日次ニュース業界別記事抽出による株価変動予測

Stock Fluctuation Forecasting Based on Text Mining of Daily Sector News

本多隆虎[1*] 和泉潔[2] 松井藤五郎[3] 吉田稔[4]
中川裕志[4] 石田智也[5] 中嶋啓浩[5] 菅原俊治[1]
Takatora HONDA[1] Kiyoshi IZUMI[2] Togoroh MATSUI[3]
Minoru YOSHIDA[4] Hiroshi NAKAGAWA[4] Tomonari ISHIDA[4]
Akihiro NAKASHIMA[4] Toshiharu SUGAWARA[1]

[1] 早稲田大学大学院 基幹理工学研究科
[1] Fundamental Science and Engineering, Waseda University
[2] 東京大学大学院 工学系研究科 & JST さきがけ
[2] School of Engineering, The University of Tokyo & PRESTO, JST
[3] 中部大学 生命健康科学部
[3] College of Life and Health Sciences, Chubu University
[4] 東京大学 情報基盤センター
[4] Information Technology Center, The University of Tokyo
[5] 野村證券株式会社
[5] Nomura Securities co.,LTD.

Abstract: In recent years, a number of researches are conducted to analyze and forecast stock values in the area of artificial intelligence. The most researches use monthly news and focus on the stock value fluctuation forecast on a monthly basis, but financial traders usually require the daily forecast of stock market prices. To forecast daily change of stock market, sector news should be taken into consideration because we think that they impact on stock market. The primary objective of our research is to investigate whether or not next day's stock values can be forecasted using text mining of the daily sector news. In our method, sentences in the sector news are resolved into the morphemes by morpheme analysis and the co-occurrence frequency are counted. Then, the derived frequency data are transformed to principal components. Finally, the relationships between stock value fluctuations and the principal components in the multiple linear regression are examined. To clarify the effectiveness of our method, we compare the forecating obtained by the daily sector news with those done by daily whole news. The experimental results shows that the accuracy of the forecasts with daily sector news is higher than that with daily whole news in the period of the big fluctuation in the sector.

1 はじめに

近年、経済市場を分析するコンテンツ数は増加傾向にある。例えば、金融経済専門のオンラインニュースや、金融機関の世界経済に関するレポートがインターネットに日々アップロードされている。その中で、株式取引を仕事とするトレーダーは多くの情報の中から市場に影響を及ぼす可能性のある情報を取捨選択して、取引の判断基準としている。トレーダーの使う指標には大きく分けて以下の 2 種類である。

- 経済指標、マーケットのテクニカル指標等の数値情報
- 市場に影響力を及ぼす要人の発言や企業の製品/商品や業績発表等のテキスト情報

*連絡先：早稲田大学大学院基幹理工学研究科
〒 169-8555 東京都 新宿区大久保 3-4-1
早稲田大学 基幹理工学研究科
情報理工学科 55N-0502B
E-mail: t.honda@isl.cs.waseda.ac.jp

特に、後者は、インターネットの普及により爆発的に情報量を増やしており、トレーダーが、瞬時にして市場に関する全ての数値情報や、テキスト情報等の非数値情報を把握するのは困難である。そこで、株式取引に有益な情報を抽出するため、情報技術をファイナンスの分野に組み込む研究が盛んに行われている。

例えば、人工市場に株価の数値情報を用いて、人工市場の上で株価データの変動を再現し、投資家の株価参照頻度を推計している研究 [1] がある。この結果株価参照頻度が高まると、株価変動の 1 次の自己相関における係数が大きくなるという結果が得られており、株価の上昇トレンドが維持され易い結果を得ている。また、[2] では、ニューラルネットや遺伝的アルゴリズム、強化学習やファジィニューロを数値情報による市場分析に応用した研究が紹介されている。[3] では価格ではなく、全トレーダーの注文を表にした「板」を情報として使用している。板情報の注文数から特徴量を抽出し人工市場と実市場の比較をすることが、二つの市場間の類似性を評価する方法として有効であることが確認している。しかし、[1]〜[3] では数値情報のみを扱っており、テキスト情報等の非数値情報は解析対象ではない。

テキスト情報等の非数値化情報を数値化し解析した研究は、日本銀行の発行する金融経済月報を分析し、実際の市場動向を予測したものがある [4]。ここでは、日本銀行の金融経済月報をテキストマイニングによって日本国債一年物、二年物、五年物、十年物の月末価格の上昇下落を予測している。解析では形態素解析、共起解析、主成分分析、重回帰分析を用いている。その結果、各市場を予測した予測的中精度は日本国債一年物で 89.0%、日本国債二年物で 84.9%、日本国債五年物で 73.5%であった。これらの結果より [4] の手法は金融経済月報から市場金利の動向を十分に予測できている。

月次テキストの代わりに、日次のテキストを用いて株価との相関を調べた研究には [5] がある。この研究では、新聞記事の株価変動への影響分析と株価変動の外部要因分析という相補的な視点から株価と新聞記事の関連性の概要抽出を行っている。結果、いくつかの銘柄に関する株価データと新聞記事の関連性を実証している。これら [4]、[5] の研究では新聞や経済分析記事と株価の相関の検証を行い有効性を確認しているが、日次株価変動は予測対象としていない。

日次株価のような短期的な市場変化は世界経済の変化、為替の値動き、産業特有のニュース等に大きく影響されやすく、それらを考慮した株価変動予測が必要である。そこで我々は、新聞記事を産業業種別、全業種でテキストマイニングし、各産業に特化した回帰式の各業界別の株価変動予測への有効性を検証した。

2 提案手法

本研究では、[4] で市場予測への有効性の確認されている形態素解析、主成分分析、重回帰分析の手法を応用した手法を用いる。形態素解析の前準備として新聞記事から産業業種関連記事のみ取りだし、そこから業種別平均株価との相関の高い記事を抽出し、株価変動を予測する。

本研究で使用したデータは日本経済新聞デジタルテキスト版 2005 年 1 月 1 日〜2008 年 12 月 31 日（ただし祝日、休刊日を除く）の期間の記事を使用した。これらの新聞記事は、通常発刊されている紙面の新聞記事からテレビ番組表、株価一覧、地域面の記事が削除されている。

株価データには野村證券株式会社の提供するインデックス値 NOMURA400[1] を使用した。NOMURA400 は、日本の株式市場の全銘柄の中から選定した上場企業 400 社を数値化した時価総額加重平均の株価指数である。NOMURA400 は 21 業種に分けられているが、本研究では、証券アナリストが特に新聞記事の影響を受けやすいと感じる 3 業種 (食品、電機・精密、メディア) のリターン値の変動を予測する。株価変動率だけでは市場全体の動きの影響を受けて株価が上下する恐れがあるため、本研究では株価変動にいかに定義する NOMURA400 業種別超過リターン値を使用した。

2.1 データ準備

2.1.1 インデックスの超過リターン変換

最初に、NOMURA400 の各業種リターン r_{ij} を次式により算出する。

$$r_{ij} = \frac{p_i(s) - p_j(s)}{p_j(s)}$$

但し、r はリターン値、i,j は日付、s は業種、$p_i(s)$ は日付 i における業種 s の NOMURA400 のインデックス値を表す。

純粋な業種別の株価変動を予測できているか検証するために市場を考慮した超過リターン r'_{ij} を次式により算出する。

$$r'_{ij} = r_{ij} - R_{ij}$$

式中の R_{ij} は東証株価指数 (TOPIX) のリターンを表している。実際のリターン r_{ij} 値は、− 1〜1 の値を取り、TOPIX のリターンは − 0.03〜0.03 の値を取るこ

[1] NOMURA400 は、野村證券株式会社が公表している指数で、その知的財産権は野村證券株式会社に帰属します。なお、野村證券株式会社は、対象インデックスの正確性、信頼性、有用性を保証するものではなく、対象インデックスを用いて行われる事業活動・サービスに関し一切責任を負いません。

とから、超過リターンは業界別リターンに市場動向の微調整を与えた値である。以下、本研究で指す各業界リターン値とは各業界超過リターン値とする。

2.1.2 関連記事抽出

本研究では、企業が関連する記事の最初の文章に、その企業の正式名称が掲載されるという特徴を利用し、新聞全記事より業種別関連記事を抽出する。具体的にはNOMURA400の各業種に含まれる企業の正式名称が、一行目に掲載されている記事のみを抽出し、解析対象とする。この作業により、関連企業の記事だけを抽出できるだけでなく、新聞の文化面、教育面等の株価に影響が薄いと思われるノイズデータを除外できる。

2.2 形態素解析

テキストデータを日本語形態素解析システムを用いることにより形態素に分解する。形態素とは、意味を持つ最小の単位のことである。本研究では、形態素解析を行った後、助詞等の付属語を除いた名詞・動詞・形容詞等を解析対象として抽出している。

2.3 共起解析

「共起」とは、ある二つの言語表現（単語等）が一定範囲（一文、段落等）で同時に出現することであり、検索エンジンや自然言語処理の分野等に利用されている。本研究では形態素解析によって分解された各単語間の共起関係を調べる。具体的には句点や空白によって区切られている文毎に共起解析をしており、同文中に隣接出現する二語は一回共起したと考える。

本研究ではまず、株価との相関が薄いと思われる一般語を共起対象から排除した。このために共起している2語のいずれかが単語辞書に掲載されている単語であることを共起判定の条件とした。単語辞書として日本経済新聞デジタルメディアが1982年から作成している新聞記事検索のための用語集「日経シソーラス」を用いた。

2.4 主成分分析

解析期間の全てのテキストを対象に共起解析を行い、そこに出現した単語の一日毎の出現状況に対し主成分分析を行い，約100個の合成変数（主成分）にまとめる。各日の主成分スコアを、2005年から2008年までについて時系列順に並べることによって、共起頻度を低次元で再現した時系列データが作成できる。これは分析対象期間のテキストデータの特徴の時系列変化を表している。線形重回帰分析の計算処理量を減らすために、主成分分析フェイズにより数千語に及ぶ共起回数の情報を一定の次元まで落とす。

また本研究では単語によって共起数が大きく違うため、共起数を標準化するために相関係数行列から主成分を求める手法を用いた。

2.5 重回帰分析

重回帰分析により、各主成分スコアの各日の動きから日次での市場価格の動きを解析し、予測対象日の株価値を予測する。具体的には第2.7節の100個の主成分スコアをAIC基準によるステップワイズ変数選択した後、主成分スコアの時系列データを説明変数として、日次の株価データを被説明変数とする線形重回帰分析を行う。AIC基準によるステップワイズ変数選択とは、回帰式作成時に予測や判別に有用な独立変数を変数増減法によって選択する変数選択法である。本研究ではAIC基準を用いることで、変数選択時に残差平方和(あてはめ誤差)の大小だけでなくモデルに含まれるパラメータ数を考慮し、ノイズデータ等の偶発的な変動に過適合しない回帰式を作成している。

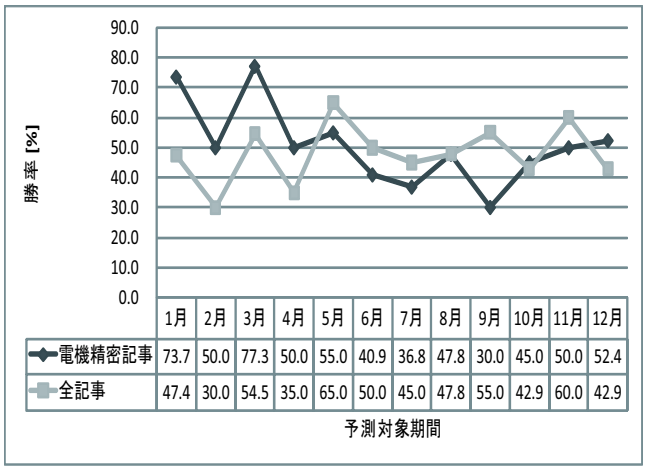
3 実験

前節の手法を用いて、2006年～2008年の株価変動予測を行った。ここで株価変動予測とは対象日前日比リターンの上昇下落を予測することをさす。予測対象月での予測成功割合を勝率と定義する。本研究では予測対象月の前一年間を回帰式作成の訓練期間とし、翌一カ月を予測している。実験結果の各月勝率は、一カ月間の勝率平均値である。

共起頻度の低い一般語を取り除くために最低共起頻度を4とし、共起回数4回未満のものは除外した。共起回数を0とした。形態素解析ではフリーツール“Chasen”[6]を使用し、主成分分析、線形重回帰分析を行うツールとして統計ソフト“R”[7]を使用した。

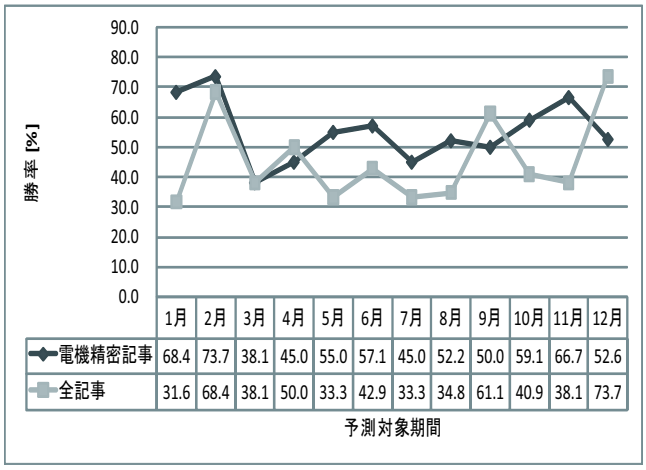
図1-図9に2006年-2008年電機・精密業界、食品業界、メディア業界予測結果を示す。

関連記事抽出と全記事抽出の勝率を比較すると、一方の勝率が連続して上回っている期間が多いことが分かる。関連記事抽出に限ると、2006年電機精密業界1月～4月、2007年電機精密業界5月～8月、2006年食品業界2月～5月、2006年メディア業界1月～5月の勝率が連続して全記事予測値を連続して上回っている。各期間の主要ニュースを調べると2006年電機精密業界1月～4月では、シャープの液晶テレビ生産拡大やソニーの2005年クリスマス商戦での好調が要因となり、電機精密業界全体の株価が上昇している。2007年電機



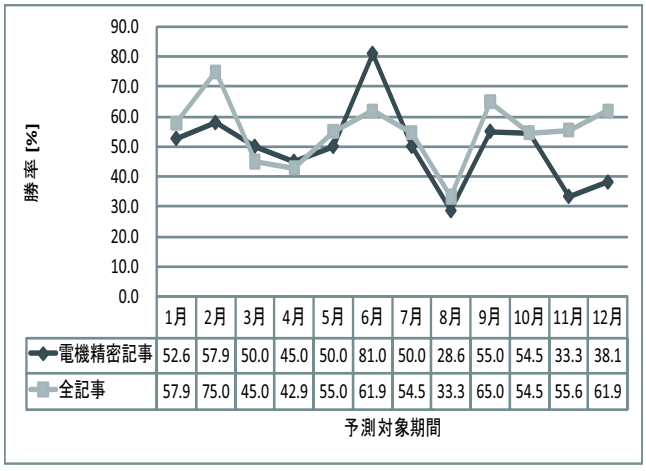
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
電機精密記事	73.7	50.0	77.3	50.0	55.0	40.9	36.8	47.8	30.0	45.0	50.0	52.4
全記事	47.4	30.0	54.5	35.0	65.0	50.0	45.0	47.8	55.0	42.9	60.0	42.9

図 1: 2006 年 電機・精密業界予測 勝率



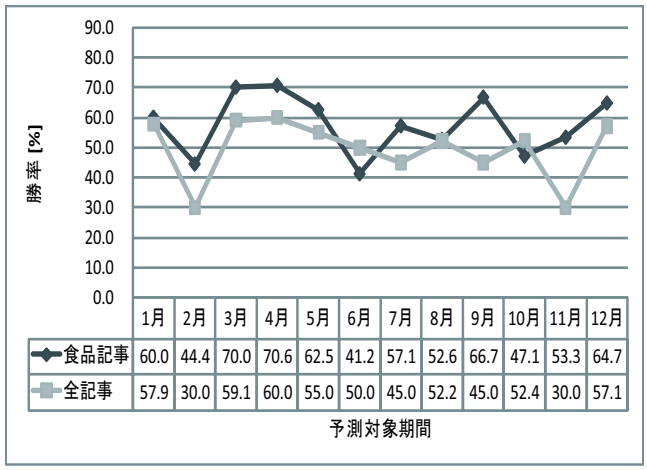
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
電機精密記事	68.4	73.7	38.1	45.0	55.0	57.1	45.0	52.2	50.0	59.1	66.7	52.6
全記事	31.6	68.4	38.1	50.0	33.3	42.9	33.3	34.8	61.1	40.9	38.1	73.7

図 2: 2007 年 電機・精密業界予測 勝率



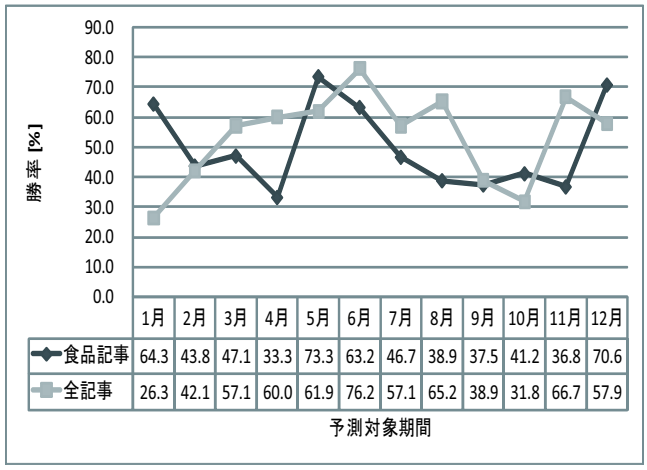
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
電機精密記事	52.6	57.9	50.0	45.0	50.0	81.0	50.0	28.6	55.0	54.5	33.3	38.1
全記事	57.9	75.0	45.0	42.9	55.0	61.9	54.5	33.3	65.0	54.5	55.6	61.9

図 3: 2008 年 電機・精密業界予測 勝率



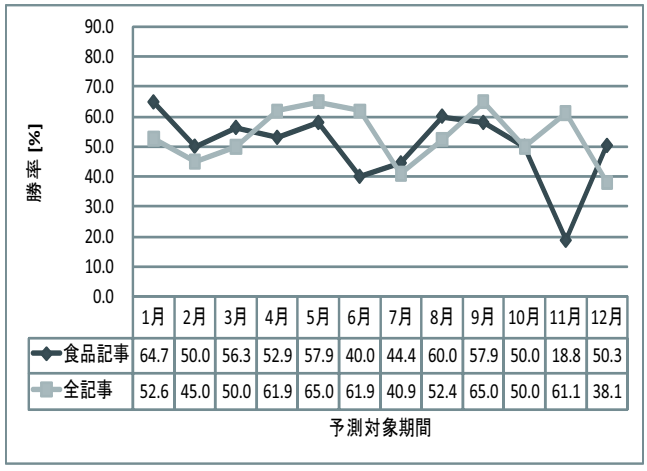
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
食品記事	60.0	44.4	70.0	70.6	62.5	41.2	57.1	52.6	66.7	47.1	53.3	64.7
全記事	57.9	30.0	59.1	60.0	55.0	50.0	45.0	52.2	45.0	52.4	30.0	57.1

図 4: 2006 年 食品業界予測 勝率



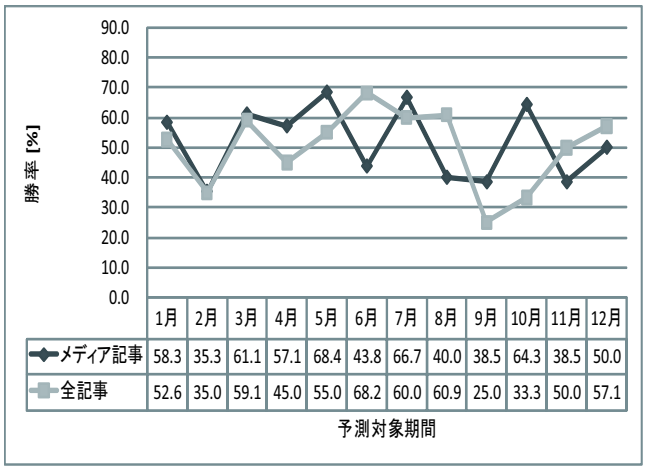
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
食品記事	64.3	43.8	47.1	33.3	73.3	63.2	46.7	38.9	37.5	41.2	36.8	70.6
全記事	26.3	42.1	57.1	60.0	61.9	76.2	57.1	65.2	38.9	31.8	66.7	57.9

図 5: 2007 年 食品業界予測 勝率



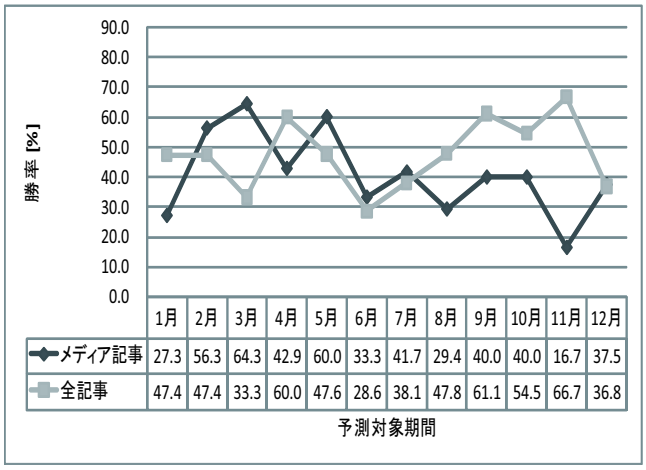
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
食品記事	64.7	50.0	56.3	52.9	57.9	40.0	44.4	60.0	57.9	50.0	18.8	50.3
全記事	52.6	45.0	50.0	61.9	65.0	61.9	40.9	52.4	65.0	50.0	61.1	38.1

図 6: 2008 年 食品業界予測 勝率



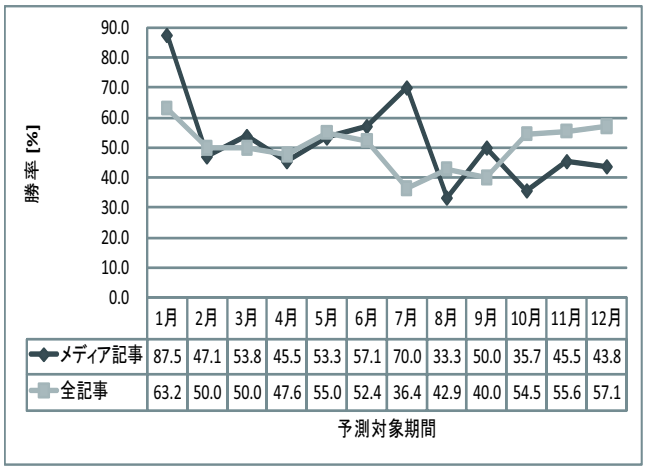
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
メディア記事	58.3	35.3	61.1	57.1	68.4	43.8	66.7	40.0	38.5	64.3	38.5	50.0
全記事	52.6	35.0	59.1	45.0	55.0	68.2	60.0	60.9	25.0	33.3	50.0	57.1

図 7: 2006 年 メディア業界予測 勝率



	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
メディア記事	27.3	56.3	64.3	42.9	60.0	33.3	41.7	29.4	40.0	40.0	16.7	37.5
全記事	47.4	47.4	33.3	60.0	47.6	28.6	38.1	47.8	61.1	54.5	66.7	36.8

図 8: 2007 年 メディア業界予測 勝率



	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
メディア記事	87.5	47.1	53.8	45.5	53.3	57.1	70.0	33.3	50.0	35.7	45.5	43.8
全記事	63.2	50.0	50.0	47.6	55.0	52.4	36.4	42.9	40.0	54.5	55.6	57.1

図 9: 2008 年 メディア業界予測 勝率

精密業界 5 月～8 月はソニー、東芝、シャープ等の主要電機メーカーの液晶テレビが好調であった他、次世代ディスクのブルーレイと HD-DVD の規格争いの山場であった。また、2006 年食品業界 2 月～5 月には、味の素、ヤクルト、アサヒビールの経営不振の発表等株価にマイナス影響の大きいニュースが多く、2006 年メディア業界 1 月～5 月は 2006 年 1 月 16 日にライブドア本社に強制捜査が入り、メディア業界に影響があったと考えられる。

実際のリターン値を見ても関連記事抽出の勝率が高い期間は株価変動の大きい日が多い。例として 2006 年電機精密業界各月のリターン値の絶対値の平均を表 1 に示す。2006 年電機精密業界 1 月～4 月の全記事予測より予測精度の高い期間は、それ以外の期間と比較して明らかにリターン値の絶対値平均が多い。これより、業界の株価変動の大きい期間については予測精度を業種別記事抽出によって高めることができたといえ、業界別記事抽出が業界別株価変動予測に有効であることが確認できた。

表 1: 2006 年電機精密業界各月リターン絶対値平均

月	日数	月	日数
1 月	0.81	7 月	0.38
2 月	0.49	8 月	0.31
3 月	0.31	9 月	0.28
4 月	0.59	10 月	0.33
5 月	0.32	11 月	0.41
6 月	0.30	12 月	0.31

逆に、業種別記事抽出の勝率が悪い期間を見てみると、3 業種とも 2008 年 9 月～12 月の勝率が全記事と比較して低い傾向が見られる。この期間は、市場全体の動きに株価全体が引っ張られた可能性が考えられ、経済全体の記事が有用に働いている期間と考えられる。この理由としては、2008 年 9 月 15 日にリーマンブラザーズ破産法適用申請があり、リーマンショックが各業界に影響を与えていると考えられる。

4 まとめ

本研究では、日次ニュースの業界別記事抽出を用いた株価変動予測の有効性について検証した。本手法により、株価変動の大きい期間では全記事を解析して回帰式を作成するよりも精度の高い株価変動予測が出来ることが確認できた。逆に、リーマンショック後のように市場全体が大きく動いている期間は全記事抽出予測の精度が高いことも確認できた。

今後は、アンサンブル学習によって、業界別記事抽出と全記事抽出予測を選択し、市場動向にあった予測を試みる予定である。また、株価と相関の高い語を抽出することで決定係数の向上を行う。

参考文献

[1] 加納 良樹, 寺野 隆雄,“人工市場による株価参照頻度の分析,” 情報処理学会論文誌, 2006.05.15

[2] 電気学会, “ニューラルネット・遺伝アルゴリズムの金融工学への応用,” 森北出版, 学習とそのアルゴリズム, 2002

[3] 西岡 寛兼, 鳥海 不二夫, 石井 健一郎,“板情報を用いた株式市場の時系列データの分析法”, 第二十三回人工知能学会全国大会報告, 2009.06.19

[4] 和泉 潔, 松井 藤五郎, 後藤 卓 “テキスト情報による金融市場変動の要因分析,” 第二十三回人工知能学会全国大会報告, 2009.06.19

[5] 小川 知也, 渡部 勇,“株価データと新聞記事からのマイニング,” 情報学基礎研究会報告 2001(20), p137-p144, 2001-03-05

[6] Chasen.http://chasen.naist.jp/hiki/ChaSen/.

[7] 統計ソフト R.http://www.r-project.org/.