べた手法では，日立製作所に関する記事が株価に影響を与えるかどうかを分類するにも関わらず，日立製作所以外の記事を訓練データとして分類を行っている．しかしながら，これは，実際には全ての企業に関する記事を訓練データに加えることは不可能なので，公正な評価を行うために必要な処置である．そのため，選択された素性には日立製作所に関連しているものが含まれている可能性は低い．したがって，３章で述べた手法により生成した分類器では正しい分類ができなかったと考える．

５．手法の改良

３章での結果と考察を踏まえて，分類対象の企業に特化した名詞（例えば，自動車関連メーカであれば「エンジン」など）を素性として追加すれば，精度が向上すると考える．そこで，分類対象の企業と関連のあるキーワードが記載されているリストを用い，それぞれの素性に

[2] 株価の変動に関連しうるかどうかの判断は，投資歴が約１０年の投資家が行った．
[3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[4] http://svmlight.joachims.org/

表２　評価結果

証券コード	抽出記事数	正解記事数	誤った記事数	正解データの記事数	再現率(%)	精度(%)
6501	26	16	10	28	57.1	61.5
6502	39	11	27	20	55.0	28.2
6758	32	17	15	36	47.2	53.1
7201	83	47	36	81	58.0	56.6
7203	43	16	27	43	37.2	37.2
7267	41	15	26	19	78.9	36.5
7751	6	2	4	5	40.0	33.3
8001	17	9	8	20	45.0	52.9
9022	3	1	2	4	25.0	33.3
9432	50	11	39	36	30.5	22.0
合計	340	145	194	292	49.7	42.6

加える．しかし，企業ごとに，その企業と関連のあるキーワードを人手で作成することは困難である．そこで，企業WEBページに着目し，企業と関連のあるキーワードを企業WEBページから抽出する．

5．1　企業WEBページからのキーワード抽出

本節では，企業WEBページから，その企業と関連のあるキーワードを抽出する手法について述べる．まず，企業WEBページの各ページを形態素解析し，名詞のみを企業ごとに抽出する．しかし，多くの企業で使用されている単語は，その企業のキーワードとならないことが多い．例えば「こと」や「サイト」といった単語はどの企業のWEBページでも多く用いられている．そこで，IDF値を用いて多くの企業で用いられている名詞を除去した．IDF値の計算には以下の式を用いる．

$$IDF(t) = \log_2 \frac{N}{Df(t, N)} \qquad (2)$$

ここで，

N: 企業WEBページを取得した企業数（今回は 4077 企業）

$df(t, N)$: N個の企業における企業WEBページにおいて，単語 t を含む企業数

IDF値の閾値を 0.8 として，閾値以下のIDF値が付与される名詞を除外した．しかし，「印刷ページ」や「印刷画面」などの，明らかにキーワードとして不適切な名詞も企業のキーワードとして抽出していた．そのため，IDF値が 0.8 以下の名詞を含んでいる名詞も除外する．例えば，「ページ」という名詞はIDF値が 0.8 以下なので，「ページレイアウト」や「印刷ページ」といった単語が除去される．表 3 に，各企業のキーワードとして企業のWEBページから抽出された名詞の一部を示す．

表３　企業のWEBページから抽出されたキーワード

企業名	抽出されたキーワード
東芝	家電，テレビ，LED，電力，CMOS
味の素	スープ，レシピ，クノール，料理
三菱商事	産業，化学，金融，開発

5．2　訓練データの追加

３章では訓練データを 10 企業としていたが，訓練データに新たに 5 企業追加し，全部で 15 企業を訓練データとした．その結果，ポジティブな記事は 352 記事，ネガティブな記事は，ポジティブな記事と同数の 352 記事を無作為に選び，訓練データは 704 記事となった．新たに追加した企業を表 4 に示す．

表４　新たに訓練データとして選択した企業

企業名	証券コード
味の素	2802
東レ	3402
神戸製鋼所	5406
三菱商事	8058
東日本旅客鉄道	9020

5．3　企業WEBページより抽出したキーワードの追加

３章で提案した手法に，企業と関連のあるキーワードを素性として追加する．さらに，素性を形態素ユニグラムと形態素バイグラムに変更して，企業ごとに素性を作成し，SVMにより分類器を生成して，株価に影響を与える記事とそうでない記事に分類する．素性の選択方法は，図 3 に示すように，例えば「味の素」の記事を分類するための分類器を生成する場合は，味の素以外の記事に含まれる形態素ユニグラムと形態素バイグラムを素性とし，さらに「味の素」の企業WEBページから抽出したキーワードを素性として追加する．

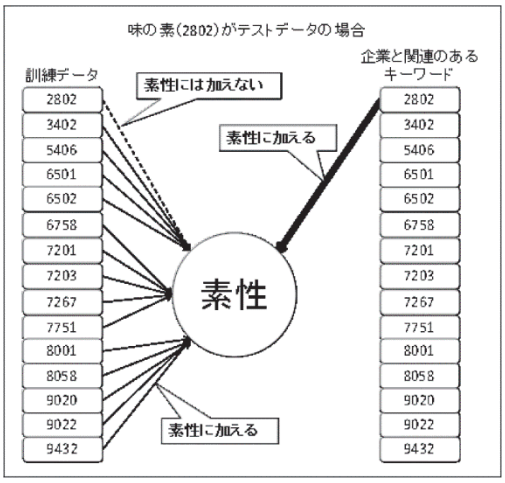

図 3　味の素をテストデータとした場合の素性選択方法の例

さらに，３章では，素性を“5 回以上出現した 2 文字以上の名詞”としていたが，改良手法では形態素ユニグラムと形態素バイグラムに変更した．例えば，「次世代太陽電池開発」という言葉の形態素ユニグラムと形態素バイグラムを取得すると以下のようになる．

形態素ユニグラム：「次世代」＋「太陽」＋「電池」＋「開発」

形態素バイグラム：「次世代太陽」＋「太陽電池」＋「電池開発」

5．4　条件付けによる絞り込み

予備実験の結果，誤りだった記事（False Positive）の表題に着目し，特徴を見つけることで条件付けにより除去を行う．結果をもとに以下の表現が，記事表題に含まれている場合に除去（すなわち，株価に影響を与える記事として分類しない）を行った．

解説，人事，特集，に聞く，社説

これらの語を除去した理由を以下に示す．コラムなどの引用した記事や解説の記事，さらに特集といった記事やインタビューの記事は，株価に影響を与えるような記事をもとに作られた記事であり，これが投資判断の材料になるとは言い難い．社説も同様に，通常の記事の背景を解説したり，解説者の主張や考えをまとめたものなので，株価に影響を与える記事ではないと考える．会社人事の記事は，新社長就任を記した記事であるが，これが株価に影響を与える可能性は皆無であると考える．以下にいくつか例を示す．

- 神戸経済特集．ベイエリアで環境事業、鉄鋼から発電へ、神鋼、来春２号機稼動．
- 日産自動車九州工場長山越敏行氏．日産再建計画の影響（この人に聞く）
- 世界自動車戦争の勝者と敗者は誰か（社説）

６．本実験と評価

改良した手法の評価を行う．評価方法は，４章と同様，交差検定を行った．結果を以下の表 5 に示す．再現率は 53.1%，精度 69.8%となり，改良前の手法に比べ，共に良い結果となった．

７．考察

本評価を行ったもののうち，日立製作所を例にとって，評価結果を考察をする．まず，株価変動に影響を与える重要な記事として適切に抽出できたものをいくつか以下に示す

- タービン用復水器生産、日立、新工場に集約．日立市に建設、工期を短縮．
- 日立、茨城新工場、台湾社と半導体合弁．次世代技術を活用．
- 日立、営業益６２％増、４ー９月、ＨＤＤが黒字に転換．
- 日立、２５６メガビットフラッシュメモリー、来年中に生産６倍に．
- 日立、グローバル化を促進、海外生産年２５％増．輸入拡大へ奨励金制度．
- エアコン用コンプレッサー、日立、中国で増産．能力４割増強．
- 日立、初の赤字．今期経常損益、１０００億円、半導体不振．

日立製作所の場合，精度は 77%であり，比較的良好な結果を得ることができた．次に，本手法では重要な記事として抽出したもののうち，不適切だったものを以下に示す．

- スターエンジニアリング（茨城・日立市）．産学交流で生ごみ処理機（挑む地場企業）
- 久慈鉄工協組、メロン水耕栽培装置を共同開発．得意技術を持ち寄る．
- 崩れる企業集団（３）投資価値で子会社選別．「モノ言う本社」へ脱皮．

上に挙げた記事は，日立製作所に関連している記事ではあるものの，それが株価に影響しているのかが判断できるものではないと考える．また，日立製作所と関連が薄い記事や，コラムを抽出してしまっていた．

本手法の再現率は 50%であった．本手法で抽出できな

表５　評価結果（本実験）

証券コード	抽出記事数	正解記事数	誤った記事数	正解データの記事数	再現率(%)	精度(%)
2802	4	3	1	11	27.2	75.0
3402	3	3	0	4	75.0	100
5406	11	10	1	29	34.4	90.9
6501	18	14	4	30	46.6	77.7
6502	10	7	3	13	53.8	70.0
6758	53	35	18	52	67.3	66.0
7201	54	37	17	66	56.0	68.5
7203	28	23	5	48	47.9	82.1
7267	18	14	4	20	70.0	77.7
7751	6	3	3	7	42.8	50.0
8001	30	14	16	24	58.3	46.6
8058	18	14	4	24	58.3	77.7
9020	2	2	0	4	50.0	100.0
9022	1	1	0	5	20.0	100.0
9432	12	7	5	15	46.6	58.3
合計	268	187	81	352	53.1	69.8

表６　証券コードと業種の対応

証券コード	業種	証券コード	業種
1300 番台	水産・農林業	7000～7400 番台	輸送用機器
1500 番台	鉱業	7700 番台	精密機器
1600 番台	鉱業（石油・ガス開発）	7800～7900 番台	その他製品
1700～1900 番台	建設業	8000～8200 番台	卸売業
2000 番台	食料品	8300～8500 番台	銀行・その他金融
3000～3500 番台	繊維製品	8600 番台	証券・先物取引業
3700～3900 番台	パルプ・紙	8700 番台	保険
4000 番台	化学・医薬品	8800 番台	不動産
5000 番台	石油・石炭製品	9000 番台	陸運
5100 番台	ゴム製品	9100 番台	海運
5200～5300 番台	ガラス・土石製品	9200 番台	空運
5400～5600 番台	鉄鋼	9300 番台	倉庫・運輸関連
5700～5800 番台	非鉄金属	9400 番台	情報通信
5900 番台	金属製品	9500 番台	電気ガス
6000～6400 番台	機械	9600～9900 番台	サービス業
6500～6900 番台	電気機器		

かった重要な記事（False Negative）をいくつか以下に示す.

- ＤＮＡチップで遺伝子回収、日立など新技術開発.
- 日立、高速光伝送システム、米社から４８億円受注.
- 日立・東海大チーム判別法を開発、ＡＬＳ患者の意思、脳の血流で読む.
- がん検査の解像度２倍、日立、北大と装置.
- 遺伝子試料を自動作製、日立基礎研が装置開発.
- 電力事業７関連会社、３社に統合、９００人削減. 日立、固定費９０億円圧縮.

上の結果を見ると，本来は株価に影響していると分類すべき記事である「大型受注」や「新技術開発」に関する記事が多数あった．再現率を上げるためにも，これらの記事を正しく抽出する必要がある.

生成された分類器を解析し，各素性の寄与度を調査した. 例えば,「がん検査の解像度２倍、日立、北大と装置.」という記事に含まれていると思われる「医療」という単語を調べると，全 56208 語中 443 番目にマイナス方向に寄与していた．これは，今回準備した 15 企業のうち，日立製作所を除いた 14 企業には「医療」という単語が出現する頻度が少なく，訓練データにおけるポジティブの記事に含まれている数が少なかったからであると考える．これを解決するには，訓練データを準備する段階で，業種に偏りのない企業を選択することで解決できると考える．また，証券コードはある程度，業種により分類されているので，それをうまく利用することで業種による偏りをなくすことができると考える．証券コードは，通常は以下の表 6 のように分類されている．したがって，各業種の企業を 1 企業でも含めば，偏りのない訓練データができ素性もより良いものが選択できると考える．その結果，SVMによる分類器も良いものが生成され，精度，再現率ともに向上すると考える.

今回は新聞記事が株価に影響を与えるか否かをSVMによる分類器で分類したが，個人投資家はその記事がどのような要因によるものかを知りたいと考える．したがって，今回の結果をさらに文書分類手法により分類すると，各記事がどのような要因によるものか分類できる．例えば，以下のようになる．

- 日立、２６００億円の赤字．今期最終損失、中間配当見送り．→　業績発表
- 日立、茨城新工場、台湾社と半導体合弁．次世代技術を活用．→　業務提携
- 新タイプの高温超電導物質、日立が発見、「非銅系」物質を合成．→　新技術開発
- ボーイング７Ｅ７炭素素材、東レ、３３００億円独占受注、１８年契約．→　大型契約
- ＮＴＴ汚職、飲食など十数回接待．牧容疑者、業者に要求か．→　不祥事

上に挙げた分類のほかにも,「事業拡大（縮小）」,「事故」などがある．

さらに，本来ならば日経平均との兼ね合いも考慮するべきであると考える．日経平均が上がったときに，対象企業の株価も上がっていた場合には，大きな出来事が起きたとは考えにくい．反対に，日経平均が上がっているにも関わらず, 対象企業の株価が下がっている場合には, 何かしらの出来事が起きたと考えるべきである. つまり, 日経平均の変動と反対の変動をしているときには，何かしらの出来事をしていると考えることができる．日経平均との兼ね合いを考慮することで，より重要な記事のみの抽出ができると考える．

8. むすび

本研究では，ある企業の株価に影響を与えるような，その企業に関連する日本経済新聞記事を自動的に抽出する手法を提案した．単に日経新聞記事内の単語を素性とするのではなく，形態素ユニグラムと形態素バイグラムを素性とし，各企業のWEBページから抽出したキーワードを素性に加えることで，各企業に特化した素性や分類器をSVMにより生成し，分類を行うことができた．評価の結果，精度は 69.8%，再現率は 53.1%となり，比較的良好な結果を得ることができた．今後の課題として，精度がある程度確立できたので，再現率の向上を目指す．また，文書分類手法などにより，記事をさらに要因ごとに分類することも行う．文書分類手法の段階に進むためには，記事数の増加，すなわち再現率の向上が必要不可欠であると考える．さらに，日経平均との兼ね合いも考慮することを必要であると考える．

謝辞

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) 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の推定，情報処理学会論文誌，Vol. 52, No. 12, pp. 3309‑3315 (2011).
2) Koppel, M. and Shtrimberg, I.: Good News or Bad News? Let the Market Decide, In Proceed- ings of the AAAI Spring Symposium on Exploring At titude and Affect in Text, pp. 86–88 (2004).
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) Jensen, D. and Allan, J.: Mining of Concurrent Text and Time Series, In Proceedings of the KDD 2000 Conference Text Mining Workshop, pp. 37–44 (2000).
5) Milea, V., Sharef, N. M., Almeida, R. J., Kaymak, U. and Frasincer, F.: Prediction of the MSCI EURO index based on fuzzy gram- mar fragments extracted from European Cen- tral Bank statements, International Conference of Soft Computing and Pattern Recognition, pp.231–236 (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) 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).
9) Sakaji, H., Sakai, H. and Masuyama, S.: Automatic Extraction of Basis Expressions That Indicate Economic Trends, Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp.977–984 (2008).
10) Vapnik, V.: Statistical Learning Theory, Wiley (1999).