2H1-OS18-6

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] Kiyoshi Izumi　池田 翔[*1] Shou Ikeda　石田 智也[*3] Tomonari Ishida　中嶋 啓浩[*3] Akihiro Nakashima　松井 藤五郎[*4] Tohgoroh Matsui　吉田 稔[*5] Minoru Yoshida
中川 裕志[*5] Hiroshi Nakagawa　本多 隆虎[*6] Takatora Honda

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

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, filtering using a term list generated by dependency parsing could improve predictability of industry-classified average prices. This is expected to be a measure of prediction confidence of text mining.

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等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．例えば，Yahoo! Finance の記事から米国の個別銘柄の 20 分後の株価動向を予測する研究 [Schumaker 10] や個別銘柄の日次の株価変動と新聞記事との関連を調べた研究 [張 08, 小川 01] などがある．既存の研究は，ニュースがもたらす数分から数時間以内の短期的な反応を分析対象とし，特定のキーワードがもたらす個別銘柄への単発的なインパクトを扱っていることが多かった．しかし金融実務者からは，複数のニュースの組み合わせがもたらす広範囲な経済状況の変化が，複数の銘柄に与える影響を推測したいという要望があった．そこで本研究は，一定期間のニューステキストの集合が，自動車や食品などの各業種ごとに与える影響を推測する手法を新たに開発した．

2. テキストマイニング手法

テキスト情報が長期的で広範な市場動向に与える影響を推測するための手法として，和泉らの研究 [和泉 10, 和泉 11] では，日本銀行が毎月発行する金融経済月報を分析し日本国債の月次価格の変動を予測している．解析には，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手法 (CPR 法) を用いている． CPR 法は市場の大きな変動の推定に対して特に有効であった．

そこで本研究においては新聞記事を用いた日次の業種別株価の変動について CPR 法をベースに手法を開発した．本研究では日銀の金融経済月報ほど形式が定まっていない新聞記事を分析するため，CPR 法をそのまま適用しても有効に市場動向を推定できないと予想された．そこで，新たに文構造を考慮した解析を行うため係り受け解析を導入した．

連絡先: 和泉 潔，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， izumi@sys.t.u-tokyo.ac.jp

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．

2.1.1 経済用語との共起頻度の計算

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 [ChaS] 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 [日本] 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．

共起解析を行う際に，過去のテキストでの係り受け関係を考慮して作成した辞書 (係り受け辞書) を使用して単語のフィルタリングを行った場合と，現在の文で隣接するかどうかだけ

で分析した場合の 2 つのオプションを用意した．両者の比較により，経済用語とよく係り受け関係にある単語に絞ることによって，市場動向推定に有効な単語のみを抽出できるようになるか調べた．

1. 係り受け辞書を用いた共起解析
   まず，過去の記事において日経シソーラスの語と係り受け関係にある単語の組で出現回数の多いものを辞書として登録する．本研究では辞書登録された組み合わせのうち上位約 1%を用いた．次に解析期間の全記事に対して，同一文中に辞書に載っている単語の組み合わせがあった場合に一回共起したとする．

2. 係り受け辞書を用いない共起解析
   解析期間の全記事において日経シソーラスの語と隣接する単語との組を共起とする．

2.1.2 構文解析による係り受け辞書の作成

上記の係り受け辞書を用いた共起解析を行う場合には，市場動向分析を行う時期よりも以前の期間 (本研究では 5 年間) のテキストデータを構文解析し辞書を作成する．構文解析には CaboCha [Cabo] を用いた．例えば,「中国の需要増で石炭や鉄鉱石の価格が高騰，鉄鋼メーカーの生産コストが上がっている.」という文の構文解析結果が図 1 になったとする．ただし，動詞・名詞・形容詞の原形のみを示している．

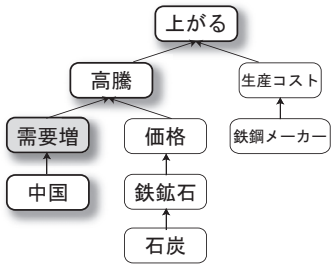

図 1: 構文解析結果の例

「需要増」という名詞が日経シソーラスに現れる専門用語だったとする.「需要増」という語と係り受け関係になり得る単語として，係り受け元の「中国」や係り受け先の「高騰」「上がる」がカウントされる．特に,「高騰」や「上がる」といった語は，文中では隣接していないので，構文解析でないと結びつきが見つけられなかった単語である．このように，過去のテキストデータに関して，係り受け関係にあった頻度の高かった単語を，各専門用語に対してリストアップする．これが各専門用語に対して意味的なつながりの強い係り受け辞書となる．

市場動向予測時の共起解析では，一文中に係り受け辞書にある専門用語と係り受け関係の強い単語の組み合わせが共起する頻度を計算する．これにより，できるだけ市場動向予測に無意味な単語の組を排除しながら，共起の範囲を広げることができると期待される．

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 時間毎である閾値 n 回以上の共起頻度があった単語の組み合わせに含まれるかどうかで，各単語が出現したか (1)/出現しなかった (0) をベクトル表示した．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．

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 [Akaike 74] 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．

3. 2009 年の主要 5 業種の予測

前節の手法を用いて 2009 年一年間の株価変動予測を行った．予測対象月の前月末までの 1 年間を訓練期間として回帰式の作成を行った．つまり 1ヶ月ごとに回帰式の更新を行った．また，係り受けによる共起解析の辞書作成期間は 2004 年から 2008 年の 5 年間とした．

3.1 共起解析の 4 条件

前述のように，本研究では共起解析の手順を変えて，実験を以下の 4 条件で行った (表 1).

*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．

表 1: 共起解析の条件

	係り受け辞書	係り受け範囲	共起範囲	使用品詞
条件 1	使用しない	−	隣接語	動詞，名詞，形容詞
条件 2	使用する	係り受け先，係り受け元，同レベル	一文	動詞，名詞，形容詞
条件 3	使用する	係り受け先	一文	動詞，名詞，形容詞
条件 4	使用する	係り受け先	一文	動詞，名詞-サ変接，名詞-副詞可

条件 1 係り受け辞書を用いずに隣接関係による共起頻度を計算した．主成分分析時の出現/非出現の閾値 n を共起頻度 4 回とした．形態素解析時に，動詞・名詞・形容詞を抽出する．

条件 2 係り受け辞書を用いた共起．構文解析時に，経済用語への係り受け元を 1 語遡った単語，構文解析時に経済用語と同じレベルに合った単語，経済用語からの係り受け先を 2 語分の範囲に出現した語を係り受け辞書に登録した．共起頻度の閾値は 3 回．形態素解析時に，動詞・名詞・形容詞を抽出する．

条件 3 係り受け辞書を用いた共起．経済用語からの係り受け先を 2 語分の範囲に出現した語を係り受け辞書に登録した．共起頻度の閾値は 3 回．形態素解析時に，動詞・名詞・形容詞を抽出する．

条件 4 係り受け辞書を用いた共起．係り受けによる辞書作成時に，係り受け関係において日経シソーラスの語の係り受け先の 2 語分の範囲に出現し，かつ品詞が動詞，名詞-サ変接続，名詞-副詞可能である語で最も先に出現した語を辞書に登録した．これらの品詞は，動きを表す語が多く株価変動に関わる品詞だと考えたためである．本条件では辞書による絞り込みが強いので，共起頻度の閾値は 1 回とした．

各条件において共起解析の結果，語数が同程度になるように共起頻度の閾値 n をそれぞれ設定している．

3.2 予測精度の評価手法

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. (係り受け辞書を使用する場合): 2004 年から 2008 年の 5 年間の新聞記事データから，日経シソーラスに含まれる各専門用語と係り受け頻度の高い単語をリストアップする．辞書に登録する係り受け回数の閾値は，係り受け先か係り受け元，同じレベルにあるかによって異なる．それぞれの頻度の上位 1%として，5 年間にそれぞれ 13 回，19 回，19 回以上出てきた組み合わせを係り受け辞書として使用した．

2. 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．

3. 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．

4. 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．

3.3 外挿予測結果

今回予測対象とした業種は，メディア，医薬・ヘルスケア，自動車，電機・精密，食品の金融実務家に挙げてもらった主要 5 業種とした．共起解析の 4 条件における予測正答率を表 2 に示す．

	条件 1	条件 2	条件 3	条件 4
1. メディア	48.35	51.61	54.76	**57.81**
2. 医薬・ヘルスケア	48.53	51.56	50.84	**54.06**
3. 自動車	54.87	49.23	52.13	**57.68**
4. 電機・精密	**55.48**	47.52	50.60	51.16
5. 食品	51.10	50.70	**51.47**	45.01
平均	51.67	50.12	51.96	**53.14**

表 2: 2009 年の主要 5 業種の市場動向予測の正答率 (%): 太字は最も精度が高かった条件．

全体的に，係り受けを用いて日経シソーラスの語と動きを表す語による解析を行うことで予測的中率を上げることが出来たと言える．より詳細に共起解析において抽出した語の組み合わせを見ると，条件 1 では言い換えなどの普通名詞同士の組み合わせが多く，条件 2 および 3 では数値なども多く抽出してしまっていた．

一方条件 4 では「円高-進む」や「原油価格-上昇」というように動きを表す組み合わせが多く抽出できていた． また，主成分分析の結果，2008 年の主成分分析の第 1 主成分の因子負荷量絶対値の上位を見ると「連結決算」「特別損失」「経常利益」など決算関連，第 2 主成分は金融政策関連の語が並ぶなど主成分が分野を表すような傾向が他の条件に比べてはっきりと見られた．

さらに条件 4 では回帰式において AIC 基準による変数選択で選択された説明変数 (主成分) と業種との間に，明確な関連性を見つけることが比較的容易であった．例えば，2009 年 5 月の予測に用いた自動車業界の回帰式における係数が大きいものとして第 30 主成分および第 32 主成分があった．第 30 主成分の因子負荷量が大きい単語として，上位から 4 番目に「販売台数」，7 番目に「世界戦略車」という自動車業界に関連の深

表 3: 条件 4 において 2009 年 5 月の予測に用いた第 30，32 主成分の因子負荷量の絶対値が大きい単語上位 10 個

	第 30 主成分		第 32 主成分	
1	道州制	0.229	土砂災害	0.258
2	終戦記念日	0.225	株価指数先物	0.237
3	参拝	0.225	**燃費規制**	-0.236
4	**販売台数**	-0.213	電話番号	0.233
5	特許権	0.209	訪問	-0.232
6	議決	0.206	健康食品	-0.228
7	**世界戦略車**	0.206	仲介	-0.216
8	実感	-0.206	品質管理	-0.216
9	IP 電話	0.202	物価上昇率	-0.213
10	システム開発	-0.202	多国籍軍	0.213

い単語が含まれていた．さらに，第 32 主成分の 3 番目に「燃費規制」という単語が見られた (表 3)．

以上の事から係り受けを用いた共起解析によって単語抽出を改善することで，主成分分析及び回帰分析においても精度を上げる事が出来たといえる．

しかし，食品業界における予測精度を向上させることはできなかった．このことに関してはいくつか理由が考えられる．一つには食品業界における株価変動の要因によるものと考えられる．食品業界の株価は，原材料の価格，商品先物価格に左右される．しかし，これらの情報は今回分析に用いた新聞記事では比較的記事になることが少ない．もう一つは今回用いた日経シソーラスによる問題が考えられる．食品業界の株価に影響を与えるような語は一般的な語が多く，日経シソーラスに含まれる経済専門用語だけに着目して共起解析を行うことが適切ではなかった可能性がある．

4. 結 論

本研究では，新たに日次の業種別株価動向推定に月次の市場分析に使われていた CPR 法を適用できることを示した．また新聞記事という日本銀行のレポートよりも構造が多様なテキストに用いるため，CPR 法の共起解析において係り受けを用いて文構造を考慮した解析を行う方法を提案し，その有効性について検証を行った．その結果，特に，株価に関連する語と「動詞」「名詞-サ変接続」「名詞-副詞可能」という動きを表す語との共起を用いて予測を行うことで，主成分分析，回帰分析においても，株価変動との相関が強くみられるようになり，より効果的な予測が出来るようになったことが確認された．

今後さらに的中率を上げるためには，様々なアプローチが考えられる．本研究においては係り受けを用いて，さらに品詞を限定することで精度を上げたが，例えば,「上昇」や「上げる」などの同じ意味を持つ単語についての表記ゆれの問題を解決する事や「ない」などの否定語が出てきた場合には逆の意味を持つようにするなどといった事も効果的かもしれない．また，予測期間，予測対象を絞ることで的中率を向上させることもできる．的中率が高い期間を分析し，例えばある業種は夏場によく当たるとか，株価の変動の動きが大きいときによく当たるなどといった事がもし分かれば，十分実務に応用できる．他にも，共起解析において基準となる株価変動に関連する語として，日経シソーラスではなく業種毎の重要語のリストを用いることで，その業種に特化した予測が出来るようになるという事も考えられる．

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

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

[Cabo] CaboCha ホームページ:http://chasen.org/~taku/software/cabocha/

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

[張 08] 張 へい, 松原 茂樹：株価データに基づく新聞記事の評価, 第 22 回人工知能学会全国大会論文集 (2008)

[和泉 10] 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)

[和泉 11] 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)

[日本] 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

[小川 01] 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, 情報学処理学会研究報告, 第 NL142-19 巻, pp. 137–144 (2001)

[Schumaker 10] 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)