情報処理学会論文誌 Vol.53 No.1 117–125 (Jan. 2012)

ファクターモデルによるインターネット株式掲示板の投稿と株式リターンの分析

諏訪 博彦[1,a)] 梅原 英一[2] 太田 敏澄[1]

受付日 2011年3月8日, 採録日 2011年10月3日

概要：Yahoo!株式掲示板には，人々の意見・気持ちが大量に投稿されている．我々は，この意見・気持ちをコラボレーション技術を用いて取り込むことによって，投資判断を支援することが可能になると考える．本論文では，Yahoo!株式掲示板の投稿数および投稿内容が株式リターンと関係しているか明らかにする．そのために，東証 1 部における投稿数および強気指数によるポートフォリオを構築し Fama-French の 3 ファクターモデルが成立しているかを検証している．その結果，投稿数が最も多いポートフォリオおよび最も少ないポートフォリオ，強気指数が最も強気なポートフォリオおよび最も弱気なポートフォリオでは，超過リターンが存在することを確認している．そこで，投稿数の最上位ポートフォリオと最下位ポートフォリオのリターンの差および強気指数の最強気ポートフォリオと最弱気ポートフォリオのリターンの差をファクターとして追加している．その結果，強気指数ファクターを追加した場合には超過リターンは検出されなかった．これにより強気指数は株価リターンと関係している可能性があると考える．

キーワード：ファクターモデル，インターネット株式掲示板，投稿内容，株式リターン

Analysis of Internet BBS Messages and Stock Return Based on Factor Model

Hirohiko Suwa[1,a)] Eiichi Umehara[2] Toshizumi Ohta[1]

Received: March 8, 2011, Accepted: October 3, 2011

Abstract: We investigate whether the number and content of messages in BBS relates to the stocks return. We verify whether Fama and French 3 factor approve a portfolio by the number and contents of messages in TSE. We find that 3 factor model may not be approved because we find abnormal returns in portfolios with the highest number of bullish postings and the highest number of bear postings as well as the highest number of postings and the fewest number of postings. Next, we investigate that we add 3 factor model to two factors: the difference of the return of a portfolio with the highest number of messages and with the fewest, and the difference of the return with most bullish and with most bearish. As a result, we believe that the contents of messages in BBS may relate to a return.

Keywords: factor model, Internet BBS, content of message, stock return

1. はじめに

インターネット株式掲示板（以下，株式掲示板）は，他の投資家の生の声を直接知ることができるツールである．我々は，株式掲示板に投稿された意見・気持ちをコラボレーション技術を用いて取り込むことによって，投資判断を支援することが可能になると考える．本論文では，株式掲示板の投稿数および投稿内容が株式リターンと関係しているか明らかにする．

株式掲示板と株式市場との関連性に焦点をあてた研究としては，Antweiler ら [1] や丸山ら [2] などがある．これら

[1] 電気通信大学
University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[2] 野村総合研究所
Nomura Research Institute, Chiyoda, Tokyo 100–0005, Japan
[a)] h-suwa@is.uec.ac.jp

© 2012 Information Processing Society of Japan

の研究によると，株式掲示板の投稿数や投稿内容は，株式市場のボラティリティや出来高を説明すること，手数料などを考慮した場合に経済的な利得を得ることは難しいが，オーバーナイトの投稿内容が翌日のリターンと関係する可能性があることなどの示唆が得られている．

しかし，これらの研究では，投稿数の多い企業を対象に分析されており，市場全体を通した分析はなされていない．株式掲示板と株式市場の関係を理解するうえでは，市場全体を分析対象とすることが必要であると考えられる．そのためには投稿数の少ない企業も分析対象に加える必要がある．そこで本研究では，東証 1 部上場企業を対象に，Fama-French の 3 ファクターモデルに投稿数および強気指数をファクターとして加え，これが株式リターンの説明変数になるかを分析する．

本論文の構成は以下である．2 章で先行研究をレビューし，3 章で分析方法について述べる．4 章で分析データについて述べ，5 章で自然言語処理と機械学習を用いた投稿の分類方法と，分類データに基づく強気指数の算出方法について述べる．6 章でファクターモデルを構築し，7 章で分析結果を述べる．8 章で考察を行い，9 章は結論である．

2. 関連研究

本章では，株式掲示板と株式指標の関係に注目している研究と，分析方法として用いるファクターモデルに関する研究についてレビューを行い，本研究の目的を述べる．

2.1 株式掲示板と株式指標の関係

Wyscocki [3] は，1998 年 1 月から 8 月の間に Yahoo!掲示板に投稿された最も投稿数の多い 50 銘柄の投稿数と企業特性や株式市場の活動との関連を時系列で調べた．その結果，前日夜の投稿数が翌日リターンと統計的に有意であるが，手数料を考慮に入れた場合，経済的利益を得ることは難しいと述べている．Tumarkin ら [4] は，インターネット関連株 73 社の RagingBull.com の 1999 年 4 月 17 日から 2000 年 2 月 18 日までの 181,133 件のメッセージに関して，投稿数や投稿内容とリターンや出来高との因果関係をイベントスタディと多変量自己回帰分析で検証した．その結果，掲示板のメッセージはリターンを予測しないと述べている．Jones [5] は，S&P100 企業に対し Yahoo! Finance の投稿前後での株式リターンが変化するか調査した．その結果，投稿後の日次出来高で有意な増加が見られ，日次リターンが投稿後に有意に低下し，日次リターンの変動性が有意に増加することを発見した．この結果は，市場や産業レベルの出来事をコントロールした後でも成立したと述べている．

Antweiler ら [1] は，ダウ・ジョーンズ工業株指数と同インターネット指数組み入れ 45 社の 150 万件以上の Yahoo! と Raging Bull のメッセージの内容を，Naïve Bayesian 法による機械学習を用いて分析した．その結果，掲示板はリターンを予測しない，強気と弱気が拮抗するとそれが取引を誘発する結果，掲示板が出来高と当日のボラティリティを説明することを示した．また合意インデックスが出来高と負の関係にあることを報告している．Das ら [6] は，5 種類の自然言語処理アルゴリズムによる多数決投票で，分類で生じる誤差が改善することを示した．またモルガンスタンレーハイテク指数の技術セクター株 24 社を対象とした 2001 年 7–8 月の 145,110 メッセージに適用した分析結果は，Antweiler ら [1] と同様の結果を得ている．

丸山ら [2] は，東証 1 部上場企業について，日本の Yahoo! ファイナンス掲示板の投稿数上位 50 社（2005–2006 年）の企業を対象に掲示板指標と株式指標の関係を実証的に検証した．彼らは，掲示板より取得した 110 万件以上の投稿データを，自然言語処理と機械学習（Support Vector Regression）を用いて，「弱気」「中立」「強気」の 3 種類に分類し，強気投稿数と弱気投稿数により作成した強気指数を開発した．その結果，第 1 に掲示板の投稿数だけでは株式リターンを予測することは難しいが，弱気投稿数が多いと翌日の株式リターンがマイナスとなる傾向があること，第 2 に投稿数がボラティリティ，出来高と関連していること，第 3 に強気指数が株式リターンやボラティリティ，出来高と関連している可能性があることを報告している．

2.2 ファクターモデル

ファクターモデルとは，各銘柄・ポートフォリオのリターンをいくつかのファクターを用いて説明するモデルである．代表的なモデルとして Sharp の資本資産価格モデル（CAPM）[7] や Fama-French の 3 ファクターモデル [8] がある．

Fama-French[8] は，市場 β，時価総額，レバレッジ，自己資本の簿価対時価比率，益利回りと株式リターンのクロスセクション変数分析を行った．彼らは，CAPM が成立しているか否かを検証するために，これらの変数を高い順に 10 分位のポートフォリオを構築し検証した．CAPM が成立するためには，ポートフォリオのリターンと市場リターンで回帰した場合に，Y 切片（以下 α と呼ぶ）が 0 でなければならない．結果として，ファクターにより α が有意に 0 から乖離していることを発見している．

Fama ら [9], [10] は，CAPM に SMB ファクター（時価総額ファクター：大型株－小型株）と HML ファクター（純資産時価総額比率の高－低）の 2 種類のファクターを追加した 3 ファクターモデルを，式 (1) として提案している．

$$R_{i,t} - Rf_t = \beta_i(Rm_t - Rf_t) + s_i SMB_t + h_i HML_t + \varepsilon_i \quad (1)$$

$R_{i,t}$　時点 t のポートフォリオ i のリターン
Rf_t　時点 t のリスクフリーレート
β_i　ポートフォリオ i の市場感応度

© 2012 Information Processing Society of Japan

Rm_t 時点 t の市場リターン
s_i ポートフォリオ i の時価総額ファクターに対する感応度
SMB_t 時点 t の時価総額ファクターによるリターン差（時価総額ファクターリターン）
h_i ポートフォリオ i の純資産時価総額比率ファクターに対する感応度
HML_t 時点 t の純資産時価総額比率ファクターによるリターン差（純資産時価総額比率ファクターリターン）
ε_i ポートフォリオ i の残差リターン

SMB ファクターは，NYSE 上場企業の時価総額のメディアン（中央値）を測定し，時価総額がメディアン以上を大型株 (B)，以下を小型株 (S) と定義されている．HML ファクターは，NYSE 上場企業で，純資産が負のものを除き，純資産時価総額比率で降順に順位付けし，30%分位，70%分位を計算する．これにより高 (H)，中 (M)，低 (L) に分類し，6 つの加重平均ポートフォリオを構築している．これを毎年 6 月末で行う．この 6 つのポートフォリオの構築後 1 年のリターンを計算する．SMB ファクターは，各月の S/H, S/M, S/L の単純平均リターンから B/H, B/M, B/L の単純平均リターンを引いた値として求めた．HML ファクターは，各月の S/H, B/H の単純平均リターンから S/L, B/L の単純平均リターンを引いたものと定義される．この結果，CAPM よりも 3 ファクターモデルの方が，説明力が高いことを示している．久保田ら [11] は，Fama-French の 3 ファクターモデルが日本の株式市場でも成立することを検証している．

Antiweiler ら [12] は，株式掲示板の投稿数に基づくファクターモデルについて検証している．彼らは，米国の Yahoo! 掲示板を 1999 年から 2001 年の期間で，上場米国企業の 3,500 万以上の投稿を分析した．CAPM や Fama-French の 3 ファクターモデルに，投稿数ファクター（投稿数の多いポートフォリオリターン－投稿数の少ないポートフォリオリターン）を加え，投稿数により分割した 5 つのポートフォリオを構築した．その結果，仮のものであるという注釈のもと，投稿数ファクターが有意な働きをしていると主張している．しかし，彼らは強気指数ファクターについては検証していない．また，日本における投稿数ファクターを検証した研究は見当たらない．

2.3 研究目的

本研究の目的は 2 つである．第 1 の目的は，投稿活動が株式市場全体（東証 1 部）と関係があるかを明らかにすることである．丸山ら [2] は，投稿数が多い上位 50 社を分析対象に，投稿数や強気指数と株式指標との関係を相関係数を用いて論じた．しかし，東証 1 部で実際に取引されている銘柄は 1,500 銘柄以上あり，株式市場と株式掲示板の関係を示したとはいいがたい．特定銘柄のみの分析では，実際の投資行動への応用という点で課題を残している．また，相関係数のみでは，株式リターンを十分に説明できているのか，どの程度説明できているのか判断できない．本論文では，ファイナンス理論で一般的に使われているファクターモデルを利用することで，ファイナンス理論の分野で使用されているファクターと投稿活動に基づくファクターとの関係を明らかにする．それにより，投稿活動が株式市場全体（東証 1 部）にどの程度関係があるかを明らかにする．

第 2 の目的は，株式市場全体（東証 1 部）で，投稿数と投稿内容のファクターのどちらが有効であるかを明らかにすることである．Antiweiler ら [12] は，米国市場において投稿量をファクターとしたファクターモデルを構築している．しかし，彼らの分析では，投稿内容の分析は行われていない．そこで，我々は，投稿数ファクターと同様に，強気指数を用いた投稿内容に関するファクターも追加する．それにより，株式市場全体（東証 1 部）で，投稿数と投稿内容のファクターのどちらが有効であるかを分析する．

我々は，投稿数ファクターと強気指数ファクターを作成し，日本の東証 1 部上場銘柄を対象に，これらが株式リターンと有意な関係があるか検証を行う．

3. 分析方法

本章では，分析方法について述べる．丸山ら [2] は，投稿数と株式リターンに関連がなく，強気指数と株式リターンに関連があると主張している．この結果に基づくと，強気指数からは超過リターンが観測される可能性がある．この超過リターンは式 (2) に示す Fama-French の 3 ファクターモデルで表現できる．

$$R_{i,t} - Rf_t = \beta_i(Rm_t - Rf_t) + s_i SMB_t + h_i HML_t + \alpha_i + \varepsilon_i \quad (2)$$

投稿数とリターンが無関係ならば，投稿数の順位で構築したポートフォリオではすべての α が有意に 0 から乖離しない．一方，強気指数とリターンに関係があれば，強気指数から構築したポートフォリオの α では，有意に 0 から乖離することになる．そこで本研究では，投稿数および強気指数がリターンと関係しているかを調べるために，α が有意に 0 から乖離するかを調査する．

さらに，投稿数および強気指数がリターンを説明するファクターになるかを調査するために，Fama-French の 3 ファクターモデルに投稿数ファクター（以下 NMQ（Noise minus Quiet）と呼ぶ）または強気指数ファクター（以下 BMB（bullish minus bearish）と呼ぶ）を加えた 4 ファクターモデルを構築する．それぞれ，NMQ モデルとして式 (3)，BMB モデルとして式 (4) で表現する．

$$R_{i,t} - Rf_t = \beta_i(Rm_t - Rf_t) + s_i SMB_t + h_i HML_t + n_i NMQ_t + \alpha_i + \varepsilon_i \quad (3)$$

© 2012 Information Processing Society of Japan

$$R_{i,t} - Rf_t = \beta_i(Rm_t - Rf_t) + s_i SMB_t + h_i HML_t + b_i BMB_t + \alpha_i + \varepsilon_i \quad (4)$$

ファクターモデルの有効性を比較する方法として，回帰分析における α の有意検定に基づく方法や Gibbons ら [13] の平均分散効率性に対する尤度比検定，Hansen ら [14] による一般化モーメント法における Hansen-Jagannathan distance を用いた方法などが提案されている．久保田ら [11] は，日本の株式市場のデータを利用して Fama-French の 3 ファクターモデルの有効性を，平均分散効率性に対する尤度比検定と一般化モーメント法を用いて分析している．その結果，どちらの結果を用いても同様の結論であったことを報告している．そこで本研究では，式 (3)，式 (4) に対する回帰分析における α の有意検定と，Gibbons ら [13] の平均分散効率性に対する尤度比検定を用いて，Fama-French の 3 ファクターモデル，NMQ モデル，BMB モデルの 3 つを比較する．

4. 分析データ

本章では，分析データについて述べる．

4.1 分析対象

我々は，2009 年 3 月末で Yahoo!株式掲示板のメッセージを取得した．分析対象は，東証 1 部上場銘柄のうち 2003/01/01 から 2008/12/31 までの 6 年間（72 カ月）でメッセージを取得できた合計 1,501 社である．ただし，企業の新規上場・上場廃止などで月次で取得可能銘柄は変わっている．たとえば，2003 年 1 月は 1,485 社，2008 年 12 月は 1,398 社である．

対象期間における株式相場は東証株価指数（TOPIX）で 2003 年から 2008 年までの 6 年間の騰落率は，23.8%，10.2%，43.5%，1.9%，−12.2%，−41.8%であった．2003 年から 2005 年が上昇相場，2006 年は高値もみ合いのボックス相場，2007 年，2008 年は下降相場である．

4.2 Yahoo!株式掲示板の概況

取得総投稿は，3,891,158 件である．また，対象期間 2003–2008 年の対象銘柄 1,501 社の平均投稿数は，36.8 件/月である．2005–2008 年の月次平均投稿数では，1 番投稿が多い銘柄はソフトバンク社であり，月間平均投稿数は 9,364 投稿である．2 位の SBI ホールディングス社の 2,864 投稿の約 3 倍の投稿数がある．投稿は一部銘柄に片寄っており，非常にアクティブな掲示板とそうではない掲示板がある．

4.3 株式リターン・ファクターリターン

各銘柄のリターンは資本異動調整済リターン（配当込）である．リスクフリーレートは新発 10 年国債利回りである．ポートフォリオのリターンは，各ポートフォリオに分類されている銘柄のリターンの単純平均で計算した．市場リターンは当該月の分析対象銘柄の加重平均リターンである．加重ウエイトは，TOPIX（東証株価指数）のウエイトを用い全体が 100%となるように調整した．時価総額ファクターリターン（SMB_t）および純資産時価総額比率ファクターリターン（HML_t）は，銘柄の性質によるリターン格差を示す指標である．毎年 8 月末営業日を基準とし，その時点における時価総額および純資産時価総額比率に基づき分析対象銘柄を分類し，その後のリターンを用いて計算している．リターンには，東証 1 部，2 部全銘柄を対象とした普通株式の時価総額加重リターンを用いている．たとえば SMB の 2008 年 2 月のファクターリターンは，2007 年 8 月 31 日時点の時価総額に基づいて大型株・小型株のポートフォリオを作成し，各銘柄の 2008 年 1 月 30 日と 2008 年 2 月 29 日の時価総額加重リターンを用いて算出している．

5. 強気指数の算出

本章では，強気指数の算出方法について述べる．

5.1 掲示板投稿データの収集および強気・弱気分類

我々は，Yahoo!掲示板から投稿データを収集し，自然言語処理と機械学習の手法を用いて，投稿データを「弱気」「中立」「強気」の 3 種類に分類する．分類方法として，丸山ら [2] が提案した手法を用いる．彼らの手法は，1. 形態素解析・ノイズ除去処理，2. 特徴ベクトルの算出，3. SVR による分類の 3 ステップで構成されている．

5.1.1 形態素解析・ノイズ除去処理

インターネット掲示板から収集した投稿メッセージは，英語などと違い日本語の文章が単語間に分かれていない．このため，形態素解析プログラム（MeCab）により文章を形態素ごとに切り分けて単語を抽出する．さらに，形態素解析の結果得られた単語群から，特徴ベクトルとして適さない単語を除去するために，以下のノイズ除去処理を行う．

① 数字，英字，記号，および日本語以外の単語の除去
② 不要語（助詞，助動詞，接続詞，連体詞，副詞，数，代名詞，感動詞，固有名詞）の除去
③ 否定語の反映

以上のことより，特徴ベクトルを算出するための単語群を抽出する．

5.1.2 分類に有用な単語の抽出と特徴ベクトルの算出

次に，各投稿の特徴ベクトルを算出する．特徴ベクトルとして，丸山ら [2] の単語辞書を用いている．丸山ら [2] は，高村ら [15] の単語感情極性対応表と独自の用語を組み合わせ，総単語数は 6,989 単語の辞書を構築している．各投稿の特徴ベクトルは，6,989 次元のベクトルであり，その個々の値は単語の重要度である．重要度とは，投稿における単語の出現頻度の関数であり，TF・IDF 法により算出を行う．TF・IDF 法に基づく投稿 t 中の単語 d の重要度 $w(t,d)$ は，式 (5) により算出する．

© 2012 Information Processing Society of Japan

$$w(t,d) = \ln(tf_{t,d}+1)\cdot \ln\left(\frac{N}{df_d}\right) \quad (5)$$

N　すべての投稿数
$tf_{t,d}$　投稿 t に単語 d が出現する頻度
df_d　単語 d が出現する投稿の数

単語ごとに算出された重要度 w を用いて投稿の特徴ベクトル f_t を以下の定義より算出する．

$$f_t = (w(t,1), w(t,2), \ldots, w(t,l))$$

l：ベクトル用総単語数（6,989 ベクトル）

投稿ごとに算出された 6,989 次元ベクトル f_t を，SVR による分類のための特徴ベクトルとする．

5.1.3 SVR による分類

掲示板の投稿を Support Vector Regression（以下 SVR と記す）を用いて，「強気」「中立」「弱気」の 3 種類に分類する．SVR の分類プログロムは LibSVM を用いる．

学習データとして，Yahoo!掲示板に付加されている機能により「投稿者の気持ち」が判別できる投稿を用いる．「投稿者の気持ち」は，投稿者が投稿ごとに「強く買いたい」「買いたい」「様子見」「売りたい」「強く売りたい」「公表しない」の 6 個の選択の中から任意のものを選び，それを公開する Yahoo!掲示板の機能である．SVR の学習データは「公表しない」を除いた投稿メッセージとし，入力値は投稿の特徴ベクトルと投稿者が実際に入力した投稿者の気持ちであり，「強く買いたい」を 1，「買いたい」を 0.5，「様子見」を 0，「売りたい」を −0.5，「強く売りたい」を −1 として学習を行う．つまり投稿の特徴ベクトルが強気を示す場合は正，弱気を示す場合は負，様子見の場合は 0 への回帰が期待される学習である．なお，この入力値を「入力極性値」と呼ぶ．

学習は，分析対象期間の全学習データ（824,447 件）を用いて銘柄ごとに行っている．なお，ソフトバンク社は投稿数が格段に多く，プログラムのメモリ容量と計算時間上の問題が起こったため，等間隔抽出により学習サンプル数を半分にして学習を行っている．SVR による出力値として，学習に基づいて予測された投稿者の気持ちが出力値として算出される．この出力値を「予測極性値」と呼ぶ．**表 1** に学習データの気持ち別投稿数および予測極性値の平均・分散を示す．「強く買いたい」「買いたい」の投稿者の気持ちが公開されている総投稿数は，417,301 件で全学習データの 50.6%である．予測極性値の単純加算平均は，0.744 である．「様子見」の投稿者の気持ちが公開されている総投稿数は，235,671 件で全学習データの 28.6%である．予測極性値の単純加算平均は，0.434 である．「強く売りたい」「売りたい」の投稿者の気持ちが公開されている総投稿数は，171,475 件で全学習データの 20.8%である．予測極性値の単純加算平均は，−0.297 である．

丸山ら [2] は，SVR の分類結果が強気にシフトする傾向があることを指摘している．今回の分析結果においてもその傾向が確認された．そこで，補正のために「強く買いたい」「買いたい」における予測極性値の単純加算平均（0.744）より大きいものを「強気」投稿，「強く売りたい」「売りたい」における予測極性値の単純加算平均（−0.297）より小さいものを「弱気」投稿，それ以外を「様子見」投稿として分類した．**表 2** に，SVR の学習データ分類精度を示す．「強く売りたい」「売りたい」の気持ちが公開されている投稿が弱気に分類される精度は 78.1%，「強く買いたい」「買いたい」の投稿が強気に分類される精度は 77.2%であった．

この学習データをもとに，気持ちを公表していない投稿を含めた全 3,891,158 件の投稿を銘柄ごとに分類し，集計した結果を**表 3** に示す．強気に分類された投稿が 1,190,467 件（30.6%），弱気に分類された投稿が 425,282 件（10.9%）であり，強気の投稿が多くなっている．本研究では，書かれた投稿内容（テキスト）に基づいて投稿を分類するために，気持ちが公表されている投稿についても予測極性値を用いて再分類し，その結果に基づいて強気指数の算出を行う．

表 1 各気持ちにおける投稿数と予測極性値の平均・分散
Table 1 Number of posts and average/variance of polarity score.

気持ち	投稿数	平均	分散
強く買いたい・買いたい	417,301	0.744	0.317
様子見	235,671	0.434	0.543
売りたい・強く売りたい	171,475	−0.297	0.612

表 2 SVR の学習データ分類精度
Table 2 Classification accuracy of SVR training data.

		分類結果			合計	精度
		弱気	中立	強気		
投稿者の気持ち	強く売りたい	9.6%	6.7%	0.9%	17.1%	54.3%
	売りたい	1.7%	1.7%	0.2%	3.7%	
	中立	2.6%	18.2%	7.8%	28.6%	63.8%
	買いたい	0.1%	12.1%	1.6%	13.9%	59.3%
	強く買いたい	0.5%	7.9%	28.4%	36.7%	
合計		14.5%	46.7%	38.8%	100.0%	
精度		78.1%	39.1%	77.2%	59.5%	

表 3 強気・弱気分類結果
Table 3 Result of classification.

投稿者の気持ち	強く売り、売り	様子見	買い、強く買い	合計
学習データ	171,475 (20.8%)	235,671 (28.6%)	417,301 (50.6%)	824,447
分類データ	弱気	中立	強気	合計
学習データの分類結果	119,305 (14.5%)	384,918 (46.7%)	320,224 (38.8%)	824,447
全データの分析結果	425,282 (10.9%)	2,275,409 (58.5%)	1,190,467 (30.6%)	3,891,158

5.2 強気指数

強気指数とは，月ごとにおける強気/弱気意見の多少を示す掲示板指標であり，式 (6) で定義する．

$$\text{強気指数}\ (t) = \ln\left(\frac{1+\text{強気投稿数}\ (t)}{1+\text{弱気投稿数}\ (t)}\right) \quad (6)$$

この指標は，強気投稿数が弱気投稿数に比べ多くなると正の値となり，同数の場合は 0，強気投稿数が弱気投稿数

© 2012 Information Processing Society of Japan

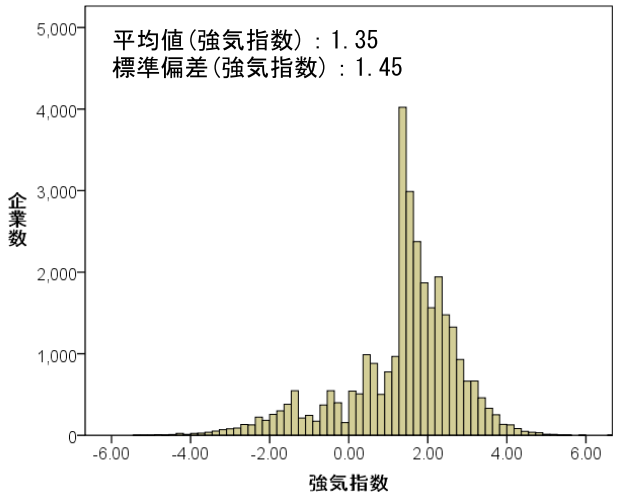

図 1 月次の強気指数の分布（2003–2008 年）
Fig. 1 Distribution of bullishness.

表 4 投稿数ポートフォリオ
Table 4 Portfolio based on number of posts.

ポートフォリオ	定義	2003年1月		2008年12月	
		平均投稿数	銘柄数	平均投稿数	銘柄数
PF-A	投稿数上位1/5の企業	54.7	155	294.1	217
PF-B	投稿数上位1/5〜2/5の企業	8.1	155	26.2	217
PF-C	投稿数上位2/5〜3/5の企業	4.0	155	8.6	217
PF-D	投稿数上位3/5〜4/5の企業	2.0	155	3.5	217
PF-E	投稿数上位4/5〜5/5の企業	1.0	155	1.3	217
PF-X	投稿なし	0.0	710	0.0	313
全社	取得可能全企業	7.2	1485	51.6	1398

より少なくなると負の値となる．ただし，1 カ月単位で強気投稿数と弱気投稿数の合計が 3 未満のデータは，強気か弱気かの判断ができないと考え除外している．**図 1** に，全企業・全月の強気指数の分布を示す．その結果，総データ数 105,779 件のうち強気指数が得られたデータは 32,989 件であった．

6. ファクターモデルの構築

ファクターモデルを構築するために，最初に，投稿数および強気指数に基づきポートフォリオを構築する．ポートフォリオ構築のために，投稿数および強気指数を月次単位（1 日から月末日）で計算する．これをもとに投稿数ファクター（NMQ）と強気指数ファクター（BMB）を構築する．

6.1 投稿数ポートフォリオの作成

投稿数ポートフォリオの作成方法は以下である．まず，投稿のまったくない企業をポートフォリオ X（以後 PF-X と記す．他のポートフォリオも同様に，PF-*と記す）として分類した．次に各月の投稿数に従い，残りの対象銘柄を 5 分割する．PF-A はその月で最も投稿数の多い企業群であり，PF-E は最も投稿数の少ない企業群である．2003 年 1 月および 2008 年 12 月の結果を**表 4** に示す．

6.2 強気指数ポートフォリオの作成

強気指標を用いて強気指数ポートフォリオを構築する．

表 5 強気指数ポートフォリオ
Table 5 Portfolio based on bullishness.

ポートフォリオ	定義	2003年1月		2008年12月	
		平均強気指数	銘柄数	平均強気指数	銘柄数
PF-1	強気指数上位1/5の企業	3.16	30	2.70	110
PF-2	強気指数上位1/5〜2/5の企業	2.25	31	1.78	110
PF-3	強気指数上位2/5〜3/5の企業	1.76	31	1.23	110
PF-4	強気指数上位3/5〜4/5の企業	1.45	31	0.36	110
PF-5	強気指数上位4/5〜5/5の企業	−0.44	31	−1.48	111
PF-XX	強気投稿数＋弱気投稿数＜3	0.00	1331	0.00	847
全社	取得可能全企業	0.16	1485	0.36	1398

表 6 ファクター間の相関係数
Table 6 Correlation coefficient between factors.

	MARKET	SMB	HML	NMQ	BMB
MARKET	−				
SMB	−0.194	−			
HML	−0.105	−0.081	−		
NMQ	0.754***	0.045	−0.043	−	
BMB	0.580***	0.140	−0.243*	0.540***	−

* p<0.05, *** p<0.001

1 カ月単位で強気・弱気投稿数の和が 3 未満のものは強気指数を判断できないと考え PF-XX として分類する．次に各月の強気指数に従い，残りの対象銘柄を 5 分割する．PF-1 はその月で最も強気指数の高い企業群であり，PF-5 は最も強気指数の低い企業群である．2003 年 1 月および 2008 年 12 月の結果を**表 5** に示す．

6.3 投稿数ファクター・強気指数ファクター

投稿数ファクター（NMQ）と強気指数ファクター（BMB）を，それぞれ式 (7) および式 (8) に定義する．各ポートフォリオのリターンは，SMB や HML と同様に各銘柄の前月末営業日と当月末営業日の時価総額加重リターンを用いて算出している．

$$\text{NMQ} = \text{PF-A のリターン} - \text{PF-E のリターン} \quad (7)$$

$$\text{BMB} = \text{PF-1 のリターン} - \text{PF-5 のリターン} \quad (8)$$

7. 分析結果

本章では，分析結果について述べる．

7.1 3 ファクターとの関係

3 ファクターモデルと投稿数および強気指数の関係を調べる．3 ファクター（市場：β，時価総額：SMB，純資産時価総額比率：HML）と投稿数ファクター（NMQ）および強気指数ファクター（BMB）の相関関係を**表 6** に示す．

表 6 より，市場ファクター（β）と投稿数ファクター（NMQ）の相関は 0.754，また強気指数ファクター（BMB）の相関は 0.580 で統計的に 0.1%有意で相関が高いことが分かる．しかし，時価総額ファクター（SMB）と投稿数ファクター（NMQ）の相関は 0.045，また強気指数ファクター（BMB）の相関は 0.140 で，有意ではない．そして，純資産時価総額比率（HML）ファクターと投稿数ファクター（NMQ）の相関は −0.043 で有意ではない．また強気指数

© 2012 Information Processing Society of Japan

ファクター（BMB）の相関は −0.243 で 5%有意となった．この結果より，株式掲示板の投稿は，時価総額や純資産時価総額比率よりも市場ファクターを強く反映していることが分かる．

7.2 投稿数および強気指数

表 7，**表 8** に，投稿数ポートフォリオおよび強気指数ポートフォリオの分析期間（2003–2008 年）における 3 ファクターモデルの係数と α の値を示す．投稿数ポートフォリオに関しては，PF-A，PF-B の α は 0 から正に有意に乖離している．PF-E は負に有意に乖離している．この結果，3 ファクター以外のファクターが存在している可能性がある．強気指数ポートフォリオに関しても，PF-1，PF-2 の α が 0 から正に有意に乖離し，PF-5 の α が 0 から負に有意に乖離する．強気指数の高い企業ほど α が高い結果になっている．強気指数が株式リターンを説明するファクターになる可能性が示唆される．次節で，4 ファクターモデル（NMQ モデル，BMB モデル）による分析結果について述べる．

表 7 3 ファクターモデルによる投稿数ポートフォリオの分析
Table 7 Analysis of portfolio based on number of posts, using 3 Factor model.

ポートフォリオ	α	β	SMB	HML	決定係数
PF-A	0.574*	1.356	0.854	0.382	0.937
PF-B	0.329**	1.100	0.656	0.325	0.978
PF-C	−0.061	0.977	0.688	0.362	0.973
PF-D	−0.166	0.859	0.623	0.271	0.955
PF-E	−0.482**	0.761	0.560	0.274	0.929
PF-X	−0.098	0.923	0.838	0.627	0.937

* p<0.05, ** p<0.01

表 8 3 ファクターモデルによる強気指数ポートフォリオの分析
Table 8 Analysis of portfolio based on bullishness, using 3 Factor model.

ポートフォリオ	α	β	SMB	HML	決定係数
PF-1	2.167***	1.388	0.918	0.383	0.931
PF-2	0.978***	1.181	1.024	0.283	0.936
PF-3	0.053	1.176	0.800	0.416	0.935
PF-4	−0.2244	1.152	0.656	0.389	0.936
PF-5	−0.804***	1.079	0.678	0.585	0.926
PF-XX	−0.180*	0.910	0.726	0.488	0.984

* p<0.05, *** p<0.001

7.3 モデルの有効性比較

表 9 は，式 (3) の NMQ モデルおよび式 (4) の BMB モデルにおける回帰分析の結果である．NMQ モデルにおいては，PF-1，PF-2 の α が 0 から正に有意に乖離し，PF-4，PF-5 の α が 0 から負に有意に乖離している．一方，BMB モデルにおいては，PF-B の α のみが 0 から正に有意に乖離している．

表 10 は，Gibbons ら [13] のファクターモデル間の有効性検定の結果である．帰無仮説は式 (9)，F 統計量は式 (10) である．彼らは，この統計量が，リターンと説明変数が正規分布で真の Y 切片が 0 であるという仮定のもとで，N と

表 10 モデル比較（平均分散効率性検定）
Table 10 Comparison of models.

	F値	自由度
3ファクターモデル	16.64	12/57
NMQモデル	14.41	11/57
BMBモデル	6.16	11/57

* 3 ファクターモデル，NMQ モデル，BMB モデルそれぞれに対して，Gibbons ら [13] 検定量の F 値である．

表 9 NMQ モデルおよび BMB モデルによる投稿数・強気指数ポートフォリオの分析
Table 9 Analysis of portfolio based on number of posts/bullishness, using NMQ and BMB model.

	ポートフォリオ		α	β	SMB	HML	NMQ	BMB	決定係数
NMQ モデル	投稿数ポートフォリオ	PF-A	−0.139	0.954	0.655	0.309	0.676		0.989
		PF-B	0.232	1.046	0.629	0.315	0.092		0.980
		PF-C	0.035	1.031	0.715	0.372	−0.091		0.975
		PF-D	0.031	0.970	0.678	0.291	−0.187		0.965
		PF-E	−0.139	0.954	0.655	0.309	−0.324		0.966
		PF-X	−0.159	0.888	0.821	0.621	0.058		0.937
	強気指数ポートフォリオ	PF-1	1.633***	1.088	0.770	0.329	0.505		0.958
		PF-2	0.622**	0.981	0.925	0.247	0.337		0.953
		PF-3	−0.228	1.018	0.721	0.387	0.267		0.946
		PF-4	−0.594**	0.943	0.553	0.351	0.350		0.956
		PF-5	−1.224***	0.843	0.561	0.542	0.398		0.954
		PF-XX	−0.141	0.931	0.737	0.492	−0.037		0.985
BMB モデル	投稿数ポートフォリオ	PF-A	0.390	1.337	0.839	0.394		0.062	0.937
		PF-B	0.403*	1.108	0.662	0.320		−0.025	0.978
		PF-C	0.204	1.005	0.710	0.344		−0.089	0.975
		PF-D	0.104	0.888	0.645	0.253		−0.091	0.956
		PF-E	−0.195	0.791	0.583	0.255		−0.097	0.931
		PF-X	−0.333	0.898	0.819	0.643		0.079	0.938
	強気指数ポートフォリオ	PF-1	0.194	1.183	0.759	0.517		0.664	0.963
		PF-2	0.68	1.150	1.000	0.304		0.100	0.937
		PF-3	0.279	1.200	0.818	0.401		−0.076	0.936
		PF-4	−0.08	1.167	0.668	0.379		−0.049	0.936
		PF-5	0.194	1.183	0.759	0.517		−0.336	0.939
		PF-XX	−0.200	0.907	0.725	0.490		0.007	0.984

* p<0.5, ** p<0.01, *** p<0.001

© 2012 Information Processing Society of Japan

$T-N-L$ の自由度を持つ F 分布に従うことを示している．

$$H_0 : \alpha_i = 0 \quad (\text{for all } i) \tag{9}$$

$$F = (T/N)\{(T-N-L)/(T-L-1)\} \{(A'\Sigma^{-1}A)/(1+R'\Omega^{-1}R)\} \tag{10}$$

ここで T は月数（72 カ月），L は回帰における説明変数の数であり，3 ファクターモデルでは 3，NMQ モデルおよび BMB モデルでは 4 である．N はポートフォリオ数であり，3 ファクターモデルでは 12 ポートフォリオ（PF-A〜E，PF-X，PF-1〜5，PF-XX），NMQ モデルでは 11 ポートフォリオ（PF-A〜D，PF-X，PF-1〜5，PF-XX），BMB モデルでは 11 ポートフォリオ（PF-A〜E，PF-X，PF-1〜4，PF-XX）である．Σ は N ポートフォリオの回帰残差の共分散行列である．なお，NMQ モデルと BMB モデルにおいて PF-E および PF-5 を分析から除外した理由は，Σ が正則行列である必要があるからである．A は N ポートフォリオそれぞれの時系列回帰の Y 切片の行ベクトルである．R は N ポートフォリオで観測したファクターリターンの平均を要素とするベクトル，Ω は N ポートフォリオ，72 カ月で観測したファクターリターンの共分散行列である．

3 モデルとも検定結果は 1%有意であり平均分散効率性は棄却される．つまり，株式リターンを説明するためには，これらのファクター以外の要因が存在していることを示している．しかし NMQ モデルは 3 ファクターモデルと比較して，ファクターを 1 つ追加しても説明力が向上しないことが分かった．一方，BMB モデルは 3 ファクターモデルよりも説明力が向上している可能性があることが分かった．この結果，投稿数はリターンを説明している可能性は低いが，強気指数は，部分的にでもリターンを説明している可能性があると考えられる．

8. 考察

本論文の目的は 2 つであった．第 1 の目的は，投稿活動が株式市場全体（東証 1 部）とどの程度関係があるかを明らかにすることである．第 2 の目的は，株式市場全体（東証 1 部）で，投稿数と投稿内容のファクターのどちらが有効であるかを明らかにすることである．

この 2 つの目的を達成するために，我々は分析対象を既存研究では対象としていない東証 1 部上場銘柄に拡張し，ファクターモデルを用いることで影響の程度を確認した．ファクターモデルを用いることで，丸山ら [2] が相関係数に基づいて議論しているため関係の有無しか議論できていないのに対し，投稿ファクターを加えたモデルが株式リターンを十分に説明できていること，既存のファクターモデルと比較して株式リターンを説明できていることが確認できた．また，投稿数が少ない銘柄に分析対象を拡張した場合においても，投稿活動が株式リターンを説明できることが確認できた．これらの結果より，株式掲示板の投稿を分析することで，日本における主要な投資市場である東証 1 部市場全体を株式掲示板の投稿活動に基づいて議論可能としており，実際の投資行動への応用可能性を示せたと考える．

また，ファクターモデルによるモデル比較により，投稿数のファクターよりも，投稿内のファクターが株式リターンを説明するために有効であることを確認できた．投稿数は銘柄への関心度を表し，投稿内容は投稿者の気持ちを表していると考えられる．株式リターンを説明するためには，どの銘柄に関心を持っているのかだけでなく，どんな関心を持っているのか，すなわち気持ちを分析する必要があることが分かった．本研究では，投稿内容として強気・弱気の 1 変数にのみ着目しているが，和泉ら [16] のように，内容分析を行うことで複数の軸を抽出することも可能であり，新たなファクターとして期待できる．

本研究では，投稿者自身が気持ちを表現している投稿についても，SVR の結果を利用して再分類している．また，SVR の分類結果が強気にシフトする傾向があることから，その補正のために分類の閾値として学習データの各分類における予測極性値の平均を用いている．気持ちの付いている投稿については，予測極性値を用いない方法や閾値をより大きく/小さくする方法も考えられる．また，特徴ベクトルを導出するための辞書に，日経シソーラスなどの既存辞書を使用することも考えられる．これらの方法や閾値などのチューニングについては今後の課題としたい．

9. 結論

本研究では，投稿数および強気指数が株式リターンと関係しているかを東証 1 部上場銘柄に対して調査した．その結果，株式掲示板の強気指数が部分的にでも株式リターンの説明変数になる可能性が示唆された．

この結果から，日本市場において投稿数は株式リターンを説明する可能性が低いが，強気指数は株式リターンを説明する可能性があることが分かった．つまり，人々がどの株に注目しているか（投稿数）よりも，人々がその株に対して何を言っているか（強気指数）が，株式リターンと関係している可能性がある．

本研究のモデルでは，各ファクターとリターンの時点は同時点として分析を行った．つまり我々の分析では，投稿活動が将来のリターンを予測しているかは分析の対象外である．今後の課題として，投稿がリターンを予測しているのか，株式相場が先行し投稿活動は相場動向を見た後で行われているのかを，月次ではなくもっと短期間のデータを用いて調査する必要があると考えている．

謝辞 日本市場における Fama-French の 3 ファクターモデルのデータを株式会社金融データソリューションズに提供いただいた．ここに記して深く感謝いたします．また，本研究の一部は，科研費（23500308）の助成を受けたものである．

© 2012 Information Processing Society of Japan

参考文献

[1] 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).

[2] 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, No.11-12, pp.110–127 (2008).

[3] Wyscocki, P.D.: Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards, Working paper, University of Michigan (1999), available from ⟨http://papers.ssrn.com/sol3/papers.cfm?abstract_id=160170⟩.

[4] Tumarkin, R. and Whitelaw, R.F.: News or Noise? Internet Postings and Stock Prices, *Financial Analysts Journal*, Vol.57, pp.41–51 (2001).

[5] Jones, A.L.: Have internet message boards changed market behavior?, *The Jounarl of Policy, Regulation and Strategy for Telecommunications*, Vol.8, No.5, pp.67–76 (2006).

[6] Das, S.R. and Chen, M.Y.: Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, *Management Science*, Vol.53, No.9, pp.1375–1388 (2007).

[7] Sharp, W.F.: Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance*, Vol.19, No.3, pp.425–442 (1964).

[8] Fama, E.F. and French, K.R.: The Cross-Section of Expected Stock Returns, *The Journal of Finance*, Vol.32, pp.427–465 (1992).

[9] Fama, E.F. and French, K.R.: Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, Vol.33, pp.3–56 (1993).

[10] Fama, E.F. and French, K.R.: Multifactor Explanations of Asset Pricing Anomalies, *Journal of Finance*, Vol.51, No.1, pp.55–84 (1996).

[11] 久保田敬一，竹原 均：Fama-French ファクターモデルの有効性の再検証，現代ファイナンス，No.22, pp.3–23 (2007).

[12] Antweiler, W. and Frank, M.Z.: Internet Stock Message Boards and Stock Returns, University of British Columbia Working Paper, 2002, available from ⟨http://strategy.sauder.ubc.ca/antweiler/public/returns.pdf⟩ (accessed 2009-05-25).

[13] Gibbons, M., Ross, S. and Shanken, J.: A Test of the Efficiency of a Given Portforio, *Econometroca*, Vol.57, No.5, pp.1121–1152 (1989).

[14] Hansen, L.P. and Jagannathan, R.: Assessing Specification Errors in Stochastic Discount Factor Model, *Journal of Finance*, Vol.52, No.2, pp.557–590 (1997).

[15] 高村大也，乾 孝司，奥村 学：スピンモデルによる単語の感情極性抽出，情報処理学会論文誌，Vol.47, No.2, pp.627–637 (2006).

[16] 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, pp.383–387 (2010).

諏訪 博彦 （正会員）

1998 年群馬大学社会情報学部卒業．2006 年電気通信大学大学院情報システム学研究科博士後期課程修了．博士（学術）．現在，電気通信大学大学院情報システム学研究科社会知能情報学専攻社会情報システム学講座助教．ソーシャルメディアに関する研究に従事．

梅原 英一

1981 年東京工業大学大学院システム科学専攻修士課程修了，野村総合研究所で DSS，トレーディングシステム，資産運用システムの開発に従事．博士（工学）．著作には『情報システムの統治組織の有効性比較』（経営情報学会誌，2008 年 9 月），『リスク情報開示ゲームの提案』（日本社会情報学会誌，2005 年 9 月）等．

太田 敏澄

1947 年生．東京工業大学経営工学科 1970 年卒業，同大学院理工学研究科 1972 年工学修士，1977 年工学博士．電気通信大学大学院情報システム学研究科教授．社会情報システム学，組織知能工学．『社会の中の企業』（共著），『都市と環境の公共政策』（共著），『環境としての情報空間』（共著），『社会情報システム学・序説』（共著），『Creative and Innovative Approaches to the Science of Management』（共著）．日本社会情報学会（JASI），日本ソフトウェア科学会，経営情報学会，日本 OR 学会等．

© 2012 Information Processing Society of Japan