テキスト情報による金融市場変動の要因分析

Analysis of Financial Markets' Fluctuation by Textual Information

和泉 潔[*1] 後藤 卓[*2] 松井 藤五郎[*3]
Kiyoshi Izumi Takashi Goto Tohgoroh Matsui

[*1]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

[*2]三菱東京 UFJ 銀行
The Bank of Tokyo-Mitsubishi UFJ, Ltd.

[*3]とうごろう機械学習研究所
Tohgoroh Machine Learning Research Institute

In this study, we proposed a new text-mining methods for long-term market analysis. Using our method, we analyzed monthly price data of financial markets; Japanese government bond market, Japanese stock market, and the yen-dollar market. 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, determination coefficients were over 75%, and market trends were explained well by the information that was extracted from textual data. We compared the predictive power of our method among the markets. As a result, the method could estimate JGB market best and the stock market is the second.

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てきた．例えば，数値情報を用いて現在の市場情報を推論するようなエキスパートシステムの構築を行う研究 [日本 93] やニューラルネットや遺伝的アルゴリズムを数値情報による市場分析に用いた研究もある [電気 02]．これらの研究は一定の成果をあげてきた．しかし，数値情報には指標化されていない情報が含まれていないので，分析対象の範囲がテキスト情報よりも狭くなる可能性がある．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 06, Seo 04, Ahmad 05, 高橋 07]．これらの研究は，1 日以内や数日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．そこで，我々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の変化を分析するための補助を目的とした解析技術を新たに開発した [三菱 08]．こういった観点から，市場参加者が特に注目する日本銀行の金融経済月報を題材に，テキストマイニング技術を用いて経済市場分析を試み，また金融経済月報が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．

2. テキストデータによる長期市場分析手法

テキストマイニングを長期的な市場分析に用いるには，2 つの重要な点がある．適切な内容と形式をもつテキストデータの選択と，テキストデータと時系列データを関連づける手法である．

最初に，本研究では日本銀行の金融経済月報をテキストデータとして選んだ．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]．金融政策の方針を決める金融政策決定会合で内容を審議し，政策決定の基礎資料とする．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いる．今回，金融経済月報を分析対象にした理由は 3 つある．第一に金融経済月報は，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からである．第二の理由は，会員制の有料マーケットリポート等のテキスト情報と違って，毎月の中旬にサイト上で定期的に発表されていて，誰でもアクセス可能な情報であることである．三番目の理由は，ブログ等のほとんど決まった形式のないテキスト情報と異なり，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．

二番目のポイントとして，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，下記の 3 つのステップからなる新たなテキスト解析技術を提案する．

2.1 共起関係に基づく主要単語の抽出と可視化

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．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．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

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

連絡先: 和泉 潔，産業技術総合研究所 デジタルヒューマン研究センター，〒 135–0064 東京都江東区青海 2–41–6, kiyoshi@ni.mints.ne.jp

*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．

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，共通であるということである．

2.3 重回帰分析による市場データの動向分析

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．分析対象期間内の金利の動きを推定するだけでなく，分析対象外のテキストデータを与えれば外挿予測を行うこともできる．この外挿予測は，月中に発表される金融経済月報から，約 2 週間後の月末の市場価格を推定することになる．

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) をサンプルデータとした．

3.1 金融経済月報による月次市場分析

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 1 年物，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，サンプル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=75.24% (日本国債 1 年物)，78.47% (日本国債 2 年物)，76.76% (日本国債 5 年物)， 74.65% (日本国債 10 年物)，85.67%(日経平均)，76.38%(円ドルレート)．

3.2 外挿予測力の市場間比較

前節で得られた 1998 年 1 月から 2007 年 12 月までの過去 10 年間の訓練データを用いた回帰式に，2008 年 1 月から 12 月までのテキストデータを入力して，各金融市場における外挿予測テストを行った．図 1a-d に，代表的な市場について，推定されたパスと実際のパスを示す．外挿期間における推定パスと実際のパスを比較すると，日本国債 2 年物と 5 年物がトレンドの方向性 (上昇と下降) および価格の全体的な水準が一致しており，最も精度の高い外挿予測を行っていた．次に日本国債 1 年物と日経平均株価が，価格の全体的な水準に乖離がみられたものの，トレンドの方向性が一致していた．日本国債 10 年物は，価格の全体的な水準が合っていたものの，価格の変化量が実際の動きよりも小さく推定された．最後に，円ドルレートに関しては，市場トレンドの方向性と価格の全体的な水準ともにうまく推定することができなかった．

上述の比較結果を確かめるために，実際に提案テキストマイニング手法が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比較を行った．まず最初に 1998 年 1 月から 2007 年 9 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 2007 年 10 月のテキストデータを入力して，2007 年 10 月末の市場価格を外挿予測によって推定した．次に 2007 年 10 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 2007 年 10 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 2007 年 11 月のテキストデータを入力して，2007 年 11 月末の市場価格を外挿予測によって推定した．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，2008 年 10 月の市場価格の外挿予測まで繰り返した．図 2 は，各市場において，上記の手続きで逐次的な外挿予測を行った各月の推定値のトレンドが，実際の市場トレンドと比較して何%の期間で正解していたかを示すグラフである．推定値と実際の値がそれぞれ前月比で上昇トレンド/下降トレンドにあるかの一致を調べた．この結果より，図 1 で調べた外挿予測の精度比較結果と同様に，日本国債 2 年物・日本国債 5 年物 > 日本国債 1 年物・日経平均株価 > 日本国債 10 年物 > 円ドルレートの順で，外挿予測力の精度が高かったことがわかる．

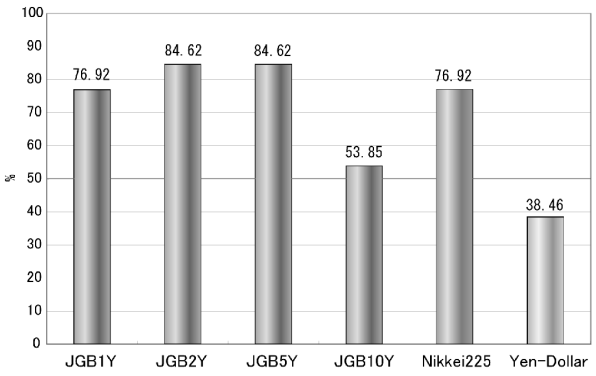

図 2: 各市場の逐次的な外挿予測における市場トレンドの正答率比較．2007 年 10 月〜2008 年 10 月までの各月末価格に関して外挿予測を行った．

3.3 市場変動の要因分析

本手法では、豊かな背景情報を含むテキストデータを用いることによって、テキスト情報で解説されている特定の経済状況と、金融市場の変動の関連性を見つけ出すことができる。日経平均株価について，さきほどの 2008 年の外挿期間につい

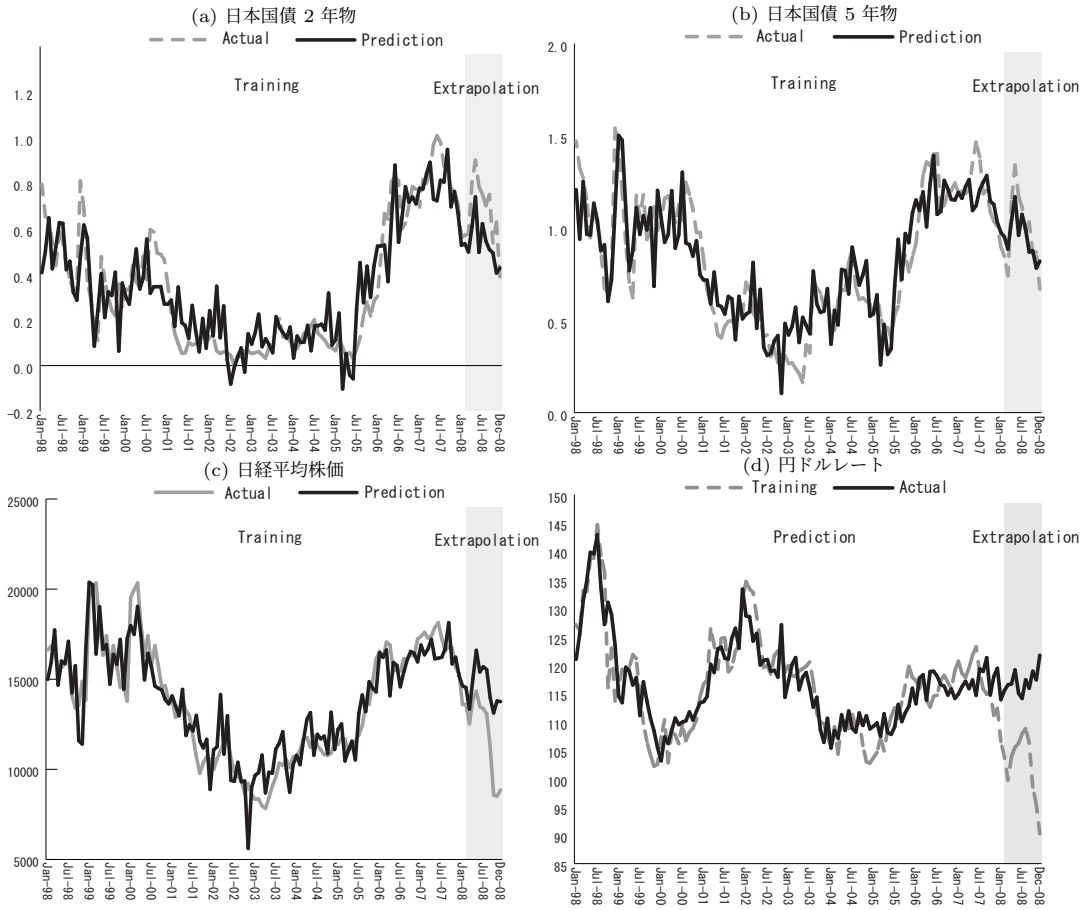

図 1: 各市場トレンドの推定．訓練期間: 1998 年 1 月〜2007 年 12 月，外挿期間: 2008 年 1 月〜12 月．

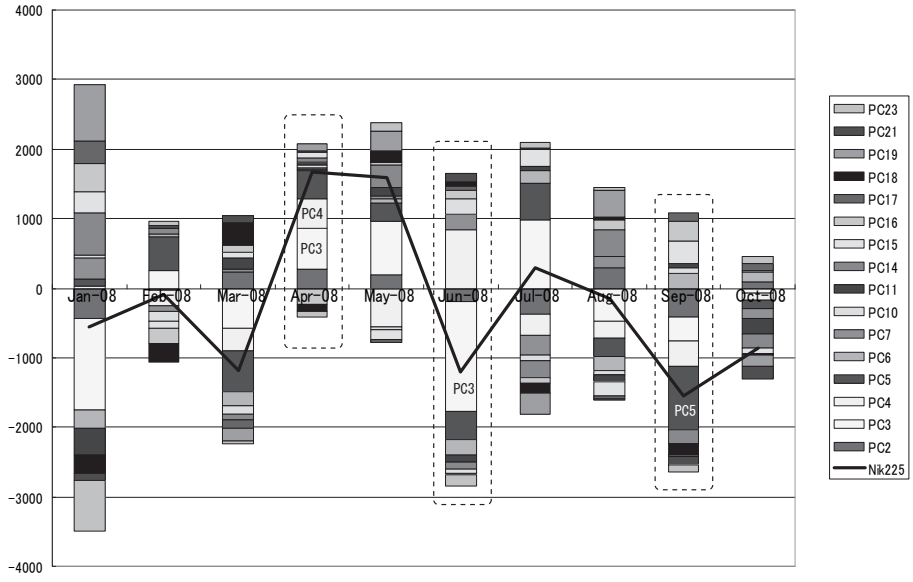

図 3: 日経平均株価の外挿期間での変動要因．線グラフはテキスト分析での推定値．棒グラフの各項目は，各主成分の前月からの変化量に回帰係数をかけた値．

て，本手法により変動が大きくなると推定された月の，変化量が大きかった主成分について調べた (図 3).

その結果，2008 年 4 月の上昇局面では，第 3 主成分と第 4 主成分に関連する単語が，その月の日銀月報での出現パターンが前月から変化していたことが分かった．表 1 に示すように，第 3 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く，第 4 主成分は「設備」「国内」「輸出」といった貿易収支に関する単語の寄与が高かった．このように 2008 年 4 月の日経平均株価の上昇は，経済的な基礎要因に関係していることが示された．

同様に，2008 年 6 月と 9 月の下降局面についても主成分の変化を調べた．その結果，同じ下降局面でも要因が異なることが示唆された．6 月の下落に最も関連した主成分は，企業活動に関する第 3 主成分であった．それに対し，9 月は「上昇」「頭打ち」「軟化」といった市況やトレンドに関する単語の寄与が高い，第 5 主成分が最も関係性が強かった．このことから，6 月の下降は経済の基礎的要因の変化を反映したものであったが，9 月の下降は株式相場の状況自体が変動要因になっていたと推測される．

表 1: 2008 年 4,6,9 月の変動要因となった主成分. 各主成分で負荷量の絶対値が上位 10 個のキーワードを示す.

主成分 3		主成分 4		主成分 5	
背景	0.655	設備	0.468	足許	-0.458
伴う	0.494	国内	-0.432	上昇	-0.436
需要	0.452	低迷	0.421	実体	-0.401
改善	-0.424	輸出	-0.411	年末	-0.394
生産	-0.421	歯止め	0.36	頭打ち	-0.394
鈍化	-0.404	掛かる	0.36	先行き	-0.381
軟調	-0.394	総合	0.36	厳しい	0.374
国債	-0.394	対策	0.36	間	-0.363
利回り	-0.394	ベース	0.358	軟化	-0.355
格差	-0.394	踏まえる	0.354	ベース	0.352

このように，提案手法によって，日銀金融経済月報で解説されている経済状況と，発表月の月末までの市場変動の関係性について，詳しく分析することができた．このような分析が可能であるのは，数値情報よりも豊かな背景情報を含んでいるテキスト情報を，市場分析に用いているからである．

4. 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．本手法により，月次の日本国債市場・株式市場・外国為替市場データの分析を行った結果，日本国債 2 年物・日本国債 5 年物 > 日本国債 1 年物・日経平均株価 > 日本国債 10 年物 > 円ドルレートの順で，外挿予測力の精度が高かったことがわかった．さらに，日経平均株価について 2008 年 4 月の上昇局面と 6 月と 9 月の下降局面の変動要因を，主成分の値の変化より分析した．その結果，4 月と 6 月の変動は経済の基礎的要因と関連しており，9 月の下落は市況自体の要因が関連したとの示唆を得た．

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)

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

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

[三菱 08] 三菱東京 UFJ 銀行：テキストマイニング手法を用いた経済市場分析の試み, *Focus on the Markets*, Vol. 24, (2008)

[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)

[電気 02] 電気学会（編）：学習とそのアルゴリズム, 第 6 章, 森北出版 (2002)

[日本 93] 日本ファジィ学会（編）：ファジィ・エキスパート・システム, 日刊工業新聞社 (1993)