情報処理学会論文誌 Vol. 52 No. 12 3309–3315 (Dec. 2011)

経済テキスト情報を用いた長期的な市場動向推定

和泉 潔[†1,†2] 後藤 卓[†3] 松井藤五郎[†4]

本研究は，金融実務家から要望が高い，数週間以上の長期的でしかも個別銘柄より広範な市場分析に，テキストマイニング技術で挑戦した．長期市場分析に有効なテキスト情報として，経済の専門家や金融機関が Web 上に発行するマーケットリポートを分析対象とした．そのために，定期的に発行されるテキストデータから時間的な特徴の変化を抽出し，テキストに関連する外部の時系列データとの関係性を見つけるテキストマイニング技術を新たに開発した．本技術を用いて実際に経済市場分析を試み，実際の市場動向をどの程度説明しているのかについて検証を行った．日本国債の 2 年物，5 年物，10 年物で運用テストを行った結果，既存のサポートベクタ回帰や計量経済モデルと比べて，どの市場でも安定して，ほぼ最高水準の運用益をあげることができた．

Long-term Financial Market Analysis Using Economic Textual Information

Kiyoshi Izumi,[†1,†2] Takashi Goto[†3] and Tohgoroh Matsui[†4]

In this study, we proposed a new text-mining methods for long-term market analysis. Using our method, we perfomed out-of-sample test using monthly price data of financial markets; Japanese government bond 2-year, 5-year, and 10-year markets. First we extracted feature vectors from monthly reports of Bank of Japan. Then, trends of each market were estimated by regression analysis using the feature vectors. As a result of comparison with support vector regression and an econometric model, the proposal method could forecast in higher accuracy about both the level and direction of long-term market trends.

1. はじめに

近年，機械学習を用いたテキストマイニング手法によって，テキスト情報と市場変動の関係性を発見し市場分析に応用する研究が増えてきた．経済指標やマーケットのテクニカル指標などの数値情報には指標化されていないような情報を，テキスト情報から素早く自動的に抽出することが期待されている．しかし，既存の研究は，数分から 24 時間以内の短期的な市場の反応を分析対象とし，特定のキーワードがもたらす市場への単発的なインパクトのみを扱っていることが多かった．そこで本研究は，金融実務家から要望が高い，数週間以上の長期的でしかも個別銘柄より広範な市場分析に，CPR 法[1] と呼ばれるテキストマイニング技術を用いて挑戦した．同じテキストデータを用いた既存のテキストマイニング手法に加えて，経済指標分析や時系列データ分析などの経済分野における既存の分析手法とも，運用テストや変動予測精度の比較を行った．それにより，テキストマイニング手法を長期的な市場分析に用いることの優位性を検証した．

2. 金融テキストマイニング研究の概観

金融テキストマイニング研究は，どのような種類のテキスト情報を分析するかによって分類できる．さらに，分析対象となるテキストは，内容や書き手の多様性/専門性の軸によって整理できる（**図 1**）．

twitter やブログなどは一番多様な内容と書き手を持つテキスト情報である．多くの書き手は経済の専門家ではなく，書かれている内容も日常的な事柄も含む非常に多様で統一性のないものである．しかし，膨大な量のテキスト情報を集めることが可能である．twitter のテキストデータから翌日のダウ・ジョーンズ工業株価平均の変動を予測する研究[2] がある．

オンラインの経済ニュースや掲示板などは，もう少し専門的なテキスト情報である．金融テキストマイニング研究で一番多く分析されているのが，ロイターや Bloomberg などのオ

†1 東京大学
The University of Tokyo
†2 科学技術振興機構さきがけ研究 21
PRESTO, Japan Science and Technology Agency
†3 三菱東京 UFJ 銀行
Bank of Tokyo-Mitsubishi UFJ, Ltd.
†4 中部大学
Chubu University
⋆1 本稿の内容は三菱東京 UFJ 銀行の公式見解を示すものではない．

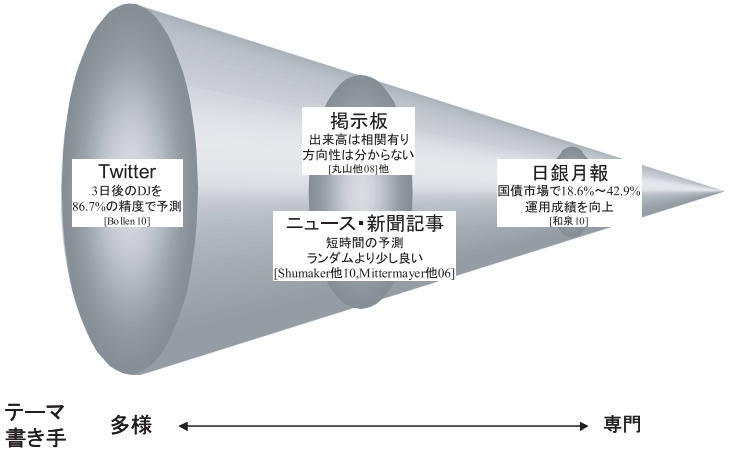


図 1　金融テキストマイニング研究の概観
Fig. 1　Framework of financial text mining researches.

ンライン上の経済ニュースのテキストである[3)–5)]．記事内容は経済に関連にありそうな事項であり，書き手も経済の専門家とは限らないが記者という人たちに限定されている．金融機関のトレーダたちもこういった記事を取引時の参考にしている．また，Web 上には，個別の株式銘柄に関する情報を書き込む掲示板サイトがある．書き手はその銘柄に興味がある人たちであり，書かれている内容も一応その銘柄に関連する事柄に限定されている．個別銘柄に関する掲示板データから翌日の株価リターンや出来高を予測する研究もある[6),7)]．

本研究では，より専門的なテキスト情報である，金融機関の発行する経済リポートを分析対象とした金融テキストマイニング手法[1),8)] を用いる．経済の専門家たちが他の専門家や投資家に市況を解説するために書いたテキストである．内容はもちろん市場に直接関係する事柄だけである．内容が専門的あるだけでなく，テキストの様式や言葉使いも，ある程度の統一性を持っている．本研究は，定期的に発行され形式も定まった経済リポートから，テキストの特徴の時間変化を抽出し，月次以上の長期的な価格時系列データの変動との関係性を発見することを目的とする．

3. テキストデータによる長期市場分析手法

本研究では長期的な市場分析に有効なテキストデータとして，日本銀行の金融経済月報を選んだ．金融経済月報は，日本銀行が金融・経済情勢を分析した資料であり，毎月半ばに，A4 で 15–20 ページの分量で公開されている[⋆1]．本テキスト情報が長期分析に有効な理由の 1 つは，書かれている内容に日本銀行の経済状況に対する態度が含まれているとされ，実際の金融市場のトレーダの多くが着目している共有の重要テキスト情報であることである．さらに，テキストの形式について，解説内容の順番や段落構成や表現について一貫性を保つように記述されていて，異なる時点間のテキスト内容の変化を比較しやすいからである．

しかし，4.3 節に後述するように，従来の金融テキストマイニング研究と同様に，このテキストデータから計算した単独のキーワードの頻度またはキーワードの組合せの共起頻度をそのまま入力して学習させても，将来の市場動向を予測することは難しかった．そこで，本研究ではある程度の一貫性を保ちながら定期的に発行されるテキスト情報を用いて，時間的な出現パターンから単語のグループ化を行った．これにより，比較的少ない次元数の特徴量を抽出し，外部の時系列データを説明する．そのために，共起解析（co-occurrence analysis）と主成分分析（principal component analysis），回帰分析（regression analysis）のステップからなる **CPR 法**を提案する．

3.1　共起関係に基づく主要単語の抽出（C）

最初に，各月 t のテキストデータ $D(t)$ の特徴を表す主要単語を抽出する．本研究では，通常の高頻度単語のほかに，KeyGraph アルゴリズム[9)] に基づいて抽出された，概念間を結ぶ橋となるキーワードも各月の主要単語に含めた．

(1)　高頻度単語の抽出：$D(t)$ 中の名詞・動詞・形容詞のうち，頻度が上位 M 個の高頻度語の集合 $HighFreq = \{w_i\}, i = 1, \cdots, M$ を抽出する[⋆2]．

(2)　橋となるキーワードの抽出：まず $HighFreq$ 内の単語 w_i，w_j 間の共起度 $co(w_i, w_j)$ を Jaccard 係数[⋆3]で計算し，上位 L 個の単語間をリンクで結ぶ[⋆4]．

$$co(w_i, w_j) = \frac{D(t) \text{ のうち } w_i, w_j \text{が共に出現する段落数}}{w_i,\ w_j \text{ の少なくとも一方が出現する段落数}} \tag{1}$$

単一のパスのみで接続される単語間のリンクを削除し，単語のクラスタ（土台）を抽出する．最後に，$D(t)$ 中のすべての語 w に対して，すべての土台 g が考慮されたと

⋆1 テキストデータは http://www.boj.or.jp/theme/seisaku/handan/gp/で毎月公開されている．
⋆2 本研究は $M = 30$ とした．
⋆3 金融経済月報では各段落が，経常収支や国内企業活動などの各テーマの完結した解説文章の単位となっている．そのため，本研究では各段落を共起範囲として Jaccard 係数を計算した．
⋆4 本研究は $L = 30$ とした．

きに語 w が用いられる条件付き確率 $key(w)$ を計算し，上位 N 個を土台（概念）間を結ぶ橋となるキーワード $HighKey$ として主要単語に加える[⋆1]．

$$key(w) = probability\left(w|\bigcap_{\forall g} g\right) = probability\left(\bigcup_{\forall g}(w|g)\right)$$
$$= \left[1 - \prod_{\forall g}\left(1 - \frac{\text{語 } w \text{ と土台 } g \text{ 中の語の共起度}}{\text{土台 } g \text{ 中の語の出現頻度}}\right)\right] \quad (2)$$

3.2　主成分分析による単語のグループ化（P）

過去の一定期間 $\{t_{-1}, \cdots, t_{-T}\}$ の各月 t の主要単語 $HighFreq(t)$ と $HighKey(t)$ に含まれる語の出現パターンに対し主成分分析を行い，主要単語をグループ化し特徴量の次元圧縮を行う．上記期間で少なくとも 1 回以上主要単語に含まれたすべての単語 w_i に対して，下記のような出現行列を作成する．

$$A(w_i, t) = \begin{cases} 1 & w_i \in \{HighFreq(t), HighKey(t)\}, \\ 0 & \text{それ以外} \end{cases} \quad (3)$$

この行列に対して，主成分分析により N_{pc} 個の合成変数（主成分）にまとめる[⋆2]．各月の N_{pc} 個の主成分スコアを対象期間について時系列順に並べることによって，N_{pc} 次元の時系列データ $x_i(t), i = \{1, \cdots, N_{pc}\}$ が作成される．これが分析対象期間のテキストデータの特徴の時間的変化を表していると考える．ここで注意してほしいのは，ここまで予測対象の時系列データはまったく用いず，純粋に単語の出現パターンのみの分析を行っていることである．つまり，ここまでの分析は予測対象に依存せずに共通である．

3.3　回帰分析による市場データの動向分析（R）

最後に，各主成分スコアの毎月の動き $x_i(t)$ から月次での市場価格の動きを解析する．具体的には，さきほどの主成分スコアの時系列データ $x_i(t)$ を説明変数として，各月の月末の価格データ $p(t)$ を非説明変数とする重回帰分析を行う．

$$\tilde{p}(t) = a_0 + \sum_{i=1}^{N_{pc}} a_i x_i(t) \quad (4)$$

回帰分析の際に，AIC 基準[10] を用いたステップワイズ選択により，説明変数の絞り込みを行った．得られた回帰式に，月央に発表された最新のテキストデータを入力すれば，半月後の月末の市場価格を推定（外挿予測）できる．

4.　運用テストによる他手法との比較

提案手法の有効性を確かめるために，日本国債市場での運用テストを行い，既存手法と運用損益の比較を行った．

4.1　運用テストの手法

今回の運用テストでは売買は月次とし，毎月の金融経済月報が発表された時点で取引ルールに従って買いまたは売りのポジションを持つ取引と，月末にポジションを解消してスクウェアに戻して損益を確定する取引を行う．取引量は毎月決まった資本量に固定し，売買量の調整は行わない．また，取引手数料は考慮しなかった．

まず，直近のデータまでを訓練データとして 3.1 から 3.3 節の手続きで回帰式を推定し，当月の新しい金融経済月報のテキスト情報を入力し月末の債券価格を予想する．当月 t に関して，$\tilde{p}(t)$ をテキストマイニングで推定した月末価格，$p'(t)$ を金融経済月報が公開された時点の価格とする．$p(t)$ を実際の月末の価格とする．次に，前月からの予想価格の変動幅 $\tilde{\Delta}(t) = \tilde{p}(t) - \tilde{p}(t-1)$ と，月報発表時に実現している変動幅 $\Delta'(t) = p'(t) - p(t-1)$ を比較し取引を決定する．

$$\begin{cases} 1 \text{ 単位の資本を買う}, & \tilde{\Delta}(t) > \Delta'(t) \text{ の場合}, \\ 1 \text{ 単位の資本を売る}, & \tilde{\Delta}(t) < \Delta'(t) \text{ の場合} \end{cases} \quad (5)$$

月末に月報発表時の取引と反対の売買を行い，損益を確定する．今月の損益 $PL(t)$ は，月報発表後の変動幅 $\Delta(t) = p(t) - p'(t)$ と，予測価格の変動幅と月報発表時点の変動幅 $\tilde{\Delta}(t) - \Delta'(t)$ の符号を比較し，月報発表後の価格変動の大きさ $|\Delta(t)|$ に比例した大きさに確定する．

$$PL(t) = \begin{cases} |\Delta(t)| & \Delta(t)(\tilde{\Delta}(t) - \Delta'(t)) > 0 \text{ の場合}, \\ -|\Delta(t)| & \Delta(t)(\tilde{\Delta}(t) - \Delta'(t)) < 0 \text{ の場合} \end{cases} \quad (6)$$

これらの手順を，毎月のデータを追加して回帰式を更新しながら，テスト期間の終わりまで逐次的に行う．

⋆1 本研究は $N = 10$ とした．

⋆2 本研究では 1998 年 1 月から 2007 年 12 月までのテキストデータを用いた主成分分析で，累積寄与率が 60%を超えた主成分数が 30 であったので，$N_{pc} = 30$ とした．

4.2　比較対象の他手法

同じ期間の運用テストを行い，以下の 4 つの既存手法と運用結果を比較した．

(1)　単語頻度を用いたサポートベクタ回帰（TF-SVR）：提案手法と同じ金融経済月報について，名詞・動詞・形容詞の基本形の毎月の頻度を計算し，サポートベクタ回帰（SVR）の入力とした．SVM-Light[⋆1]を用いて線形カーネルで回帰し，毎月末の価格を予想し，4.1 節と同様にして取引を行う．

(2)　共起頻度を用いたサポートベクタ回帰（Co-SVR）：提案手法の主要単語の抽出（3.1 節）まで行い，主成分分析による単語のグループ化を行わずに，SVR の入力とした．後は，TF-SVR と同様にして取引を行う．

(3)　数値指標を用いた計量経済モデル（BOJ）：日本銀行が提唱した日本国債利回りを 6 つの経済指標[⋆2]を用いて回帰分析したモデル[11] を月次に拡張した式[12]．前月までの数値指標を用いて回帰し，他と同様のルールで取引を行う．

(4)　時系列外挿モデル（EXT）：過去の価格チャートから線形的な外挿を行う順張りモデル．前月末から月報発表時まで価格が上昇していたら買いポジションを有する．下降していたら売りポジションになる．

4.3　運用テスト結果

運用テストの期間は 2008 年 1 月から 2010 年 5 月までであり，各月の取引を決定するために 1998 年 1 月から前月までのテキスト情報・価格・数値指標を訓練データとして用いた．テストを行った市場は，日本国債の 2 年物，5 年物，10 年物である．式 (6) の損益をテスト期間全体で平均し年率に計算した結果を**表 1** に示す．全体的には，金融経済月報を用いた手法（CPR，TF-SVR，Co-SVR）は，数値指標（BOJ）や線形外挿（EXT）よりも高い利益を出せた．このことから，テキスト情報が長期市場分析に有用な情報を含んでいたことが分かる．詳細に見ると，提案手法（CPR）はどの市場でも安定して，ほぼ最高水準の運用益をあげることができた．他の既存のテキストマイニング手法（TF-SVR，Co-SVR）は少し不安定であり，市場によってはテキストマイニング以外の手法を下回る運用成績を示した．つまり，本手法により，安定した外挿予測に有効な集約された特徴量を抽出できたことが分かる．

表 1　運用テストでの平均損益（年率）
Table 1　Profit and loss in the operational test (annual rate).

	CPR	TF-SVR	Co-SVR	BOJ	EXT
日本国債 2 年	**59.60**	40.85	36.49	50.24	41.70
日本国債 5 年	**223.18**	88.56	215.37	−39.90	143.70
日本国債 10 年	243.02	**248.14**	233.70	31.79	−47.84

単位はベーシスポイント（0.01%）．太字は各市場での最大利益．

⋆1 http://svmlight.joachims.org/

⋆2 消費者物価指数 3 年前比年率，鉱工業生産 3 カ月前比年率，無担保コールオーバーナイト，米国実質金利，ドル円 3 カ月前比年率，CD・TB（3 カ月物）レート格差．

5.　CPR 法の詳細検証

なぜ提案手法が既存手法よりも良い運用成績を示すことができたのか，変動の予測精度と抽出された単語グループの中身の 2 点から検証する．

5.1　変動の予測精度

今回の運用テストでは取引量の調整は行わなかったので，重要なのは，売買の方向性を決める将来の価格変動の方向性を正確に推定することである．テスト期間の 2 年 5 カ月間のうち，各手法で推定した変動の方向（上昇/下降）が合っていた月の割合を見ると，意外にも CPR 法は特に予測精度が高いわけではなかった（**表 2**）．正答率が 50%である帰無仮説について Z 検定を行った結果，どの場合も有意水準 5%では帰無仮説を棄却することはできなかった．

ところが，月報発表後の価格変動の大きさ $|\Delta(t)|$ が上位 25%内の月（テスト期間 29 カ月中 7 カ月）に限定すると，CPR 法の正答率が飛躍的に高くなっていた（**表 3**）．表 2 と同様の検定を行った結果，日本国債 2 年物と 5 年物については CPR 法のみが，日本国債 10 年物に関しては Co-SVR 法が有意水準 5%で正答率 50%の場合と有意差が見られた．このような時期は，金融経済月報に込められた日本銀行の態度変化や市場へのメッセージに対して，国債市場が大きく反応した月だと考えられる．つまり提案手法は，テキスト情報から市場動向の予兆を比較的うまく抽出することができたのである．そして，価格変動が大きかっ

表 2　テスト期間全体での価格変動の正答率 (%)
Table 2　Accuracy of price change prediction during the test period (%).

	CPR	TF-SVR	Co-SVR	BOJ	EXT
日本国債 2 年	55.17	**65.52**	58.62	62.07	55.17
日本国債 5 年	58.62	44.83	58.62	55.17	**62.07**
日本国債 10 年	55.17	**62.07**	58.62	55.17	44.83

太字は各市場での最高精度．

表 3　高変動期間での価格変動の正答率 (%)
Table 3　Accuracy of price change prediction during the highly fluctuated period (%).

	CPR	TF-SVR	Co-SVR	BOJ	EXT
日本国債 2 年	**85.71**	71.43	71.43	71.43	57.14
日本国債 5 年	**85.71**	57.14	42.86	28.57	42.86
日本国債 10 年	71.43	71.43	**85.71**	42.86	57.14

太字は各市場での最高精度.

た月の取引では損益 $PL(t)$ も大きくなるので，変動期の予測精度が高ければテスト期間全体での運用益も大きくなる．

5.2　抽出された単語グループの内容分析

実際に，本手法で抽出された単語グループが経済分析的に意味ある分類になっているのかを調べた．1997 年 1 月から 2007 年 12 月までのテキストから抽出された主成分と各主成分で負荷量の絶対値が上位のキーワードを**表 4** に示す．この時期の名詞・動詞・形容詞の基本形は全部で 2,927 個であり，そこから 3.1 節の手法で 273 個の主要単語が抽出された．さらに，3.1 節の主成分分析で 30 個の主成分に集約された．

抽出された主成分に関連する単語を見て，市場分析時によく使われる経済要因[13] に分類した（**表 5**）．たとえば第 1 主成分は，「横ばい」「圏内」「緩やか」といった動きを表す単語と関連するので，「市場の地合い」要因に分類される．第 3 主成分は，「需要」「改善」「生産」といった単語の寄与が高いので，「生産・在庫」要因に分類される．各主成分は市場分析に使われる経済要因との対応関係が明確であり，比較的経済的な意味のある主成分に集約されていたことが分かった．

特に興味深いのは，本手法は数値データには現れにくい事件性の高い要因をテキストデータからうまく抽出できていることである．たとえば，第 10 主成分は「金融システム不安」に関わる単語の寄与が高く，第 12，14，15，16 主成分は「国際政情不安」というニュース的な情報を表している．これらの要因が経済指標にすべて反映されているかは分からず，もし反映するとしてもニュースよりもかなり時間がたってからである．また，チャート時系列に反映されるのも，価格が動いてからであり，しかもそれが事件で動いたのか経済的なファンダメンタルズで動いたのかは判別が困難である．本テキストマイニング手法はこれらの情報を抽出できたので，既存の計量経済モデルや時系列外挿モデルよりも，市場の大きな変化の兆候をうまくとらえることができたと思われる．

表 4　テキストから抽出された主成分と各主成分で負荷量の絶対値が上位のキーワード
Table 4　Keywords with larger loadings to each principal components.

主成分 1	主成分 2	主成分 3	主成分 4	主成分 5	主成分 6	主成分 7	主成分 8
横ばい	リスク	背景	設備	足許	量的	調整	歯止め
圏内	軟調	伴う	国内	上昇	停滞	雇用	掛かる
環境	国債	需要	低迷	実体	持続	関連	総合
資金	利回り	改善	輸出	年末	強い	厳しい	対策
伸び	格差	生産	歯止め	頭打ち	実施	銀行	中小
基調	根強い	鈍化	掛かる	先行き	歯止め	量的	見込む
緩やか	投資	軟調	総合	厳しい	掛かる	停滞	収益
民間	窺う	国債	対策	間	総合	持続	ベース
金融	横這い	利回り	ベース	軟化	対策	強い	指標

主成分 9	主成分 10	主成分 11	主成分 12	主成分 13	主成分 14	主成分 15	主成分 16
マクロ	システム	年末	同時	作用	雇用	もと	不透明
ギャップ	銀行	頭打ち	テロ	進行	縮小	効果	生産
超過	不安	受ける	事件	昨秋	受ける	同時	金利
市況	済	間	社債	公共	イラク	テロ	調達
国際	傾向	軟調	機械	ベース	情勢	事件	イラク
プラス	個人	国債	米国	不安	必要	結果	情勢
商品	幅	利回り	システム	済	不透明	支出	低調
均す	大幅	格差	財	結果	賃金	アジア	銀行
考える	伴う	根強い	発行	季節	アジア	財	長期

主成分 17	主成分 18	主成分 19	主成分 20	主成分 21	主成分 22	主成分 23	主成分 24
減少	アジア	着実	乏しい	賃金	製品	調査	一部
反動	米	高め	流通	消費	年末	本年	受ける
金利	前年	反動	需給	一部	頭打ち	不透明	圧力
わが国	効果	昨年	減少	発行	状況	乏しい	既往
弱まる	伴う	マクロ	自動車	不透明	必要	流通	弱い
相場	年末	ギャップ	明確	需要	減少	減少	緩和
部品	頭打ち	超過	維持	既往	その後	イラク	需給
たどる	その後	雇用	弱い	サービス	マクロ	情勢	不透明
強まる	不安	調査	好影響	持ち直し	ギャップ	高水準	最終

主成分 25	主成分 26	主成分 27	主成分 28	主成分 29	主成分 30
後退	米価	押し上げ	米価	ドル	住宅
調査	一時	働く	一時	相場	米価
本年	調査	個人	発行	方向	一時
意識	本年	需要	強まる	イラク	既往
発行	圧力	着実	意識	情勢	一部
米	高水準	輸出	後退	基調	為替
サービス	最終	収益	当面	米	伸び
緩和	作用	要因	電気	発行	後退
既往	進行	相場	アジア	テンポ	変化

表 5　主成分と経済要因との関係性
Table 5　Relationship between principal components and economic factors.

経済要因	主成分
1. 景気	
a. 景況	25
b. 設備投資	4
c. 貿易収支	9, 17, 27, 28
d. 企業活動	8, 20, 22, 23
e. 生産・在庫	3, 16, 20, 22
f. 雇用	7, 14, 19, 21
g. 住宅	30
h. 個人消費	10, 27
2. 物価	26, 28, 30
3. 金融政策	1, 2, 6, 10, 17
4. 政情	12, 14, 15, 16, 18, 23, 24, 29
5. 市場の地合い	1, 2, 5, 11, 13, 19

6. ま と め

金融経済月報の担当者は，表現の一貫性を保つことに最新の注意を払っている．ある経済状況を表現するのに，過去の似た状況時の月報を参考にして言葉を選ぶことがよくあるそうだ．本研究の手法は，そのような形式の安定した専門的なテキスト情報から，特定の状況を表現するのによく使われる単語をグループ化して抽出することに成功した．それにより，金融テキストマイニングの先行研究でよく使われる，単語頻度をそのままサポートベクタ回帰に入力する手法より高精度に，長期市場予測を行うことができた．

特に，今回分析した国債市場は，中央銀行である日本銀行の動向に敏感であるといわれている．なぜなら，長期国債の市場動向を基準に長期金利が決定されるが，それらは中央銀行が設定する短期の政策金利と密接に関連するからである．つまり，本研究で用いた CPR 法は，金融経済月報に記述されている日本銀行の態度変化に関する兆候や場合によっては意図的なメッセージをうまく抽出できたとも考えられる．先行研究[1)]で本手法を国債市場以外の日本の株式市場や外国為替市場に適用した結果，株式市場では有効であったが，外為市場ではあまり有効でなかった．これは，日本銀行の動向に対する市場の感応度から考えると当然の結果と思われる．ただ，海外の市場分析に海外の中央銀行が発行している経済リポートを用いて本手法を適応可能であるので[14)]，今後多国間比較を行えば外国為替市場も精度の良い分析が可能となるであろう．

表 4 を見て分かるように，現状では主に名詞と動詞から構成される単語グループしか抽出できていない．「激しい上昇」や「弱いドル」のような修飾語による表現の調整まで取り扱えていない．経済分析ではこのような表現の調整は意味が大きい場合があるので，今後の課題としたい．また，本手法は市場分析だけに限らず他の分野のテキスト情報にも応用可能である．経済以外の分野でも本手法の有効性を試していきたい．

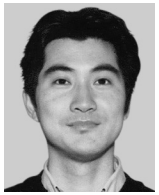
謝辞　本研究の一部は，科学研究費補助金特定領域研究「情報爆発 IT 基盤」の助成を受けています．お礼申し上げます．

参 考 文 献

1) 和泉　潔，後藤　卓，松井藤五郎：テキスト分析による金融取引の実評価，人工知能学会論文誌，Vol.26, No.2, pp.313–317 (2011).
2) Bollen, J., Mao, H. and Zeng, X.-J.: Twitter mood predicts the stock market, *CoRR*, Vol.abs/1010.3003 (2010).
3) Mittermayer, M.A. and Knolmayer, G.: Text Mining Systems for Market Response to News: A Survey, Working paper (2006).
4) Schumaker, R.P. and Chen, H.: A Discrete Stock Price Prediction Engine Based on Financial News, *IEEE Computer*, Vol.43, No.1, pp.51–56 (2010).
5) Seo, Y.-W., Giampapa, J.A. and Sycara, K.: Financial News Analysis for Intelligent Portfolio Management, Technical Report CMU-RI-TR-04-04, Carnegie Mellon University (2004).
6) 丸山　健，梅原英一，諏訪博彦，太田敏澄：インターネット株式掲示板の投稿内容と株式市場の関係，証券アナリストジャーナル，Vol.46, No.11・12, pp.110–127 (2008).
7) Antweiler, W. and Frank, M.Z.: Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance*, Vol.59, No.3, pp.1259–1294 (2004).
8) 和泉　潔，後藤　卓，松井藤五郎：テキスト情報による金融市場変動の要因分析，人工知能学会論文誌，Vol.25, No.3, pp.383–387 (2010).
9) 大澤幸生：チャンス発見のデータ分析―モデル化+可視化+コミュニケーション→シナリオ創発，東京電機大学出版局 (2006).
10) Akaike, H.: A new look at the statistical model identification, *IEEE Trans. Automatic Control*, Vol.19, pp.716–723 (1974).
11) 日本銀行調査統計局：情勢判断資料（1997 年秋）(1997), 入手先⟨http://www.boj.or.jp/research/past_release/js/js1997d.pdf⟩.
12) 三菱東京 UFJ 銀行：日銀長期金利モデル検証 (2002), 入手先⟨http://www.bk.mufg.jp/report/focus2002/FotM_32.pdf⟩.
13) 住友信託銀行マーケット資金事業部門：投資家のための金融マーケット予測ハンドブッ

ク（第 4 版），日本放送出版協会 (2009).
14) 余野京登，和泉　潔，後藤　卓，松井藤五郎，陳　ユ：英文経済レポートのテキストマイニングと市場分析，2010 年度人工知能学会全国大会（第 24 回）(2010).

(平成 23 年 3 月 29 日受付)
(平成 23 年 9 月 12 日採録)



和泉　　潔（正会員）

東京大学大学院工学系研究科准教授．1993 年東京大学教養学部基礎科学科第二卒業．1998 年同大学院博士課程修了．博士（学術）．同年より 2010 年まで電子技術総合研究所（現，産業技術総合研究所）勤務．2010 年より現職．マルチエージェントシミュレーション，特に社会シミュレーションに興味がある．人工知能学会，電子情報通信学会，電気学会各会員．

後藤　　卓

三菱東京 UFJ 銀行融資企画部．1997 年名古屋大学工学部情報工学科卒業．同年株式会社東海銀行（現，株式会社三菱東京 UFJ 銀行）入社，2010 年より現職．1998 年より ALM および債券運用業務に従事し，2001 年から 2007 年まで日本および英国にてプロップ・トレーディング業務に従事．帰国後，円貨資金証券部，市場企画部を経て現在に至る．

松井藤五郎

1997 年名古屋工業大学知能情報システム学科卒業．2003 年同大学院工学研究科博士後期課程電気情報工学専攻修了．博士（工学）．2003〜2009 年東京理科大学理工学部経営工学科助教．2009 年とうごろう機械学習研究所設立．2010 年より中部大学生命健康科学部臨床工学科兼工学部情報工学科講師．機械学習およびデータ・マイニングに関する研究に従事．人工知能学会，ACM，AAAI 各会員．