4E1-OS-15-4

新聞記事のテキストマイニングによる長期市場動向の分析

Analysis of Long-term Market Trend by Text-Mining of News Articles

藏本 貴久[*1] Takahisa Kuramoto
和泉 潔[*1] Kiyoshi Izumi
吉村 忍[*1] Shinobu Yoshimura
石田 智也[*2] Tomonari Ishida
中嶋 啓浩[*2] Akihiro Nakashima
松井 藤五郎[*3] Tohgoroh Matsui
吉田 稔[*4] Minoru Yoshida
中川 裕志[*4] Hiroshi Nakagawa

[*1] 東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo
[*2] 野村證券株式会社※
Nomura Securities Co., Ltd
[*3] 中部大学 生命健康科学部
College of Life and Health Sciences , Chubu University
[*4] 東京大学 情報基盤センター
Information Technology Center, The University of Tokyo

Abstract: In this study, we developed a new method of the long-term market analysis by using text-mining of news articles. Using our method, we conducted extrapolation tests to predict stock price averages by 19 industry and two market averages, TOPIX and Nikkei225 for about 10 years. As a result, 8 sectors in 21 sectors (about 40%) showed over about 60% accuracy, and 15 sectors in 21 sectors (over 70%) showed over about 55% accuracy. We also developed a web system of financial text-mining based on our method for financial professionals.

1. はじめに

経済の変動は経済指標や株価に代表される数値データ，ニュースや新聞に代表されるテキストデータなど，膨大な情報が相互に作用し合って決定されるため非線形な挙動を示す．それに対して投資家はテクニカル分析やファンダメンタルズ分析のような分析手法を駆使して投資を行っているが，それでも長期の変動の予測は非常に困難なのが現状である．これは情報量の膨大さ，情報間の関係性の複雑さに起因しており，その情報全てを瞬時に解釈して適切な投資を行うことは不可能に近い．長期の変動においては，変動を決定する情報もさらに膨大になり，その関係性も不明瞭になるため，投資判断は益々難しくなる．

これまでの研究ではニューラルネットワーク[高穂 2002]や遺伝的プログラミング[伊庭 2006]といった機械学習の手法によって経済の変動を自動的に分析することを目指したものが存在する．しかし，数値情報の解析結果のみで長期的な投資判断を下すことは難しい．その理由として，人間が解釈を行えないほど複雑な最適化が生じることや，数値外の情報が欠落してしまっていることが挙げられる．そこでテキスト情報を用いて経済の変動を分析する研究が近年現れてきた．和泉らは CPR 法という手法を開発し，テキストを解析することで経済動向の分析を行っている[和泉 2011]．テキストを用いる利点としては解釈が容易であること，数値に含まれない情報も分析できることの 2 点が挙げられる．

本研究ではテキスト情報を用い，投資家からの要望が高い長期的な経済動向の分析を行う．さらに膨大な情報の中から投資に有益な情報を抽出し，その情報間の関連性を明確にすることで投資活動に役立つ情報を提供する．

※本研究の内容は野村證券株式会社の公式見解を示すものではありません．

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. 分析手法

本研究では和泉らが提唱する CPR 法を用いて分析を行う．CPR 法とは共起解析，主成分分析，回帰分析の三段階からなる分析手法である[和泉 2011]．従来の CPR 法で入力テキストとしていた日本銀行の金融経済月報は比較的形式の決まった文書であるため，非常に扱いやすいものであった．本研究では新聞記事という形式が金融経済月報よりも定まっていない文章を用い，より広範で長期的な分析が可能となるように CPR 法を拡張した．

2.1 共起解析

本研究で用いたテキスト情報は日本経済新聞である．経済紙であるため，経済の変動を決定する要因が掲載されていると考えられる．

まず地方面を除く全記事に対して ChaSen[ChaS]を用いて形態素解析を行い，動詞・名詞・形容詞を抽出する．そして同一文中に隣接して出現した単語のうち，少なくとも一方に日経シソーラス[日経]に収録されている経済専門用語が含まれる組合せのみを数え上げる．日経シソーラスとは日本経済デジタルメディアが公開している新聞記事検索のための辞書であり，約 1 万 3 千語が収録されている．単語ごとの出現頻度を数え上げるよりも，隣接した共起関係の出現頻度を数え上げることで経済動向に関する情報をうまく抽出でき，解釈も容易になると考えられる．従来の CPR 法では KeyGraph アルゴリズム[大澤 2006]を用いて重要語を求めているが，本研究は新聞記事を用いた長期予測であるため，日経シソーラスの語との共起をとることで網羅的な情報を抽出した．この数え上げを 1 ヶ月間の記事で繰り返し行い，閾値以上なら出現，閾値未満なら不出現とし，出現パターンを定義した．なお，閾値は 30 とした．1 ヶ月間に出現した共起関係のうち少なくとも一方に日経シソーラス

の単語を含むものは約 10 万組存在し，閾値以上の共起関係の組合せは 400 から 500 組であった．

2.2 主成分分析

2.1 節の共起解析を過去 3 年間(36 ヶ月)の新聞記事に対して行い，閾値以上ならば 1，閾値未満ならば 0 とし，1 ヶ月ごとの単語の出現パターンを結合した行列を作成する．この時，訓練期間で少なくとも一回は出現した単語数は約 2 千語であり，36 行 2000 列の行列が作成されたことになる．この行列に対して主成分分析を行い，各月 15 個の主成分で記事を評価した．すなわち，1 ヶ月間の新聞記事を 15 次元のベクトルで評価し，そのベクトルを結合することで新聞記事の特徴量の時系列データが得られたことになる．

2.3 回帰分析

株価データは NOMURA400 の 19 業種，さらに日経平均，TOPIX を用いた．NOMURA400 とは，野村證券金融工学研究センターが提供しているデータであり，日本株式市場の全銘柄から選定された市場代表性の高い投資ユニバースである．構成銘柄は日本株式市場の全銘柄の中から，アナリストの意見を基に選定された市場代表性の高い 400 銘柄である[野村]．そのうち，化学，鉄鋼・非鉄，機械，自動車，電機・精密，医療・ヘルスケア，食品，家庭用品，商社，小売り，サービス，ソフトウェア，メディア，通信，通信建設，住宅・不動産，運輸，公益，金融の 19 業種が対象である．日本の市場の分析には適している指標だと判断した．また，市場全体の動きを把握するための指標として日経平均，TOPIXも予測指標に加えた．指標i の時刻t における株価を$p_{i,t}$とすると，単位期間Δt (1 ヶ月，2 ヶ月，3 ヶ月)でのリターン$r_{i,t}$は下式で定義できる．

$$r_{i,t} = \frac{p_{i,t+\Delta t} - p_{i,t}}{p_{i,t}}$$

過去 3 年間の新聞記事から得られた主成分スコアと株価データを用いて次の回帰式を推定する．

$$r_{i,t} = a_{i,0} + \sum_{k=1}^{15} a_{i,k} x_{k,t}$$

$x_{k,t}$ とは時刻tにおける第k主成分の主成分スコアである．回帰分析の際，AIC 基準[Akaike 74]におけるステップワイズ選択を行い，説明力の低い説明変数は回帰式に含めていない．

3. 外挿予測結果

2 節の手法を用いて外挿予測の精度検証を行った．回帰式の推定および外挿予測の手順は以下の通りである．

1. 1998 年 1 月 1 日~2000 年 12 月 31 日の 3 年間(36 ヶ月)の新聞記事で説明変数の訓練データを作成する．

2. 各月の説明変数で翌月末の終値を被説明変数とする重回帰式を推定する．ただし，この時推定するのは 2.3 節で示したリターンの値である．

3. 推定された回帰式に直近のテキストデータ(2001 年 1 月 1 日~2001 年 1 月 31 日)から得られた主成分スコアを代入することで翌月末(2001 年 2 月末)の終値の推定を行う．

4. 訓練データの作成開始日を 1 ヶ月ずつ遅らせることで回帰式を毎月更新していき，2010 年 10 月末までの推定を行う．

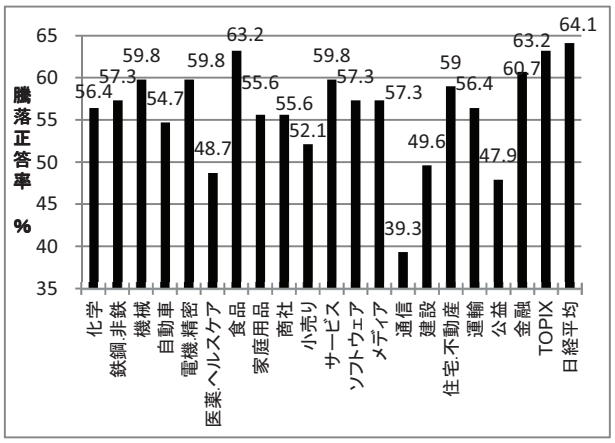
1 ヶ月後の予測においては 9 年 9 ヶ月の 117 回の推定を行った．また，2 ヶ月後・3 ヶ月後の予測の場合，リターンにおける単位期間Δtを調整することで推定を行った．

3.1 1 ヶ月後の外挿予測結果

外挿予測の騰落正答率の結果を図 1 に示す．この結果は外挿予測を行った回数のうち，騰落が一致していた回数の割合を百分率で表したものである．

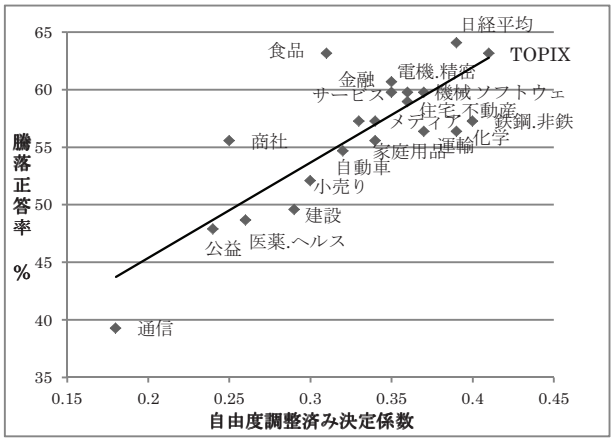
1 ヶ月後の予測の騰落正答率は市場全体の動向を表す TOPIX や日経平均といった市場平均株価で 63.7%であった．また，投資判断への適用可能性の目安となる 55%以上の正答率は 7 割以上の業種(予測指標 21 のうち，15 指標)で達成することができた．55%という騰落正答率は AI を用いた実際のファンドも目指している数値である[AERA]．本研究では日次よりも長期の月次予測で，しかも約 10 年間という長期間の外挿予測テストで安定した結果を示している．これは投資家から要望の高い長期予測の精度を大幅に改善したということである．

図 1: 1 ヶ月後の外挿予測における騰落正答率



1 ヶ月後の予測においては，訓練期間の重回帰分析の自由度調整済み決定と外挿期間の騰落正答率に正の相関が見られた．その様子を図 2 に示す．図中の直線は一次の近似曲線である．

図 2: 予測正答率と自由度調整済み決定係数の相関



自由度調整済み決定係数とは回帰式の当てはまり度を表す指標であり，回帰式が過去の変動をよく説明できているほど騰落正答率が高くなる傾向が見られた．これは訓練期間で当てはまりのよい回帰式を作成できているほど外挿予測の精度も高まるということであり，

内挿の段階で外挿予測の精度を大まかに推し量ることができる. また, 図2からも明らかなように, 本研究では過剰適合が生じていない. 過剰適合とは, 訓練期間で当てはまりの高いデータを作成するあまりに汎用性に欠け, 未来のデータの予測には適さない形となってしまうことである. 新聞という経済動向に関する網羅的な情報を扱うテキストデータから, 主成分分析によって合成変数を作成し, さらに回帰分析の際にステップワイズ選択を用いることで過剰適合が回避されたと考えられる.

一方で, 通信や公益や建設, 医療・ヘルスケアに関しては自由度調整済み決定係数, 外挿予測の騰落正答率ともに低い水準であった. これらの産業が内需産業と呼ばれ, 新聞記事からこれらの産業に関する特徴量がうまく抽出できなかったためであると考えられる.

実際に抽出された共起単語のペアを見てみると, 「長期金利-上昇」や「デフレ-克服」などといった経済全体に作用すると考えられる共起, 「米国-トヨタ自動車」や「住宅-融資」のように個別の業種に影響を及ぼすと考えられる共起が見られた. これらの出現パターンを用いて回帰式を推定しているため, 特徴量がうまく抽出できる業種とそうでない業種が存在する. それが図 2 の自由度調整済み決定係数に表れている. しかし, テキストという人間が解釈を行うことのできる情報を用いて株価予測を行っているため, 各々の推定に対してどのような情報が実際に効果を持っていたのかを視覚的に確かめることができる. これは過去の変動に対して分析が可能である事, そして未来の株価の推定の際に用いることのできる可能性を示している.

3.2 予測期間別・高変動期の予測正答率

3.1 節では 1 ヶ月後の外挿予測について述べたが, 本手法によってどの程度先の未来まで推定できるのかを検証した. 期間別の外挿予測の騰落正答率を図 3 左の斜線棒グラフで示す. 期間別とは予測期間の長さであり, 2 節で述べたように 1 ヶ月後, 2 ヶ月後, 3 ヶ月後の 3 つの期間について予測精度を検証した.

予測期間が長くなるにつれ, その間に起こる事象の数が増えるためテキストと経済変動の相関は弱まる. 回帰式に用いられた主成分の見てみると 2 ヶ月後・3 ヶ月後の予測では, 1 ヶ月後の予測の 1.5 倍から2.0倍の数の主成分が用いられていた. これは説明力の弱い説明変数を多く用いることで経済変動に合致する回帰式を作成していたことになり, 過去の変動の説明はできても未来の変動に対する予測力が落ちる. 実際に, 期間が長くなるにつれて予測精度が悪くなる傾向が見られた. 逆に, 1 ヶ月後の予測に関して 2 ヶ月後, 3 ヶ月後の予測よりも少ない主成分による推定で好成績を収めたということは, それだけ経済の変動に直結した主成分を作成できていたということである.

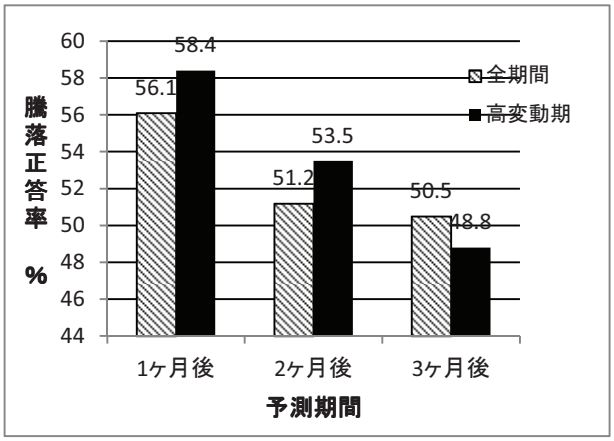


図 3: 予測期間別予測正答率・高変動期の予測正答率

予測期間が長くなるにつれて騰落正答率が下がる一方で, 変動の大きな時期における予測正答率では 1 ヶ月後・2 ヶ月後の予測で 2 ポイント以上の改善が見られた(図3 右の黒棒グラフ). このことから本手法は 2 ヶ月後の予測まで有効であるということができる. なお, 高変動期の予測とは下式によって定まる, 時刻 T から始まる訓練期間 36 ヶ月におけるリターンの標準偏差 σ の 0.1 倍よりも絶対値が大きな変動を推定した場合のみ予測を行ったものである.

$$\sigma^2 = \sum_{t=T}^{T+35} (r_{i,t} - \overline{r_{i,t}})^2$$

この閾値によって定めた高変動期は全予測期間のおよそ 3 分の 1 であった. 予測した回数のうち, 騰落が一致した割合を図 3 の黒い棒グラフで示している.

4. 提案手法のシステム化

実際の金融関係者の使用を目的として, 本研究で提案した手法のシステム化を行った. 3 節までの実験で投資家から要望の高い長期予測の有効性を確かめた. その上で投資家が円滑に投資を行うための解釈を行い, 過去の変動に対しても分析できるシステムを提案する. 過去の株価変動やテキストの情報を提示するのみでなく, 分析に基づいた情報を提供する.

4.1 提案システム概要

本システムは forecast タブ, tag cloud タブ, 共起関係出現パターン一覧で構成されている. 図 4 が forecast タブ, 図 5 が tag cloud タブ, 図 6 が共起関係出現パターン一覧である. 本節ではそれぞれの図に対して機能を説明する.

○forecast タブ
指標ごとの予測の騰落を株価天気予報の形で図 4 左の表にまとめている. 順に 1 ヶ月後, 2 ヶ月後, 3 ヶ月後の予測を表示したものであり, 一目で経済動向が把握できるようになっている. 図 4 右のグラフは指標ごとにタブで切り替えられるようになっており, 実際のリターンと回帰で推定されたリターンを訓練期間の 36 ヶ月+予測の 1 ヶ月分表示している. ユーザはグラフ上部にある a から u のタブを切り替えることで個別の指標を分析することも可能である. また, グラフ上の各点をクリックするとポップアップが表示され, 各時点において変動に最も寄与した主成分が表示される. 変動に最も寄与した主成分とは, その時点において回帰係数と主成分スコアの積の絶対値が最も大きい主成分のことである. pre/next のリンクをクリックすることで前月/翌月の分析結果に移動する.

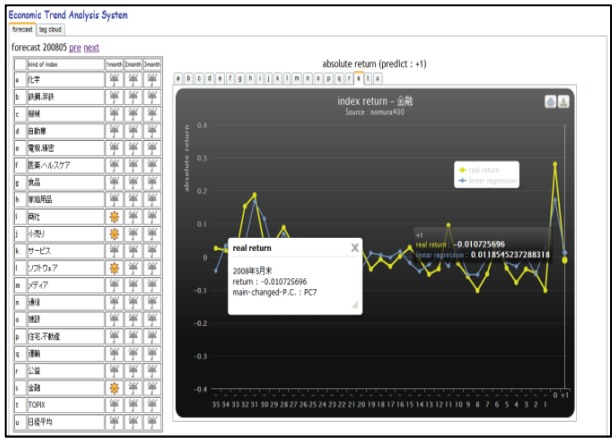


図4: forecast タブ

○tag cloud タブ
この時期の回帰式の説明変数となった各主成分に対して，因子負荷量の大きな単語を含む共起単語ペアの一例を主成分ごとに示している．図 4 のグラフのポップアップで表示される影響力の大きな主成分と，この表を見比べることで，どのような単語がどの指標のどの変動に効いていたのか調べることができる．また，各主成分の主要な共起単語を一覧でまとめているのは過去の分析にも対応するためであり，ここを一見すればどのような経済関連の記事があったのかを大まかに把握することができる．また，各主成分名（PC1，PC2，…，PC15）はリンク構造を持っており，クリックすることで図 6 の共起関係出現パターンを見ることができる．

pc	1	2	3	4	5	
PC1	純一郎 - 首相	買収 - 融資	信用 - 米国	数 - 株価指数先物	米州 - 総局	…
PC2	CSR - 推進	室長 - 広報	室長 - 取締役	大相撲 - 春場所	支社 - 東北	…
PC3	内閣 - 安倍	安倍 - 総裁	担保 - 新株予約権	発射 - ミサイル	ミサイル - 問題	…
PC4	郵政民営化 - 法案	郵政民営化 - 衆院	参院 - 郵政民営化	電車 - 脱線事故	JR - 尼崎	…
PC5	木村 - 建設	復党 - 問題	制度 - 減価償却	シリア - 北朝鮮	研究 - 研究所	…
PC6	連結経常利益 - 前期	地裁 - 東京	会談 - 首相	ロンドン - 同時	取締役 - 専務	…
PC7	甲子園 - 野球	シーア - イスラム教	英 - 首相	英 - 投資ファンド	イラン - ウラン濃縮	…
PC8	軍事 - 政権	指定 - 国家	テキサス - ウエスト	支援 - 国家	給油 - 海上自衛隊	…
PC9	下落 - 株価	違反 - 証券取引	純利益 - 連結決算	決算 - 純利益	提案 - 株主	…
PC10	第三者割当 - 増資	偽造 - する	財政改革 - 三位一体改革	経費 - 事務所	ガソリン - 税	…
PC11	実施 - 配当	実施 - 核実験	条約 - 憲法	年金 - 番号	年金 - 支給	…
PC12	遺伝子 - 組み換え	内閣 - 発足	内閣 - 改造	債権 - ローン	統一 - 地方	…
PC13	団体 - 業界	担当 - 太平洋	本部 - 捜査	強制 - 捜査	指名 - 首相	…
PC14	本体 - 子会社	する - 子会社	自動車 - メーカー	利回り - 高い	従業 - 取締役	…
PC15	一級 - 建築士	更新 - 最高値	関連 - 内需	ワシントン - 小竹	核 - 北朝鮮	…

図 5： tag cloud タブ

○共起関係出現パターン一覧
訓練期間・外挿期間における主要単語ペアの時系列的な出現パターン情報を示している．主要単語とは各主成分において因子負荷量が上位 15 位までのものとした．その主要単語を含む共起単語ペア全ての出現パターンを本システムでは表示し，上にあるほど因子負荷量が大きい単語を含むペアとなっている．この出現の有無と株価の騰落を合わせて分析することで，どのようなペアの出現の有無で騰落が決まっているのかを把握することができる．

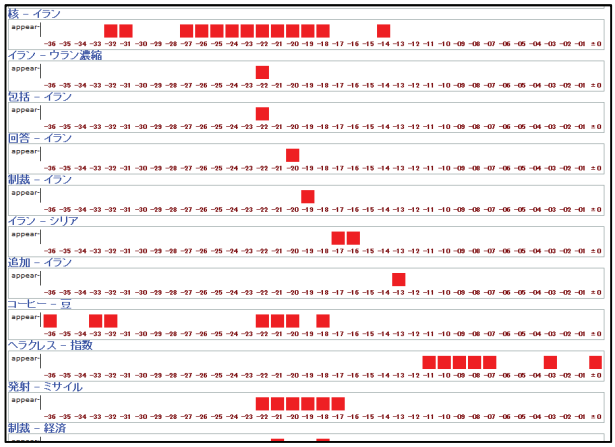


図 6： 共起関係出現パターン一覧

4.2 本システムを用いた経済動向分析の利用例

最初に forecast タブの晴/雨の表一覧で今後の経済トレンドを一目で把握できるようになっている．グラフについては過去の変動に対して内挿関数を作っているが，自由度調整済み決定係数は比較的小さい．その一方で少ない主成分の数で変動を説明できているため，図 5 や図 6 の主成分中の因子負荷量が大きい単語を含む共起ペアがどのように株価の騰落に影響しているのかを把握できる．

本システムでは過去の株価データとテキストデータの相関性を可視化することで，解釈を可能にしている．過去の相関から未来の予測の解釈に役立てることが本システムの目的である．

5. まとめ

本研究では CPR 法を応用して約 10 年間という長いスパンでの外挿予測を行った結果，1 ヶ月後の予測において TOPIX や日経平均といった市場平均株価では 60%以上の騰落正答率を収めた．また，業種別平均株価を含めた騰落正答率の平均でも 55%以上という高い精度を示した．これは実務で求められる水準の正答率である[7]．さらに，本研究で用いた手法を基に，投資家が解釈を行うことを前提とした経済動向分析システムも構築した．このシステムにより，今まで不明瞭であった経済変動とテキスト情報との関係性が明らかとなった．投資家は本システムを用いて解釈を行い，投資活動に役立てることが可能である．

本研究で用いた手法にはまだ改善の余地がある．例えば形態素解析の段階でも連語が分解されてしまっている点や，大容量のデータとして新聞記事のデータだけに限定してしまっている点である．実際の投資には様々な情報を用いて解釈を行うため，単一の情報ではなく複数指標に対応したシステムを提案していきたい．

参考文献

[AERA] AERA’2012.2.13，pp.62，朝日新聞出版．

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

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

[高穂 2002] 高穂洋，荒井隆行，大竹敢，田中衛 ： ニューラルネットワークによる時期の株価予測－株価予測におけるフィルタリングによる特徴量抽出－，電子情報通信学会技術研究報(NPL)，Vol.102，No.432， pp.13-16 (2002) .

[伊庭 2006] 伊庭斉志 ： 進化論的手法を用いた金融データの予測，日本信頼性学会誌，Vol.28，No.7，pp.471-480 (2006) .

[和泉 2011] 和泉潔，後藤卓，松井藤五郎： 経済テキスト情報を用いた長期的な市場動向推定，情報処理学会誌，Vol.52，No.12，pp.3309-3315 (2011) .

[大澤 2006] 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).

[日経] 日経シソーラス，日本経済新聞デジタルメディア http://vip-test2.nikkei.co.jp/help/contract/price/02/help_KIJI_thes.html

[野村] 野村證券金融工学研究センターホームページ ： http://qr.nomura.co.jp/jp/n40/index.html※※

※※ただし，NOMURA400 の提供は 2012 年 3 月末日をもって終了しています．