2010 年大会速報論文特集

Short Paper

テキスト分析による金融取引の実評価

Implementation Tests of Financial Market Analysis by Text Mining

和泉 潔
Kiyoshi Izumi
東京大学大学院 工学系研究科 & JST さきがけ
School of Engineering, The Univ. of Tokyo & PRESTO, JST
kiyoshi@ni.mints.ne.jp, http://kinba.sakura.ne.jp/

後藤 卓
Takashi Goto
三菱東京 UFJ 銀行 市場企画部*1
The Bank of Tokyo-Mitsubishi UFJ, Ltd.
takashi_6_gotou@mufg.jp

松井 藤五郎
Tohgoroh Matsui
中部大学 生命健康科学部 & 工学部
College of Life and Health Sciences & College of Engineering, Chubu University
TohgorohMatsui@tohgoroh.jp, http://とうごろう.jp

keywords: financial market, text-mining, time-series data, regression analysis.

Summary

In this study, we propose a new text-mining method for long-term market analysis. Using our method, we performe out-of-sample tests using monthly price data of financial markets; Japanese government bond market, Japanese stock market, and the yen-dollar market. First we extract feature vectors from monthly reports of Bank of Japan. Then, trends of each market are estimated by regression analysis using the feature vectors. As a result of comparison with support vector regression, the proposal method could forecast in higher accuracy about both the level and direction of long-term market trends. Moreover, our method showed high returns with annual rate averages as a result of the implementation test.

1. は じ め に

金融市場のトレーダー達は，市場に影響を及ぼす多様な情報を取捨選択し，現在の市場の状況を分析・予測している．しかし，送られてきた情報の全てを，現場のトレーダーが自分で目を通して市場分析に用いることは不可能に近い．そのため，いくつかの情報技術を市場分析に適用する研究が行われ，一定の成果をあげてきた．特に近年，テキスト情報による市場分析に関して，ロイターなどのオンラインの経済ニュースに対する市場の反応を推測する研究もでてきた [Mittermayer 06a, Seo 04, Ahmad 05, 高橋 07]．しかし，これらの研究は，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ており，より長期的な市場動向の分析には用いられてこなかった．本研究では，オンラインのテキスト情報から，数年にわたる比較的長期の市場動向分析を支援す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．

†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．

2. テキストデータによる長期市場分析手法

今までテキストマイニングが長期的な市場分析にほとんど適用されなかった主な原因は次の 2 点である．

a. テキスト情報の様式: 従来の経済テキストマイニング研究では，ニュース記事 [Fung 02, Mittermayer 06b] や掲示板 [丸山 08, Antweiler 04] などの形式が多様なテキスト情報を使用したため，異なる時点間のテキストを比較して長期的な時間変化を抽出することが困難であった．この問題に関して，本研究では長期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．この情報によって，日本銀行が，当面の経済動向をどう分析しているか対外的に明らかにしており，実際の金融市場のトレーダーが多かれ少なかれ着目している共有の重要テキスト情報である．金融経済月報が長期分析に有効な理由は，解説内容の順番や段落構成等がほぼ定式化されていて，月ごとのテキスト内容の変化が比較しやすいからである．このテキスト情報を用いた長期市場分析を一般的な数値データを用いた分析と外

*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．

挿予測誤差を比較した結果，適切な分析手法を用いれば推定精度を大幅に改善でき，本テキストが長期市場分析に適していることが明らかになっている [和泉 08].

b. テキスト分析と外部時系列データの関連性抽出: 二番目の問題は，市場分析用にテキストデータと時系列データを関連づける適切な手法がなかったことである．

二番目の問題に関して，本研究ではテキストデータと時系列データを関連づけるために，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．

2·1 共起関係に基づく主要単語の抽出と可視化 (C)

最初に，各月のテキストデータに KeyGraph [大澤 06] を適用し，共起関係を解析した．分析対象の範囲は，毎月の金融経済月報の中から，図表内のテキストを除いた本文と脚注のテキストとした．具体的にはまず，日本語形態素解析システムである Chasen [ChaS] による形態素解析を行い，出現頻度順に名詞・動詞・形容詞等を抽出した．次に，Jaccard 係数（= p(A and B)/p(A or B); ただし A,B は抽出した単語）を段落毎に適用し，段落毎に同時に出現する単語と単語を繋ぎ，共起グラフを作成する．この際に，段落の区切りは金融経済月報テキストでの改行とし，セクションタイトルは一つの段落として扱った．その後，単結合（A,B 間のみの結合部分）を切断し，結合による「島」を作成する．またその後，各単語間の共起度に基づき，上位順に「橋」を作成する．これらによって，各月のテキストデータから主要単語をノードとするネットワークを構築した．

2·2 主成分分析による単語のグループ化 (P)

KeyGraph で作成したネットワークに出現した単語のパターン (単語を月毎の出現状況に従いパターン分類したもの) に対し主成分分析を実施し，30 個の合成変数（主成分）にまとめる．ここで，主成分の数が 30 個であったのは，1998 年から 2007 年までの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 であったからである．各月の 30 個の主成分スコアを，分析対象期間について時系列順に並べることによって，30 次元の時系列データが作成される．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．主成分分析の際には，単語に関して品詞を区別せずに分析を実施する．ここで注意してほしいのは，ここまで市場データは全く用いず，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，債券市場や株式市場，外国為替市場などの分析対象となる市場の種類に依存せずに共通である．

2·3 重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 30 個の主成分スコアの時系列データを説明変数として，月次の市場データを被説明変数とする重回帰分析を行う．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，約 2 週間後の月末の市場価格を推定 (外挿予測) できる．

3. 金融経済月報のテキストマイニング

上述の手法を用いて，日本国債市場 (金利)・株式市場 (日経平均株価)・外国為替市場 (円ドルレート) の月次変動を分析した．1998 年 1 月から 2007 年 12 月までの 10 年間 (120ヶ月) の金融経済月報のテキストと各市場データ (月末終値) を訓練データとした．毎月のテキストのファイルサイズは 15 から 25KB で、文字数は約 8 千から 1 万 2 千個であった．

3·1 訓練データの説明力評価

最初に，KeyGraph アルゴリズムと主成分分析を用いて，30 次元の特徴量を金融経済月報のテキストデータから抽出した．抽出された主成分には大きく分けて 2 つのタイプがあった．一つは市場の動きに関する特徴量である．例えば，1 番目の主成分は,「横ばい」「圏内」「緩やか」といった動きを表す単語から構成されていた．他にも，5 番目の主成分は,「上昇」「頭打ち」「軟化」といった単語の寄与が高かった．もう一つのタイプは，経済のファンダメンタルズに関する特徴量である．例えば，2 番目の主成分は「リスク」「国債」「利回り」といった金利に関する単語から構成されていた．他にも，3 番目の主成分は,「需要」「改善」「生産」といった企業活動に関する単語の寄与が高かった．

次に，これらの 30 次元の特徴量の時系列データを用いて，各市場データの回帰分析を行った．回帰分析の際に，AIC 基準 [Akaike 74] 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．日本国債の 2 年物，5 年物，10 年物の金利について，23-25 個の説明変数による回帰式を得ることができた．日経平均については 18 個，円ドルレートに関しては 13 個の説明変数が選択された．決定係数 R^2 をみると，訓練データについて十分な説明力を持つことがわかった．R^2=85.67%(日経平均)，76.38%(円ドルレート)，78.47%(国債 2 年物)，76.76%(国債 5 年物)，74.65%(国債 10 年物).

3·2 サポートベクタ回帰 (SVR) との予測力比較

1998 年から 2007 年の訓練データから推定された回帰式に，2008 年 1 月から 12 月の新たなテキスト情報を入力し外挿予測を行った．提案手法の予測力を評価するために，従来の金融テキストマイニング研究でよく用いら

れているサポートベクタ回帰 (SVR) と予測力を比較した[*2]．提案手法と同じ 1998 年から 2007 年の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の入力とした．訓練期間の入力データは 2927 次元となった．非説明変数も提案手法と同様に，テキストが発表された月の終値 (日経 225, 円ドル，国債 2 年物，国債 5 年物，国債 10 年物) である．SVM-Light[*3]を用いて線形カーネルで回帰し，2008 年の 1 年間を外挿予測した．

提案手法と SVR でどちらの予測値が，実際の市場価格に近かったのかを評価するために，外挿期間の平均 2 乗平方根誤差 (RMSE) を比較した．その結果，国債 2 年物と 5 年物に関して，提案手法の方が有意に予測誤差が小さかった (図 1)． 次に，変動の方向性の予測力を評価するために，外挿期間の 12ヶ月間において，前月からの価格変動の予測値が実際の価格変動と符号 (上下) が合っていた月数の割合 (正答率) を比較した．その結果，円ドル以外の全ての市場において，提案手法の方が SVR より高い正答率を示した (図 2)． これらの結果より，我々の提案する手法が，長期の市場動向の水準や方向性の両方に関して，SVR よりも高い精度で予測できることが明らかになった．SVR は提案手法である CPR 法の共起解析 (C) と主成分分析 (P) を行わずに，回帰分析 (R) の部分により複雑な手法を用いた分析であると考えられる．なので，今回の結果は共起解析と主成分分析の組み合わせが，金融経済月報の分析の有効性向上に貢献していることも示している．

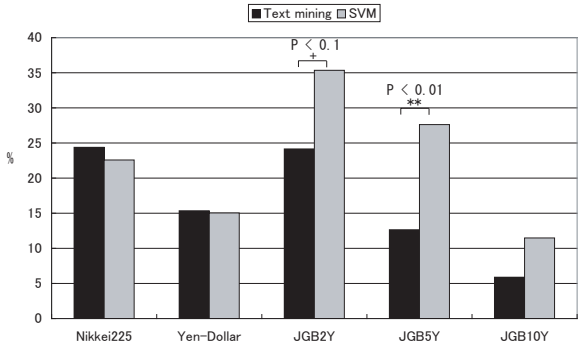


図 1 外挿期間の平均 2 乗平方根誤差 (RMSE) の比較．縦軸は，各市場の外挿期間の平均価格に対する RMSE の比率 (%) で表示．

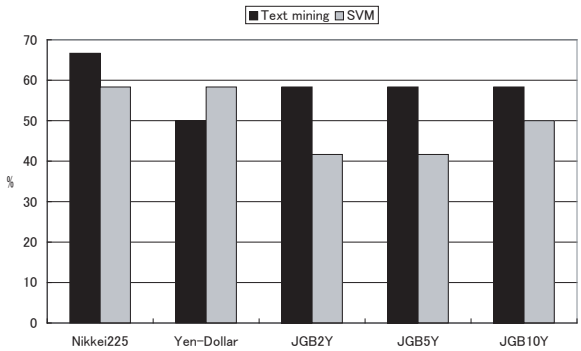


図 2 外挿期間の変動予測の正答率

4. 外挿予測力の運用テスト

上述のテキストマイニング手法の有効性を確かめるために，実際に現場で使われる場面と同様に，直近のデータまでを訓練データとして毎月新しいデータを追加して新たに分析を更新した場合の外挿予測力のテストを行った．

4·1 運用テストの手法

運用テストの期間は 2008 年 1 月から 2009 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データと価格データを訓練データとして用いた．

§1 取引ルール

取引決定に次の 2 通りのルールを用いて検証した．ただし，$\tilde{p}_t$ をテキストマイニングで予測した月末価格，p'_t を金融経済月報が公開された時点の価格，p_t を実際の月末の価格とする．また，t は 1ヶ月ごとの月次の時点を表す．

(1) 取引ルール 1(価格水準の比較): 予測価格 $\tilde{p}_t$ と発表時の価格 p'_t を比較して，予測価格が高ければ ($\tilde{p}_t > p'_t$)，価格 p'_t で 1 単位の資本を買う．低ければ ($\tilde{p}_t < p'_t$)，1 単位の資本を売る．

(2) 取引ルール 2(価格変動の比較): 予測価格 $\tilde{p}_t$ の前月からの変動幅 ($\tilde{\Delta}_t = \tilde{p}_t - \tilde{p}_{t-1}$) と，発表時の価格 p'_t の前月末価格 p_{t-1} からの変動幅 ($\Delta'_t = p'_t - p_{t-1}$) を比較して，予測価格の変動幅が高ければ ($\tilde{\Delta}_t > \Delta'_t$)，価格 p'_t で 1 単位の資本を買う．低ければ ($\tilde{\Delta}_t < \Delta'_t$)，1 単位の資本を売る．

今回のテストでは，各市場での売買は月次であり，金融経済月報が発表された時点で取引ルール 1 またはルール 2 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．

§2 評価ルール

損益は，取引ルール 1 に関しては，実際の月末の価格 p_t が月報発表時点の価格 p'_t に比べて，予測価格 $\tilde{p}_t$ と同じ方向に変動した場合 ($(p_t - p'_t)(\tilde{p}_t - p'_t) > 0$) に，$|p_t - p'_t|$ の利益が得られる．異なる方向に変動した場合

*2 先行研究では，テキスト情報から市場の上下を分類するため，サポートベクタマシンが使われた [Fung 02, Mittermayer 06b]. 今回は分類ではなく，価格の値を回帰するのでサポートベクタ回帰を比較対象とした．

*3 http://svmlight.joachims.org/

$((p_t - p'_t)(\tilde{p}_t - p'_t) < 0)$，$|p_t - p'_t|$ の取引損が月末に確定する．

取引ルール2に関しては，実際の月末価格 p_t と月報発表時点の価格 p'_t の変動幅 $\Delta_t = p_t - p'_t$ が，予測価格の変動幅 $\tilde{\Delta}_t$ と月報発表時点の変動幅 Δ'_t の差と同じ方向の場合 $(\Delta_t(\tilde{\Delta}_t - \Delta'_t) > 0)$，$|p_t - p'_t|$ の利益が得られる．異なる方向に変動した場合 $(\Delta_t(\tilde{\Delta}_t - \Delta'_t) < 0)$，$|p_t - p'_t|$ の取引損が月末に確定する．

§3 データ更新

これらの手順を，毎月のデータを更新しながら逐次的に行う．

まず最初に1998年1月から2007年12月までのテキストデータ(日銀金融経済月報)と市場データを訓練データとして回帰式を推定し，その式に2008年1月のテキストデータを入力して，2008年1月末の市場価格を外挿予測によって推定した．2008年1月の金融経済月報は2008年1月22日に公開されたので，その時点から2008年1月31日までの外挿予測となる．2008年1月22日時点の価格でポジションを持ち，2008年1月31日時点の価格で損益を確定する．

次に2008年1月のテキストデータと市場データを訓練データに追加して，1998年1月から2008年1月までのテキストデータと市場データを訓練データとして回帰式を推定し，その式に2008年2月のテキストデータを入力して，2008年2月末の市場価格を外挿予測し取引を行う．同様にして毎月のデータを追加して回帰式を逐次的に更新しながら，月末の市場価格を外挿予測するテストを，2009年5月の市場価格の外挿予測まで繰り返した．

4·2 運用テスト結果

取引ルール1とルール2でテストを行った結果を図3に示す．各市場での運用結果の評価に，シャープレシオと呼ばれる指標を用いた．シャープレシオは下記の式で表され，高い運用利回りをより低いリスク(収益のブレ)で得られたかを表す標準的な指標である．

$$\text{シャープレシオ } S = \frac{R - R_f}{\sigma}, \qquad (1)$$

ただし，R は外挿期間中の平均運用利回り，R_f は無リスク資産の利回りで今回は無担保コールレートの値を用いた．σ は運用利回りの外挿期間中の標準偏差を表す．

取引ルール間の比較 まず，取引ルール1と取引ルール2で比較すると，全般的に変動幅を比較した取引ルール2の方が運用成績が良かった．このことから，日銀金融経済月報と提案テキストマイニング手法では，市場価格の水準自体よりも市場変動の方向性に関する情報の方がよく抽出できていることを表していると思われる．

市場間の比較 次に各市場ごとに平均年率損益リターンを比べてみると，日経平均株価・日本国債5年物 > 日本国債2年物・日本国債10年物 > 円ドルレートの順で，運用成績が良かった．これは，日銀の動向が各市場に対してどれほど影響力を持ちうるのかということを表した結果だと思われる．特に，外国為替レートに関して，日本経済中心のレポートの分析では限界があることは当然といえば当然である．

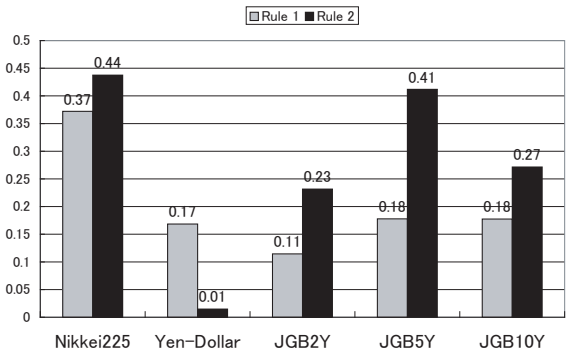


図3 提案手法の運用テスト結果．テスト期間: 2008年1月から2009年5月，訓練期間: 1998年1月から前月．縦軸はテスト期間のシャープレシオ．

方向性の予測 表1に示すテスト期間での価格変動の正答率を見ると，平均リターンの高かった日本国債5年物の正答率が，他の市場に比べて高かったわけではない．それでも，平均リターンが高かったのは，大きな価格変動があった月で，きちんと価格変動の方向性を正答できたからである．図4に示すように，方向性を間違えた月は，価格変動の絶対値が小さく，それほど大きく市場が動かなかった月が多かった．逆に，図4の左端の4つと右端の4つのように，前月に比べて大きく下降また上昇した月は，提案手法による変動予測は合っていた．つまり提案手法は，市場が大きく動くときに，テキスト情報から市場動向の予兆を抽出することができたのである．

表1 テスト期間での価格変動の正答率 (%)

	日経225	円ドル	国債2年	国債5年	国債10年
ルール1	70.6	64.7	47.1	47.1	52.9
ルール2	70.6	52.9	52.9	52.9	52.9

5. ま と め

本研究では，テキストデータを用いた長期的な市場分析の新たな手法を提案した．従来の金融テキストマイニング研究でよく用いられているサポートベクタ回帰と比較した結果，提案手法が長期の市場動向の水準や方向性の両方に関して，より高い精度で予測できることが明らかになった．

金融市場の長期的な動向には，様々な要因が関係している．その中には，国内外の政情や社会全体が持つ経済

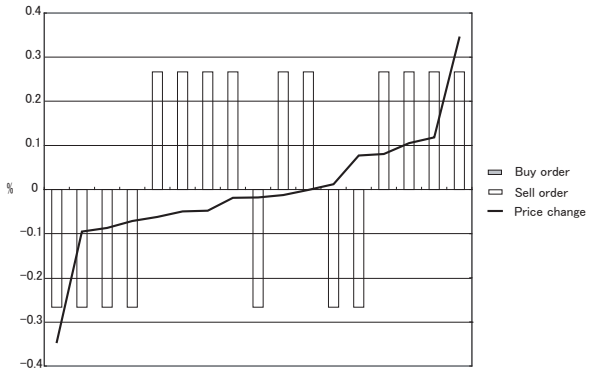


図 4 国債 5 年物に関する実際の月次価格変動 (線グラフ) と提案手法による価格変動の予測の方向性 (棒グラフ). 横軸は前月比で大きく減少した月から大きく上昇した月の順番に並べ替えた.

の先行きへの感想といった，テキスト情報で表されやすい要因が含まれる．逆に，経済指標やチャート分析などの数値的な情報からの方がうまく抽出できる要因もある．本研究の方向性として，長期市場分析におけるテキスト分析手法の限界と有効な条件を明らかにして，他の分析手法と統合した新たな手法を開発することが考えられる．そのために，今回用いたテキストデータの性質と提案手法の有効性の関係について，より詳細な分析が必要となる．例えば，テキストデータの持つ情報の一部や訓練期間の長さを変えた場合に，本手法の有効性がどのように変化するかを測る感度分析が考えられる．

本研究では，分析に好条件であると思われるテキスト情報を用いたが，今後は本手法をマーケットリポートやブログ等のより条件の厳しい情報に適用を試みる予定である．またテキストマイニングに市場分析と，市場シミュレーションを統合することによって，市場参加者の行動によるフィードバックを考慮したより動的な市場分析を行うことを目指す．

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 ◇

[Ahmad 05] Ahmad, K., Gillam, L., and Cheng, D.: Textual and Quantitative Analysis: Towards a new, e-mediated Social Science, in *Proc. of the 1st International Conference on e-Social Science* (2005)

[Akaike 74] Akaike, H.: A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, Vol. 19, pp. 716–723 (1974)

[Antweiler 04] 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)

[ChaS] ChaSen ホームページ：`http://chasen.naist.jp/hiki/ChaSen/`

[Fung 02] Fung, G., Yu, J., and Lam, W.: News Sensitive Stock Trend Prediction, in *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 481–493 (2002)

[和泉 08] 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, 2008 年度人工知能学会全国大会 (2008)

[丸山 08] 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, 11・12, pp. 110–127 (2008)

[Mittermayer 06a] Mittermayer, M. A. and Knolmayer, G.: Text Mining Systems for Market Response to News: A Survey, Working paper (2006)

[Mittermayer 06b] Mittermayer, M.-A. and Knolmayer, G. F.: NewsCATS: A News Categorization and Trading System, in *Proceedings of the 6th IEEE International Conference on Data Mining*, pp. 1002–1007 (2006)

[大澤 06] 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)

[Seo 04] 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)

[高橋 07] 高橋 悟, 高橋 大志, 津田 和彦：株式市場におけるヘッドラインニュースの効果についての研究, ファイナンス学会第 15 回大会, pp. 373–383 (2007)

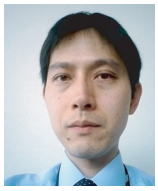
〔担当委員：中岩 浩巳〕

2010 年 7 月 13 日　受理

著　者　紹　介

和泉　潔(正会員)

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．情報処理学会，電子情報通信学会，電気学会各会員．



後藤　卓

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，2008 年より現職．1998 年より ALM 及び債券運用業務に従事し，2001 年から 2007 年までプロップ・トレーディング業務に従事．うち 2002 年から 2007 年までロンドン勤務．帰国後，円貨資金証券部を経て現在に至る．

松井　藤五郎(正会員)

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 年より現職．機械学習およびデータ・マイニングに関する研究に従事．情報処理学会，ACM，AAAI 各会員．