株式掲示板における投稿者の評価手法

海野 洋平† 柳井 佳孝†† 山名 早人‡¶

†早稲田大学理工学部 〒169-85555 東京都新宿区大久保 3-4-1
††早稲田大学大学院理工学研究科 〒169-85555 東京都新宿区大久保 3-4-1
‡早稲田大学院理工学術院 〒169-85555 東京都新宿区大久保 3-4-1
¶国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †{unno, believe, yamana}@yama.info.waseda.ac.jp

あらまし 近年，投資家のための情報支援サイトや，情報交換の場として，掲示板やブログ，コミュニティサイトなどが多く提供されている．こうした Web 上のファイナンス系テキストデータから有用な情報を抽出するためには，内容よりも書き手の信頼度が重要となる．本研究では，信頼度の高い書き手を抽出することを目的として，Yahoo!掲示板に寄せられた，銘柄に関する意見が述べられた投稿テキスト，株価時系列データを用いて，投稿者の信頼度評価を行う手法を提案する．投稿者の信頼度は，投稿者の投稿テキストから投資戦略を読み取り，投稿翌日の株価をもとに評価する．検証実験では，特定銘柄における約 1 年間の投稿データから投稿者の投資戦略を抽出して，日次の収益率を用いて投稿者の評価を行った．提案手法による評価の結果，特定の銘柄において，ベースを上回る信頼度の高い(勝ち組)投稿者を 119 名抽出できた．さらに，裏切り戦略を行うという前提で，有用な負け組投稿者を 210 名抽出することができた．また，全ての投稿者の投資戦略を評価した結果，評価値平均が評価期間の平均収益率を下回っており，裏切り戦略が株価予測の観点から，有効であることが示唆された．

キーワード Web とインターネット，知識発見，データマイニング

An Evaluation Technique of Contributors in a Stock Bulletin Board

Yohei UNNO† Yoshitaka YANAI†† and Hayato YAMANA†‡

†Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555 Japan
††Graduate School of Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555 Japan
‡National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
E-mail: †{unno, believe, yamana}@yama.info.waseda.ac.jp

Abstract In this paper we propose an evaluation technique of contributors in a stock bulletin board to extract winners who predict correct stock strategy. In such stock bulletin boards, many contributors exist and contribute many investment strategies. However, some of them will be in-correct. In this paper, using recent 1-year contributed texts in Yahoo! JAPAN bulletin board, we propose a detection scheme to find out trustful contributors who predict investment strategy in correct. In the evaluation, we use contributor's investment strategy, which is able to be detected by their written texts, and daily rate of stock return. As a result, 119 useful contributors who exceed the base line, called winners, are extracted. Furthermore, 210 useful contributors, who exceed the base line only when we take betrayal investment strategy against them, are also extracted. As the result of evaluating all the investment strategies of all contributors, we have found that betrayal strategy against many strategies written in bulletin boards, is effective in the viewpoint of excess income, because the average earning rate has been less than the average of stock return in the evaluation period.

Keyword Web and Internet, Knowledge Discovery, Data

1. はじめに

研究の背景として，近年，相次ぐネット証券参入や，手数料の自由化による低コスト化，証拠金制度の緩和，株式分割，投資単位の引き下げなどによって，個人の投資機会は大幅に増加している．株式分布状況[1]によると，2005 年度の個人株主数は前年度比 268 万人増の 3807 万人となっている．

個人の投資機会の増加に伴って，Web 上に，投資家のための情報支援サイトや，掲示板やブログ，コミュニティサイトなどの情報交換の場が多く提供されるよ

うになっている．その中でも，Yahoo! Japan の提供する「Yahoo!掲示板＞株式カテゴリ」[2]には，1 日に 1000 件近い投稿が寄せられているトピックが確認されており，最も活発に情報交換が行われている場所と言える．銘柄別に設置されたトピックには，Yahoo_ID を持つ投稿者から銘柄に関する情報や意見が投稿されているが，銘柄とは関係のない情報(ノイズ)も多く確認された．そのような投稿から，有用な情報を抽出することは知識発見の観点からも重要である．

ファイナンス系テキストデータからの情報抽出では，従来研究として，新聞記事やアナリストレポートなどの形式的な文章を扱ったものが確認された．[3]では，経済新聞の記事テキストを対象として，分類器によるインパクト記事の自動抽出を行っている．ファイナンス系テキストからの有用情報抽出においては，有用性は株価もしくは経済動向の予測精度に大きく依存しているため，テキストの内容以前に投稿者の信頼度の高さが求められる．本研究では，平均を上回る投資戦略を持つ投稿者は優秀(勝ち組)であり，その投稿は投資家にとって信頼度の高い情報であるという仮定のもと，掲示板投稿者から勝ち組投稿者の選出を行った．

2 節では，知識発見・情報抽出における，従来研究の問題点と，本稿で解決する点について，3 節で専門用語について説明，4 節では具体的な提案手法の流れと，具体的な評価方法を述べる．5 節では，4 節で述べた評価手法を用いた検証実験を行い，6 節では，まとめと今後の課題について述べる．

2. 関連研究

Web 掲示板からの情報抽出では，主に意見文・評価文の抽出を行う研究がされている[4 -11]．しかし，従来手法をファイナンス系テキストデータにそのまま適用することは有用性の観点から難しい．つまり，ファイナンス系テキストの場合，従来手法で意見文を抽出しても，テキスト内容が投稿者の利害・ポジションに大きく左右されるため，投資家にとって利用価値が低い．本稿では，投稿者単位で評価・抽出を行うことで，投稿者の利害を考慮する必要のない情報を抽出できる．

ファイナンス系テキストデータに対する知識発見・情報抽出の関連研究では，超過収益を得ることが１つの大きなテーマになっている．四季報中の企業アウトルック説明文を用いて，キーワードによるグループ分け(肯定/否定/中立)された銘柄と株価収益の，関連性を検証している研究[12]や，アナリストレポートを対象とした，特定のキーワードに対する株価収益や，数値データ(市場の平均予想値)との関連性を検証している研究[13]がある．

ファイナンス系テキストデータからの有用情報抽出では，経済新聞を用いた分類器によるインパクト記事の自動抽出を行っている研究[3] がある．[3]では，経済新聞記事を対象として，掲載されている個々の企業の記事の内容を解析し，インパクトのある記事を株価変動率により選択し，特徴量を取り出して分類器を生成している．しかし，新聞記事のように形式的でノイズの少ない文章から抽出する手法はそのまま掲示板のような煩いテキストには応用できない．つまり，ノイズが多く特徴量が少ない掲示板の投稿テキストでは，分類器による抽出は難しい．そこで，本稿では，テキストの特徴量に依存しない，投稿者単位での有用テキスト抽出を行った．

3. 用語定義

リスク，収益率，空売り，シャープレシオなどの株式関連用語について用語定義を行う．

リスク：不確実性，本稿では，リスクを定量的に表すために標準偏差を用いる．

収益率：投資収益率，投資した資本当りの利益率．

空売り：借りてきた株券を用いて売却を行うこと．株価の下落により収益を得ることが出来る．借入先に貸株料を払う．

シャープレシオ：リスクに対する収益の割合を表す指標．リスクが低く，収益が高いほどシャープレシオは高くなる．

4. 提案手法

「Yahoo!掲示板 > 株式カテゴリ」に寄せられた投稿テキストと，株価時系列データを用いて，投稿者の信頼度を評価する手法を提案する．投稿テキストは，投稿者名(Yahoo_ID)，投稿日，「投稿時の気持ち」(任意)，投稿内容を含み，テキストに含まれる「投稿時の気持ち」を投資戦略に対応付けることで株価時系列データによる評価を行っている．(例：「買いたい」と投稿した場合投資戦略を半全力買い戦略として，翌日の株価が上がっていればプラスの評価をする)

投稿者の評価は以下の手順で行う．

① 投稿毎に任意で付与された，5 段階の「投稿時の気持ち」{強く買いたい/買いたい/様子見/売りたい/強く売りたい}を抽出して，投資戦略を対応づける．
② 日単位かつ投稿者毎に投資戦略を対応づけ，収益率と投資戦略から評価を集計する．
③ 投稿者ごとに得られた収益率から，リスクに対するリターンの割合を示すシャープレシオを算出する．
④ 一定の投稿日数を満たし，かつ，評価期間のベースを上回ったものを，勝ち組投稿者として抽出する．

4.1 提案手法概要

投資戦略とは，投稿時に任意で付与できる五段階の「投稿時の気持ち」を定量化したものであり，直近約1年分のデータが入手可能である．

日付 x に，投稿者 i が投稿戦略 y の投稿を寄せた時に，投稿者 i に加算される評価 $E_i(x,y)$ を(1)式によって与える．評価 $E_i(x,y)$ は，投稿者の投資戦略を翌日の株価収益率 R_{x+1} によって評価したものである．投稿者の収益率を意味する．

$$E_i(x,y) = y \times R_{x+1} \quad \cdots(1)\text{式}$$

$$R_x = \frac{(x+1)\text{日始値} - x\text{日始値}}{x\text{日始値}} \times 100 - c \cdots(2)\text{式}$$

x：投稿の日付．株式市場の都合上，1日を朝九時から翌日の朝九時とする．朝九時までの投稿は前日のものとして扱う．

y：本稿では，投資戦略を以下のように対応づけた．空売り可能な貸借銘柄の場合，「強く買いたい」：$y=1$，「買いたい」：$y=0.5$，「売りたい」：$y=-0.5$，「強く売りたい」：$y=-1$，と設定した．1は全力買い，−1は全力空売りを意味する．

R_x：始値で買い，翌日の始値で売る収益率(%)．日付 x に投稿があった場合，収益率は翌日の始値で買い，翌々日の始値で売る収益率 R_{x+1} を評価に用いる．

c：貸株料，取引手数料などの取引コスト．検証では簡単のため，0とする．

次に，評価期間 $T: x \in T$，評価日数 cnt_i，評価累計 E_i，投稿者 i のリスク(標準偏差) $\sigma(E_i)$，シャープレシオ $Sr(E_i)$ を求める．

$$\sigma^2(E_i) = \sum_{\forall x \in T} \frac{1}{cnt_i} (Average(E_i) - E_i(x,y))^2 \cdots(3)\text{式}$$

$$V(E_i) = \sqrt{\sigma^2(E_i)} \quad \cdots(4)\text{式}$$

$$Sr(E_i) = \frac{E_i - \text{無リスク資産の利回り}}{V(E_i)} \quad \cdots(5)\text{式}$$

無リスク資産の利回り：無視できる程小さいので，0とする．

$Average(E_i)$：投稿者 i の平均評価値．

cnt_i：投稿者 i の評価日数．投資戦略を対応付けた投稿者に対して日単位で行った評価の回数．

投稿者の優劣は，単純なリターンの比較だけではなく，リスクとのバランスを考慮した上で決定する必要があるため，シャープレシオを求めている．

投稿の流動性を確保するため，一定以上の評価数（銘柄によって調整）を持つ投稿者を対象に，シャープレシオによるベースとの比較を行う．ベースには，一般的な投資戦略として，評価期間の平均日次収益率である ALL と，モメンタム(順張り）戦略((6)式)，リターンリバーサル（逆張り）戦略((7)式)を用いる．

順張り：$R_{x-1} > 0 \quad (x \in T) \Rightarrow x$ 日に買い…(6)式

逆張り：$R_{x-1} < 0 \quad (x \in T) \Rightarrow x$ 日に買い…(7)式

順張り戦略とは，前日の収益率 R_{x-1} がプラスの時に買う戦略，逆張り戦略とは，前日の収益率 R_{x-1} がマイナスの時に買う戦略である．

5. 投稿者評価実験

4節の提案手法を用いて実際に投稿者の評価を行った．実用的な数の投稿者が抽出されるかどうか，また，抽出された投稿者の成績の，株価予測可能性について検証することを目的とする．

5.1 実験データ

4節で述べた評価手法を用いて，実際に投稿者の評価実験を行った．実験では，投稿数の多い銘柄として，ソフトバンク(9984.T1)の投稿テキストを評価対象に用いた．評価期間：2005/10/3～2006/10/31，投稿数：約35万件(「投稿時の気持ち」の付与された投稿は約75,000件），投稿者(Yahoo_ID)数：9,995となった．

また，(6)，(7)式を用いて，評価期間中のベースを求めた結果を表1に示す．

表1：ベース

9984.T1	ALL	順張り	逆張り
平均日次収益率(%)	0.145	-0.288	0.587
平均標準偏差(%)	4.18	3.73	4.57
シャープレシオ	0.0346	-0.0772	0.129

評価期間では逆張り戦略が有効であることより，ベースとして逆張りのシャープレシオ $Sr_{base} = 0.129$ を用いる．このベースを上回る成績を上げた投稿者を信頼できる(勝ち組)投稿者として抽出する．

5.2 実験結果

評価日数 $cnt_i \geq 2$ の条件を満たす投稿者を選出した結果，3813名の投稿者が対象となった．対象となっ

た全ての投稿者を(1)式，(4)式により評価し，横軸に評価日数，縦軸に投稿者毎の平均評価値 $Average(E_i)$（収益率）・標準偏差 $\sigma(E_i)$ をとったプロットを図 1，図 2 に示す．

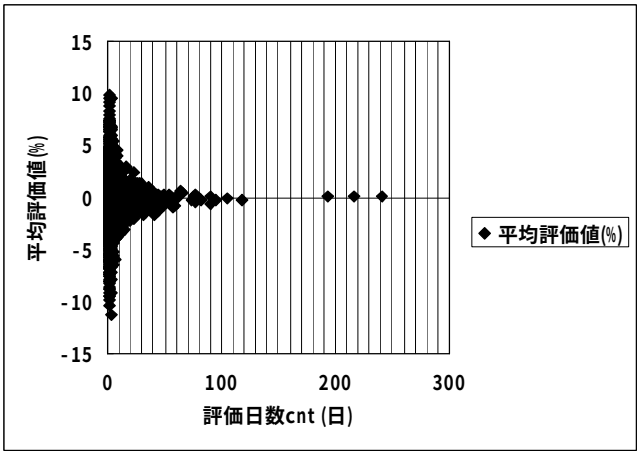

図 1：$cnt_i \geq 2$ の平均評価値プロット図

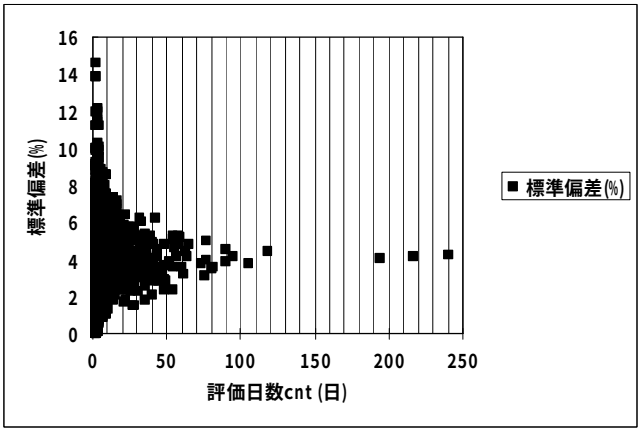

図 2：$cnt_i \geq 2$ の標準偏差値プロット図

平均評価値 $Average(E_i)$ は投稿者の全ての投資戦略を評価した値の平均，標準偏差 $\sigma(E_i)$ はそのばらつきを表している．図 1，図 2 より，評価日数が 10 前後を超えたところで平均評価値・標準偏差の値が安定していることが分かる．評価日数を増やせば標準偏差の信頼度が高まるが，評価対象の投稿者が少なくなってしまうため，評価日数の下限を 10 に設定して検証を行う．

$cnt_i \geq 10$ の条件を満たす投稿者を選出した結果，543 名の投稿者が評価対象となった．対象となった全ての投稿者を評価し，横軸に評価日数，縦軸に投稿者毎の平均評価値 $Average(E_i)$（収益率）・標準偏差 $\sigma(E_i)$・シャープレシオ $Sr(E_i)$ をとったプロットを，図 3 に示す．

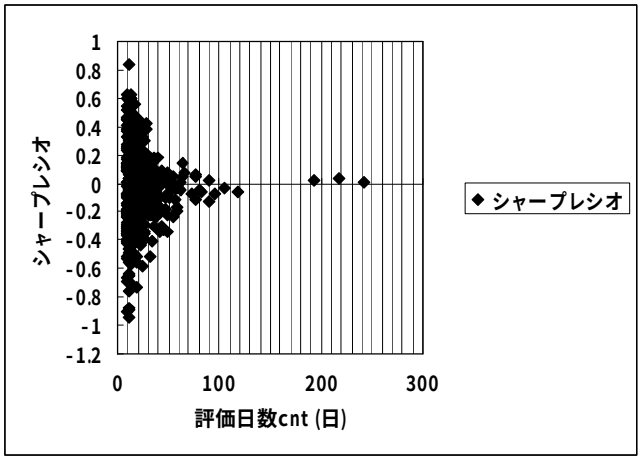

図 3：$cnt_i \geq 10$ のシャープレシオプロット図

図 2，図 3 より，標準偏差は，ALL 戦略の標準偏差値(4.18)付近に収束しているが，シャープレシオは評価日数の増加に伴い，マイナスの値に収束していることが分かる．次に，平均評価値，標準偏差，シャープレシオを評価日数 cnt_i 毎に集計した結果を表 2 に示す．

表 2：評価結果の平均値

評価日数下限	平均評価値(%)	平均標準偏差(%)	シャープレシオ
2	-0.313	2.93	-0.107
10	-0.182	3.92	-0.0465
50	-0.139	4.10	-0.0338

次に，$Sr_{base} = 0.129$ に対して，$cnt_i \geq 10$，$Sr(E_i) > Sr_{base}$ を満たす勝ち組投稿者を抽出した結果，116 名の投稿者を抽出した．同様に，$cnt_i \geq 10$，$-Sr(E_i) > Sr_{base}$ を満たす負け組投稿者を抽出した結果，206 名の投稿者を抽出した．勝ち組，負け組別の評価結果を表 3 に示す．

表 2,表 3 より，表 3 の平均評価値($cnt_i \geq 2$)が全体の平均収益率を下回っているため，投稿者の投資戦略を鵜呑みにしても中・長期的に負けてしまうことが分かる．また，表 2 より評価日数の増加に伴って平均評価値が上がっていることから，負け組は市場(掲示板)から徐々に退場していると考えられる．表 3 の結果より，実用的な数の信頼度の高い投稿者が抽出された．

表 3：$cnt_i \geq 10$における勝ち組と負け組の評価結果

	平均収益率(%)	平均標準偏差(%)	シャープレシオ	該当人数(名)
勝ち組	1.16	3.95	0.293	116
負け組	-1.12	3.85	-0.291	206
全体	-0.182	3.92	-0.0465	543

5.3 T 検定

前節ではシャープレシオを用いて勝ち組と負け組みの分類を行ったが，図 2 の標準偏差は表 3 の平均標準偏差に比べ高くなっている部分が見受けられる．つまり，抽出された投稿者の中には，株価の急騰や急落によって偶然，勝ち組や負け組みに分類された場合が考えられる．本節では前節で評価・抽出した結果に加えて，統計上有意な水準にある成績を持つ投稿者の抽出を行った．

5.4 T 検定による検証

2005/10/3～2006/10/31 における日次の収益率(翌日比）と，各投稿者の複数回の投稿データをもとに導出した投資戦略に基づく投稿日単位の収益率，上記の 2 つの収益率に対して T 検定を行った．対象となる投稿者はデータ数が 10 以上, つまり, 投稿日数 $cnt_i \geq 10$ を満たす 543 名である. 検証手法は, エクセルの TTEST 関数を用いて，両側分布，等分散の 2 標本を対象とする T 検定を行った．前提として，

- 帰無仮説 H_0 :「母平均値の差は偶然である」
- 対立仮説 H_1 :「母平均値に有意な差がある」
- 有意水準 5% で両側検定を行う．

横軸にシャープレシオ，縦軸に有意確率をとったプロット図を図 4 に示す．

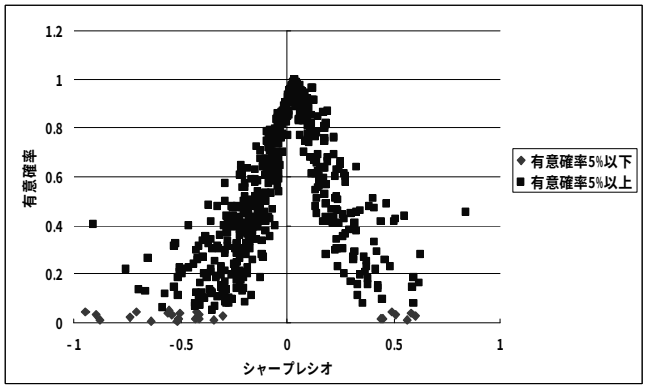

図 4:投稿者毎のシャープレシオ-有意確率プロット図

シャープレシオの絶対値の増加に伴って，有意確率が減少傾向にあるのが分かる．しかし，統計上有意な水準にある成績を持つ投稿者の抽出を行うためには T 検定による抽出を行う必要がある．帰無仮説 H_0 を棄却するためには有意確率 5%(図 4 中 0.05)以下でなければならない．結果として，有意水準 95%で「母平均値に有意な差がある」と言える成績を持つ投稿者は，勝ち組から 7 名，負け組から 18 名抽出された．

6. おわりに

評価実験の結果，勝ち組投稿者を 119 名，負け組投稿者を 210 名抽出することができた．さらに，負け組投稿者の投資戦略に対して裏切り戦略をとることで，有用な投稿者となるだけでなく，勝ち組の成績を上回ることが分かった．また，T 検定による統計上有意な投稿者を検証した結果，勝ち組から 7 名，負け組みから 18 名の投稿者が抽出された．

表 1，表 2 より，評価対象となった全ての投稿者 ($cnt_i \geq 2$)の評価値平均は,評価期間の平均収益率を下回り，絶対値では上回っている．よって，負け組みの投資戦略に対しての裏切り行為が，株価予測の観点から有効である可能性が示唆された．今後の課題として，より多くの銘柄において投稿者の評価を行い，優秀な投稿者を抽出して検証を行うことが挙げられる．そのためには，投稿の少ない銘柄においても充分な評価数を確保する必要がある．今後は，関連研究[2]を応用して,「特定のキーワードをあらかじめ用意して,出現数から投資戦略を決定する」方法等も組み合わせながら，既存の投資戦略の有無に依存されずに投稿者の評価を行う手法を考えていく予定である．

文　献

[1] “平成 17 年度株式分布状況調査の調査結果”:http://www.nse.or.jp/j/img/toukei/h17_youyaku.pdf

[2] “株式掲示板 > 株式カテゴリ”: http://messages.yahoo.co.jp/yahoo/Business___Finance/Investments/Stocks/index.html

[3] 酒井浩之, 増山繁経, “済新聞記事内容の個々の企業におけるインパクトの判定”, 情報処理学会研究報告 自然言語処理, Vol. 2006, No. 94, pp. 43-50, 2006.

[4] 立石健二，石黒義英，福島俊一，“インターネットからの評判情報検索”，情報処理学会研究報告，NL-144-11，pp.75-82，2001.

[5] 岡村 剛，角 康之，西田 豊明，“電子掲示板からの興味ある会話の抽出支援”，インタラクション 2005，情報処理学会主催，2005.

[6] 河野 勇介，嶋田 和孝，遠藤 勉，“Web 掲示板からの評価文の抽出と P/N 分類”，第 13 回電子情報通信学会九州支部学生会，2005.

[7] 峠泰成，山本和英，“手がかり語自動取得による

Web 掲示板からの評価文抽出”，言語処理学会第 10 回年次大会，S3-7，pp.107–110，2004.

[8] 藤村滋，豊田正史，喜連川優，“電子掲示板からの評価表現および評判情報の抽出”，人工知能学会第 18 回全国大会，*3F1-03*，2004.

[9] 藤村滋，松村真宏，岡崎直観，石塚満，“電子掲示板上の評判情報に基づく意思決定支援”，人工知能学会全国大会予稿集，2B1-05，2003.

[10] 松尾 豊，大澤 幸生，石塚 満，“電子掲示板における会話からのトピックの発見と要約”，人工知能学会全国大会，3D1-07，2002.

[11] 松尾豊 ， 大沢幸生 ， 石塚満，“電子掲示版における会話からのハイライト部分の抽出”，第 46 回人工知能基礎論研究会，2002.

[12] 栗田昌孝，真壁昭夫，“テキストマイニング～文脈や語感のニュアンスの定量化”，第四回行動経済学ワークショップ，2005.

[13] 加藤英明，天気と株価の不思議な関係，p. 239，東洋経済新報社，2004.