# テキスト分析による業種別平均株価の動向推定

# Trend Estimation of Industrial Stock Price Indexes by Text Analysis

和泉 潔[1,2*]　牧野 龍彦[1]　石田 智也[3]　中嶋 啓浩[3]
松井 藤五郎[4]　吉田 稔[5]　中川 裕志[5]　本多 隆虎[6]
Kiyoshi Izumi[1,2]　Tatsuhiko Makino[1]　Tomonari Ishida[3]　Nobuhiro Nakashima[3]
Tohgoroh Matsui[4]　Minoru Yoshida[5]　Hiroshi Nakagawa[5]　Takatora Honda[6]

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

**Abstract:** In this study, we proposed a new text-mining method for stock price indexes using newspaper articles. Using this method, we conducted extrapolation tests to evaluate the prediction accuracy for the year 2009. As a result, 11 sectors in 19 sectors (57.8 percent) showed over 52% accuracy. The prediction accuracy showed seasonality in some sectors. This is expected to be a measure of prediction confidence of text mining.

## 1　はじめに

金融市場では常に様々な情報が溢れている．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．市場の分析に用いる情報には大きく分けて 2 種類がある．一つは，経済指標，マーケットのテクニカル指標等の数値情報である．もう一つは，市場に関わる要人の発言，中央銀行や他の市場参加者の解析記事などのテキスト情報である．これらの多様な情報が瞬時にトレーダー達のもとに，オンラインで送られてきているのである．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 [6] 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 [3]．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もともと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．しかも，指標を集計して発表するには，どうしてもタイムラグが生じてしまうので，分析への反映も遅れがちである．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 [1,8,9]．このようなテキストマイニング技術を金融市場分析に用いることは，近年金融実務家の間でも注目され始めている．筆者らは日本銀行の金融経済月報を題材に，月次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 [4,5]．本稿は，この技術をより期間の短い日次の業種別の平均株価に適用することを行った．分析対象とするテキスト情報も，より構造が多様な新聞記事を用いた．

*連絡先：東京大学大学院 工学系研究科 システム創成学専攻
〒 113-8656 文京区本郷 7-3-1
E-mail: izumi@sys.t.u-tokyo.ac.jp

## 2 テキストマイニング手法

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** [4, 5] を適用する．

### 2.1 共起関係に基づく主要単語の抽出 (C)

今回の分析対象テキストは，日本経済新聞 本紙に掲載された地方面を除く記事である．分析対象範囲は 24 時間以内に配信された記事，つまり当日の本紙朝刊，本紙夕刊の見出しと本文である．各 24 時間での記事数は 300 から 500 個であり，テキストファイルサイズは 300 から 500KB になった．文字数では約 10 万から 17 万個，単語数ではのべ約 5 万から 9 万語であった．

本手法の第 1 ステップとして，各期間で配信された記事の集合から，記事テキストにおける共起関係によって特徴量を計算する．最初に，Chasen [2] による形態素解析を行い，名詞・動詞・形容詞を基本形に変換して抽出した．次に，各形態素の組み合わせに関して，同じ文の中で隣接して出現した回数を数える．これは，単一の形態素の頻度よりも，形態素の組の共起頻度の方が，記事が表す経済状況に関する情報をうまく抽出できると考えたからである．例えば，単に「介入」という単語の頻度を見るよりも,「介入-実施する」や「大規模-介入」のような単語の組の方が状況の変化をよく表すことができると考えられる．共起頻度を計算する際に，できるだけ市場分析に関連するような用語のみを抽出できるように，単語の組の少なくとも一方が経済に関する用語を含む組み合わせのみを対象とした．経済に関する用語は，日本経済新聞デジタルメディアが作成している日経シソーラス [7] に収録されている約 1 万 3 千語に含まれている用語とした．本ステップの最後に，各期間に配信された全ての記事での共起頻度を合計する．共起頻度の合計がある閾値以上の単語の組み合わせに現れた単語を，その期間での主要単語と定義する．今回は閾値は 4 回とした．

### 2.2 主成分分析による単語のグループ化 (P)

前ステップで抽出した各期間での主要用語の出現パターンから，主要単語のグループ分けを行う．今回は，過去 1 年間 (約 250 営業日) の新聞記事データでの出現パターンから主成分分析を行った．各 24 時間毎でどの主要単語が出現したか (1)/出現しなかった (0) をベクトル表示し，過去 1 年間のベクトルを結合して行列を作成する．この行列に対して主成分分析を実施し，100 個の合成変数 (主成分) にまとめる．各 24 時間での 100 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，100 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，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，各主成分スコアの各期間の動きから日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 100 個の主成分スコアの時系列データを説明変数として，各日の終値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，訓練に使われていない最新のテキストデータを入力すれば，当日の終値を推定 (外挿予測) できる．

本研究では，日本の株式市場での業種毎の株価指数を予測対象とした．用いた指数は NOMURA400[1] である．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を数値化した時価総額加重平均の株価指数である．今回は表 1 の 19 業種を分析対象とした．

回帰分析の非説明変数 $r'_{i,t}$ は業種 $i$ の日次超過リターンとした．超過リターンとは，業種 $i$ の前日比での価格変動率 $r_{i,t}$ が，基準株価の変動率 $R_t$ をどれくらい上回ったか，もしくは下回ったかを示すものである．本研究では基準株価として TOPIX を用いた．

$$r'_{i,t} = r_{i,t} - R_t \tag{1}$$

過去 1 年間 (約 250 営業日) の新聞記事データと株価データを用いて，各業種毎に次の回帰式を推定した．

$$r'_{i,t} = a_{i,0} + \sum_{j=1}^{100} a_{i,j} x_{j,t} \tag{2}$$

ここで．$x_{j,t}$ は期間 $t$ の新聞記事データから計算された第 $j$ 主成分のスコアである．回帰分析の際に AIC 基準に基づくステップワイズ選択を行い，説明力の低い主成分は説明変数として使用しなかった．その結果，各業種での回帰式は大体 30 個程度の主成分を用いた式となった．過去 1 年間のデータから得られた 2 式に，新たな新聞記事の主成分スコアを入力することにより外挿予測ができる．

[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，完全性，信頼性，有用性を保証するものではなく，本論文に関し一切責任を負いません．

| | |
|---|---|
| 1 | NOMURA400 サービス |
| 2 | NOMURA400 ソフトウェア |
| 3 | NOMURA400 メディア |
| 4 | NOMURA400 住宅不動産 |
| 5 | NOMURA400 公益 |
| 6 | NOMURA400 化学 |
| 7 | NOMURA400 医薬・ヘルスケア |
| 8 | NOMURA400 商社 |
| 9 | NOMURA400 家庭用品 |
| 10 | NOMURA400 小売り |
| 11 | NOMURA400 建設 |
| 12 | NOMURA400 機械 |
| 13 | NOMURA400 自動車 |
| 14 | NOMURA400 通信 |
| 15 | NOMURA400 運輸 |
| 16 | NOMURA400 金融 |
| 17 | NOMURA400 鉄鋼非鉄 |
| 18 | NOMURA400 電機・精密 |
| 19 | NOMURA400 食品 |

表 1: 分析対象の株価指数 (順不同)

## 3 分析１: 2009 年の外挿予測

上述の手法を用いて，2009 年の 1 年間を対象に外挿予測精度を評価した．手順は下記のようになる．

1. 1 年前から前月末までの新聞記事データと価格指標データを用いて，共起解析・主成分分析・回帰分析を行い，各 19 業種について 2 式を求める．例) 2008 年 1 月 1 日から 2008 年 12 月 31 日までのデータで訓練.
2. 求めた回帰式に，翌月の各日のテキストデータの主成分スコアを入力し，その日の超過リターンを推測する．例) 2009 年 1 月 1 日から 2009 年 1 月 31 日.
3. 訓練期間と外挿期間を 1ヶ月ずつ移動して，上記の手続きを繰り返す.
   - 訓練期間
     2008 年 1 月 1 日から 2008 年 12 月 31 日
     ↓
     外挿期間
     2009 年 1 月 1 日から 2009 年 1 月 31 日.
   - 訓練期間
     2008 年 2 月 1 日から 2009 年 1 月 31 日
     ↓
     外挿期間
     2009 年 2 月 1 日から 2009 年 2 月 28 日
     ...
   - 訓練期間
     2008 年 12 月 1 日から 2009 年 11 月 31 日
     ↓
     外挿期間
     2009 年 12 月 1 日から 2009 年 12 月 31 日.

変動の方向性の予測力を評価するために，外挿期間において超過リターンのの予測値が実際の超過リターンと符号 (上下) が合っていた日数の割合 (正答率) を比較した．結果を図 1 に示す．その結果，とりあえずの予測精度の目標とした 52%を超えた業種は，19 業種中 11 業種 (57.8%) であった.

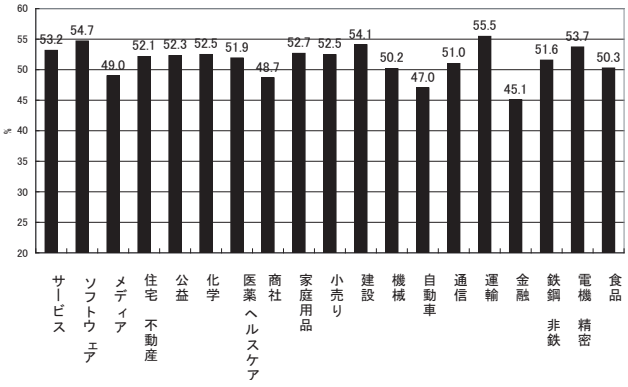


図 1: 2009 年の外挿結果

## 4 分析 2: 2003-2008 年の予測力の変化パターン

本研究での回帰式は 1ヶ月毎に更新されるわけであるが，月によって外挿予測の正答率にある程度のばらつきがある業種があった．このような予測力の時系列変化に業種によってはパターンがあるかを調べた.

前節と同様の手法によって，2003 年から 2008 年までの 6 年間の 19 業種の超過リターンを予測した． 外挿した 2003 年から 2008 年までの 6 年間の予測正答率の平均は，常に正答率が高い，もしくは低いという業種は見られなかった．しかし，時期・業種によっては予測正答率に差が見られた．テキストマイニングの予測力の季節性を調べるために，各四半期ごとに 6 年間の業種別予測正答率平均を計算した．19 業種の各四半期の平均正答率から主成分分析を行い，予測正答率の傾向によって業種の分類を行った.

その結果，図 2 に見られる 2 つの主成分が抽出できた．第 1 主成分は第 1 四半期 (1-3 月, 1Q) と第 2 四半期 (4-6 月, 2Q) の予測正答率が高いと正の値になり，第 4 四半期 (10-12 月, 4Q) の予測正答率が高いと負の値になる．1 年間の前半と後半のどちら予測精度が高いか

を表している．第 3 四半期 (7-9 月, 3Q) の予測正答率が高いと負の値になり，それ以外の時期の予測正答率が高いと正の値になる．夏場に予測精度が高いかを表している．

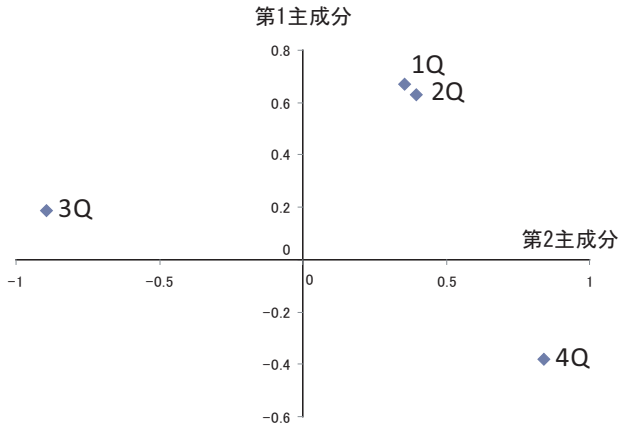


図 2: 各主成分の因子負荷量

これら 2 つの主成分に対する 19 業種の得点を図 3 に示す．その結果，以下のようなことが分かった．

- 自動車および商社は第 2 主成分の値が負になる．夏場にテキストマイニングの予測正答率が上がる．
- 電機精密および化学は第 2 主成分の値が正になる．夏場以外にテキストマイニングの予測正答率が上がる．
- 家庭用品は第 1 主成分の値が正になる．年の前半にテキストマイニングの予測正答率が上がる．
- 公益は第 1 主成分の値が負になる．年の後半にテキストマイニングの予測正答率が上がる．

前述の結果を確かめるために，特にテキストマイニングの予測正答率が 80%近くと高かった，自動車の 2004 年 7 月の外挿予測について調べてみた．この月の外挿予測に用いられた主成分を調べてみると,「ガソリン」や「ナフサ」「出光興産」といった原油関係の用語を含む主成分が複数見られた．記事の中身を見ても，夏場は通常でもガソリンの需要が高い時期であり，さらに原発停止やイラク相場などの要因により原油高となっていた．このように，自動車業界の株価は特に夏場に原油価格の動向に関連するニュースに敏感に反応していることを示していた．

## 5　まとめ

本研究では，新聞記事データを用いた業種別株価指数の分析の新たな手法を提案した．本手法を用いて，2009

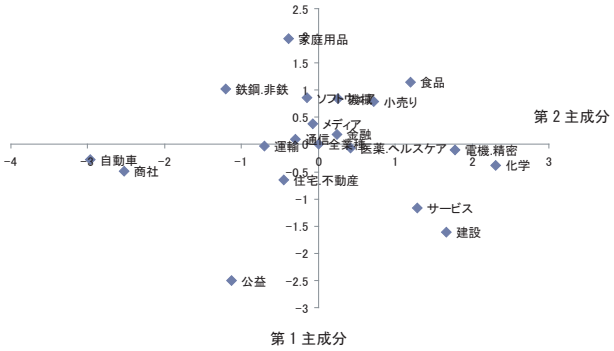


図 3: 四半期ごとの業種別予測正答率の平均をもとに主成分分析を行った結果．縦軸が主成分 1 の得点，横軸が主成分 2 の得点．

年の 1 年間を対象に外挿予測精度を評価した結果，予測精度の目標とした 52%を超えた業種は，19 業種中 11 業種 (57.8%) であった．また，予測正答率は時期・業種によって予測正答率の季節性が見られた．これにより，期間毎のテキストマイニングによる予測の信頼度を測る指標になることが期待できる．

今後は，テキストマイニング手法の改善により予測正答率の向上を目指す．例えば，係り受けなどの文章構造を考慮した単語頻度の計算が考えられる．それと同時により詳細な予測精度の変化パターン分析を行うことにより，テキストマイニングの信頼度を計算し，他の予測手法とハイブリッドする手法の開発を目指したい．

## 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」の公募研究 B01-12「金融市場分析のための経済情報抽出と意思決定支援に関する研究」(No. 21013049) の助成を受けています．お礼申し上げます．

## 参考文献

[1] K. Ahmad, L. Gillam, and D. Cheng. Textual and quantitative analysis: Towards a new, e-mediated social science. In *Proc. of the 1st International Conference on e-Social Science*, 2005.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 日本経済新聞デジタルメディア. 日経シソーラス. http://telecom21.nikkei.co.jp/help/contract/price/00/help_KIJI_thes.html.

[8] Young-Woo Seo, Joseph Andrew Giampapa, and Katia Sycara. Financial news analysis for intelligent portfolio management. Technical Report CMU-RI-TR-04-04, Carnegie Mellon University, 2004.

[9] 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.