

株式掲示板におけるユーザ行動異常検知を用いた 相場操縦発見手法に関する研究

宮崎 邦洋^{1*} 松尾 豊¹

¹ 東京大学工学系研究科技術経営戦略学専攻

1 背景

現在、日本市場において個人投資家の存在が大きくなっている¹。この流れは日本の経済にとって好ましいものであり、政府としては個人の資金をさらに市場に向わせたい旨をその提言の中で述べている²。

1.1 Web 上における風説の流布

一方、インターネットの普及により、ソーシャルメディア上で風説の流布が行い易くなったため、個人投資家が安全に投資を行えなくなるような要因も増していると言える。実際掲示板上の風説の流布によって告発されたケースも複数ある³。

そのような行為をする者を相場操縦士や仕手師と呼ばれるが、この犯罪行為の標的にされやすいのもまた個人投資家である。何故なら個人投資家は機関投資家と比較して情報劣位にあり、非合理的な投資行動を取りやすいからである [1, 2]。しかし先述したように、東京市場の活性化という観点においては個人投資家の存在はとても大きいものである。現在の個人投資家売買比率増加の傾向を持続させるためにも、市場の公正性・信頼性を保つことはより重要になってくる。

もちろん証券取引等監視委員会は、市場の監視を行っているが、大量の情報が生成されるソーシャルメディアにおいて、すべての書き込みを監視することは容易ではない。

そのような問題に対して、Web マイニングの技術が有効である。Web マイニングではソーシャルメディア上に所謂ビッグデータを抽出・構造化・分析することで、現実世界のデータからでは得難い知見をより簡単に得ることができる。

この技術を用いることで、かつて困難であったソー

シャルメディアの監視を効率的に行えるようになる。

以上の内容を踏まえ、本研究では、インターネット株式掲示板において相場操縦を発見することを目的とする。そのため、掲示板におけるユーザの行動を分析し、相場操縦行為を行った可能性のあるユーザを発見する手法を提案する。

本研究が提案する手法は、急騰急落直前のユーザの行動が、通常時と比べどれくらい異常であるかを測定し、その異常具合をスコア化・ランク付けするものである。提案手法の使用用途については、金融庁など相場操縦を発見する主体を支援するシステムなどが想定される。

2 関連研究

2.1 相場操縦の研究

相場操縦とは一般に裁定取引機会の創出と利用と定義されている [3]。この相場操縦に関しては、主に直接的に取引戦略によって行われるもの（見せ玉等）と、間接的に情報操作などによって引き起こされるものがある。

本研究は後者に関係するものだが、[4]では、1600年から今日の情報化社会に至るまでにおいて、技術の発展によって株式銘柄に関しての情報が世の中に容易に拡散されやすくなり、それによって投資家が影響を受けられやすくなっていることが示されている。その技術の中でもインターネットの発達により、投資家が世界中に情報を発信できることや、世界中から情報を得ることができるようになったことに大きく貢献している。

インターネットにおける風説の流布において株式掲示板はその主たる媒体であるが、その株式掲示板を使用した相場操縦の研究については、Delortらがオーストラリアの株式掲示板に関する研究を行っている [5]。この研究では、HotCopper⁴という掲示板について、2008年のオーストラリア証券取引所上場銘柄 1825社に関して、100万余りの書き込みの分析から、相場操縦を検証し

*E-mail : miyazaki@weblab.t.u-tokyo.ac.jp

¹NTT データ経営研究所による調査
<http://www.jpx.co.jp/files/tse/news/31/b7gje6000003lhav-att/b7gje6000003lhjw.pdf>

²財務省「金融・資本市場活性化に向けての提言」:
http://www.mof.go.jp/about_mof/councils/kinyuukaigou/report/gaiyou.pdf

³証券取引等監視委員会 HP「不正取引について」
<http://www.fsa.go.jp/sesc/support/hukousei/hukousei.htm>

⁴<http://hotcopper.com.au/>

た。このラベルは複数あるが、Delort らは”ramping ” (相場操縦の疑い) とラベル付けされた投稿を分析し、これらの投稿とリターン、ボラティリティ、出来高に相関があることがわかった。

このように、インターネットを用いた風説の流布の研究は行われてはいるが、実際に個人行動を分析した上で相場操縦士の可能性があるユーザの発見に関してはまだ研究は行われていない。

2.2 Web 上の犯罪捜査の研究

Web マイニングには、犯罪捜査に特化した Web Crime Mining と呼ばれる分野がある。この分野の研究については、[6] が関連研究をまとめている。

Web Crime Mining の手法は様々なものがあるが、その一つが異常検知である。この異常検知は得られたデータの中での外れ値を探すものである。この手法は、その通常時と異なる動きを検出する声質から、ネットワークへの侵入（アクセスの急激な増加）や、詐欺の検出に用いられてきた。しかし、ソーシャルメディアにおけるユーザの行動の異常度を発見することには用いられていない。

本研究の新規性をまとめると以下の通りである。

- 風説の流布を行うユーザの発見をテーマにした研究であること。
- ユーザ行動の指標に異常値を用いたこと、特に、複数ユーザの共起異常を分析したこと。

3 手法

3.1 実験手法

まず、本提案手法をまとめたシステムの概要は図 1 のようになっている。このシステムは、まず株価デー

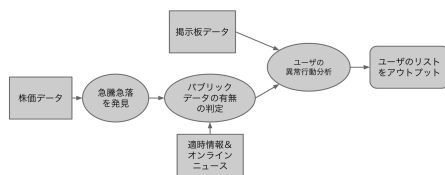


図 1: システムの概要図

タから急騰急落を発見し、その後パブリック情報と照らし合わせパブリック情報の影響を除去する。その後、急騰急落直前の書き込みを分析し、検出した異常行動にもとづいてユーザのランキングを作成するものであ

る。

それぞれの項目について、以下で説明する。

3.2 急騰急落の発見方法

今回の急騰急落の発見においては、忘却型自己回帰モデルを用いた計算アルゴリズムである、ChangeFinder アルゴリズムを用いた。ChangeFinder は、時系列データの各時点における外れ値スコア（本研究ではアノマリースコアと呼ぶ）を算出し、このスコアが高いほど、変化の割合が高いと判断する。

アノマリースコア算出のベースとなる考えは、自己回帰モデル（AR モデル）を用いた予測値と実測値との乖離の計測にある。スコアリングには AR モデルを用いており、ある時点 $t-1$ までのデータを学習し、時点 t での予測値と実測値の外れ値具合をアノマリースコアとして算出している。ここでアノマリースコアには対数損失を利用している。

このアルゴリズムの計算量はデータ数に対して線形オーダーに抑えられるため、4000 銘柄に関して一日あたり数百～数千の投稿を分析することに適していると言える。

また、実際山西らは金融データにおいてこのアルゴリズムを用い、急騰急落の発見において有効であることを示している [7]。

3.3 パブリックデータの導入について

今回取得したい急騰急落は株式掲示板の書き込みによるものであり、考えられる他の急騰急落の主要因であるパブリック情報については、基本的に株価はこれからの業績に対する期待値で動いているため、その影響を除外する必要がある。今回その除外方法として、ChangeFinder によって急騰急落を発見後、その 30 分前にその企業に関するパブリックデータがないかを検索した。

3.4 ユーザの異常行動の発見について

ユーザの行動は大別して 2 つあり、個別アカウントによるものと複数アカウントによるものである。

3.4.1 個別アカウントの異常行動の発見

今回はユーザが取れる行動として、書き込み行動、返信行動、評価行動の異常度を測定した。

書き込み頻度の異常

書き込み行動の異常度として、あるユーザが単位時間あたりに書き込む件数である書き込み頻度の異常度を測る。普段そこまで多く書き込んでいないユーザが、ある瞬間に急激に書き込んでいるとすれば、それは周りのユーザを煽ろうとする意思があると推察される。

その異常度の尺度としては、ポアソン分布を用いた。式は以下である。定数 λ に対し、自然数を値にとる確率変数 X が

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (1)$$

を満たすとき、確率変数 X はパラメータ λ のポアソン分布に従う。ここで、 e はネイピア数であり、 $k!$ は k の階乗、また λ は所与の区間内で発生する事象の期待発生回数である。今回は、2014/12 2015/5 の場中の時間帯を 30 分スロットに区切り、その中でユーザの書き込み頻度を通常時の書き込み頻度のポアソン分布のモデルとした。ここで、 P の値が小さければ小さい程、それだけそのユーザの急騰急落前の書き込み行動が異常であることを示す。

返信行動、評価行動の異常

返信行動の異常に関しては、急騰急落前の書き込みの中で、どれくらい多く返信をしているかの異常度を測るものである。通常時に比べ、ある時間帯に多く返信行動をしてれば、その瞬間にできるだけ多くのユーザに働きかける意思があったものと思われる。返信行動異常の尺度に関しては、割合の偏りを測るのに一般的に使用されるカイ二乗値を用いた。急騰急落時の投稿数（返信以外）と返信数を $C_{a\hat{r}}$ 、 C_{ar} 、通常時の投稿数（返信以外）と返信数を $C_{o\hat{r}}$ 、 C_{or} とそれぞれ定義すると、このとき、 χ^2 値は次のように定義される。

$$\chi^2 = \frac{(C_{ar} - C_{or})^2}{C_{or}^2} + \frac{(C_{a\hat{r}} - C_{o\hat{r}})^2}{C_{o\hat{r}}^2} \quad (2)$$

この値が大きければ大きいほど、通常時に比べ多く返信行動をしているということである。

また同様の手法を用い、評価行動の異常度も測定した。

3.4.2 複数アカウントの異常行動の発見

掲示板では、個別のアカウントだけではなく、複数のアカウントによって相場操縦が行われる場合がある。それは、複数人による共謀行為であったり、掲示板の匿名性を活かしたなりすまし行為の場合もある。

複数アカウントの異常行動の指標には、共起関係の指標である Jaccard 係数を用いた。ここで書き込み頻度異常を求める際と同様に、場中を 30 分のスロットで

区切る。そして、あるユーザ u の書き込みスロット群を S_u とする。そのときユーザ u_1 と u_2 の Jaccard 係数の定義は式 3 のようになる。

$$Jaccard = \frac{|S_{u_1} \cap S_{u_2}|}{|S_{u_1} \cup S_{u_2}|} \quad (3)$$

この Jaccard 係数に関して、通常時 (o) と急騰急落時 (a) を比較するために、それぞれの Jaccard 係数を $Jaccard_a$ 、 $Jaccard_o$ のように定義し、その割合をとることで、式 4 のように Suspicious_score を計算した。

$$Suspicious_score = \frac{Jaccard_a}{Jaccard_o} \quad (4)$$

ここで、Suspicious_score が高ければ高いほど、そのユーザ群が急騰急落時に多く同時に書き込んでいる可能性が高いことを示す。

3.4.3 ユーザネットワークの作成と分析

また、ここで取得されたユーザ群はあくまでペアのものであるが、実際の事件のように 3 つ以上のアカウントを使用している場合も想定される。そこで、この Suspicious_score で検出したペア群によってネットワークを作成し、そのネットワークを分析することで、3 人以上で同時に書き込んでいるユーザの検出を試みた。

4 使用したデータ

4.1 株価データ

実際に全ての銘柄において風説の流布が行われるわけではないため、本研究では特に風説の流布が行われやすい 320 銘柄において実験を行った。銘柄の選定条件は以下である。

- 発行株数が 5 千万株以下であること
- 資本金が 50 億円以下であること
- 株価が 300 円以下であること

この条件については [8] を参考にした。これらの銘柄について、分足のデータを東京証券取引所から購入し、使用した。

4.2 インターネット掲示板データ

本研究において使用した掲示板は、日本最大級の掲示板である textream である。

今回使用するデータは、2014/12～2015/5 までの半年間のデータであり、かつ 9:00～14:59 の場中のデータである（昼休みの時間帯を除く）。使用した 320 銘柄において、総ユーザ数 8,601、総コメント数 91,549 であった。

4.3 パブリック情報データ

表 1 に使用した記事の情報をまとめる。オンラインニュースに関しては、日本の市場の株価に影響を与える主要なメディアを選択した。

表 1: 使用したパブリックデータ

情報ソース	件数
適時開示情報	8,877
日経新聞	7,864
ブルームバーグ日本版	4,082
ロイター通信日本版	6,976
CNN 日本版	1,002
朝日新聞	3,992
読売新聞	5,211
産経新聞	14,560
ヤフーファイナンス	9,850
合計	62,396

表中の件数は、2014/12～2015/5 もので、かつ場中（9:00～14:59 除昼休み）に公開されたものである。

5 結果

本章では、前章で説明した実験手法によって実際に得られた結果について述べる。

5.1 書き込み頻度異常度分析の結果

書き込み頻度の異常分析においては、表 2 のような結果を得た。

表 2: 書き込み頻度の異常上位ユーザ

user	anomary_slot	anomary_comment	other_slot	other_comment	poisson_probability
not*****	1	10	2	2	7.60E-35
koj*****	1	7	9	15	1.34E-17
けん	1	8	24	46	1.67E-16
dea*****	2	13	4	11	1.11E-15
クロ	1	4	3	1	2.46E-14

この表は、左から

- ユーザのアカウント名 (user)
- ユーザが書き込んだ急騰急落のロット数 (anomary_slot)
- 急騰急落ロットにおいて書き込んだ書き込み数合計 (anomary_comment)

- ユーザが書き込んだ通常時のロット数 (other_slot)
- 通常ロットにおいて書き込んだ書き込み数合計 (other_slot)
- 通常書き込みをポアソン分布とした際の、急騰急落書き込み頻度の発生確率 (poisson_probability)

を示している。

一番右列のポアソン分布を用いた確率によって昇順にソートされている。この表を見ると、急騰急落前のみ異常な頻度で書き込んだユーザが取得できていることがわかる。例えばランキング 1 位のユーザは、通常時 2 ロットに対して 2 回の書き込みというペースで書き込んでいるにも関わらず、急騰急落前には 30 分回で 10 回の書き込みを行っている。

実際にここで抽出されたユーザが、風説の流布と思われる書き込みをしているかどうかを確認するため、書き込み内容を確認する。

実際にこのユーザの急騰急落直前の書き込みを見てみると、表 3 のようなものが続く。発言の多くは「もっ

表 3: 書き込み頻度異常上位ユーザの急騰急落直前の書き込み

user	stock_ticker	date	comment_time	comment
not*****	9478	1/29/2015	14:28	掲示板を使った投資家へのマインドコントロールだとい。
koj*****	6819	2/5/2015	13:06	もっと買いまーす!
けん	6495	1/20/2015	14:29	【ヤバイ】逃げときます! ヤバイ感じなんで...
dea*****	3266	1/26/2015	13:26	そもそも下げる理由が見当たらない。他の新興銘柄と違って確固たる材料が出てるしね。

と買いまーす」や「そもそも下げる理由が見当たらない。」など株価の評価や他人を煽るものが多く、これらの発言を見るだけでも他人に対して影響を及ぼそうとする様子が窺える。

5.2 返信行動異常度、評価行動異常度分析の結果

次に、銘柄に対しての返信行動に関する異常行動分析において得られた結果を示す。表 4 のようになっている。

表 4: 返信行動異常上位ユーザ

user	anomaly_reply_c	anomaly_comment_c	other_r_c	other_c_c	chi-square
drtetsu	22	49	158	542	565.49
tre*****	16	36	178	479	557.14
momoco	6	15	97	358	414.00
キャッシュハンタ007	2	2	161	240	393.04
	4	12	65	252	285.82

この表は、左列から

- ユーザのアカウント名 (user)
- ユーザが書き込んだ急騰急落の返信付き書き込み数 (anomaly_reply_c)
- 急騰急落スロットにおいて書き込んだ書き込み数合計 (anomaly_comment_c)
- ユーザが書き込んだ通常時の返信付き書き込み数 (other_r_c)
- 通常スロットにおいて書き込んだ書き込み数合計 (other_slot)
- 通常書と急騰急落時を比較したカイ二乗値 (chi-square)

となっている。この表は一番右列のカイ二乗値で降順にソートされている。

この表を見ると、上位ユーザは急騰急落直前に通常時に比べ多く返信行動をしていることがわかる。例えば、このランキングで1位のユーザは、通常時542回の書き込みのうち158回が返信行動であるのに対し(約29%)、急騰急落直前では、49回の書き込みに対し、22回の返信行動(約45%)を行っている。

次に、銘柄に対しての評価行動に関する異常行動分析において得られた結果を示す。表5のようになっている。

表 5: 評価行動異常上位ユーザ

user	anomaly_senti_c	anomaly_comment_c	other_s_c	other_c_c	chi_square
xko*****	3	3	115	132	235.15
star_holder	9	9	129	130	224.25
元気	4	4	63	75	122.47
わごん	2	3	18	112	120.30
漫 株太郎	5	5	43	64	87.97

この表の列の項目は、返信行動の項目のうち「返信」を「評価」に置き換えたものである。この表を見ると、上位ユーザは急騰急落直前に通常時に比べ多く評価行動をしていることがわかる。例えば、このランキングで1位のユーザは、通常時132回の書き込みのうち115回が評価行動であるのに対し(約87%)、急騰急落直前では、3回の書き込みに対し、3回の評価行動(100%)を行っている。

5.3 複数アカウントの異常行動

次に、複数アカウントの異常行動に関して分析結果について述べる。

表 6: 共起関係異常度

user1	user2	a_jaccard	o_jaccard	Suspicious_score
pro*****	sup*****	0.125	0.011	11.70
kur*****	ibu*****	0.333	0.029	11.47
pro*****	チャネラ	0.091	0.009	9.82
v*****	トム12	0.333	0.036	9.17
ibu*****	sup*****	0.167	0.022	7.60

5.3.1 Jaccard 係数を用いた共起関係異常度の分析

表6は、ユーザの共起関係異常度を分析して得られた結果の表である。

表6は、左列から、

- ユーザ1のアカウント名 (user1)
- ユーザ2のアカウント名 (user2)
- 急騰急落時のユーザ1とユーザ2のJaccard係数 (a_jaccard)
- 通常時のユーザ1とユーザ2のJaccard係数 (o_jaccard)
- 急騰急落時と通常時のJaccard係数の比率 (Suspicious_score)

という項目になっている。ここで、ランキング上位のユーザペアを見てみると、急騰急落時のJaccard係数が0.125であるのに対し、通常時は0.011しかなく、通常時に比べ、急騰急落直前に共起関係が非常に強くなっていることがわかる。

このランキングによって、急騰急落時のみに異常に同タイミングで書き込んでいるユーザペアが検出できた。

5.3.2 ネットワーク分析を用いた相場操縦グループの分析

ここでは、共起行動によって抽出されたユーザのペアリストから、ユーザの共起グループの発見を試みる。図2が、今回得られた分析によって得られたネットワーク図である。

ここで、各ノードはユーザを示す。ノードの大きさは anomaly_slot の大きさを示している。エッジは今回得られた Suspicious_score を示しており、太いほど値が高い。この図により、異常な共起関係をとっていたユーザがグループを作成していることがわかった。

6 評価

6.1 システムの評価手法

今回は有識者として金融の業務経験者の方々にアンケートをとることで、風説の流布と思われる書き込み

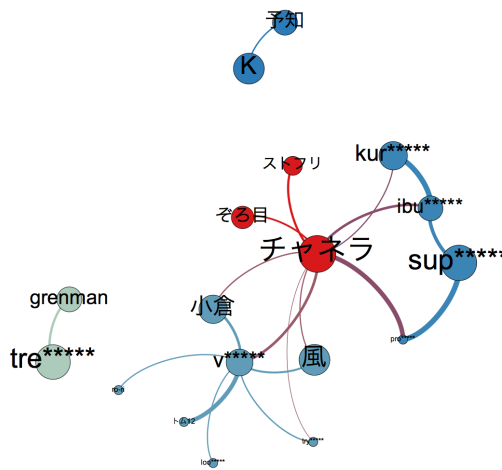


図 2: Jaccard 係数の異常度により作成されたネットワーク図

の評価を行った。

評価を行う分析は、本手法のベースになっている「書き込み頻度の異常」である。以下のタイプ別の書き込み 100 件ずつの 4 つの群から、一番「風説の流布」と思われる書き込みの集合であると思われるものから 1 ~ 4 位の順位をつけてもらうことにする。

- リスト上位ユーザの急騰急落前の書き込み (D)
- リスト上位ユーザの通常時の書き込み (B)
- 非リスト上位ユーザの急騰急落前の書き込み (C)
- 非リスト上位ユーザの通常時の書き込み (A)

6.2 結果

アンケートの結果を 7 に示す。

表 7: アンケート結果の集計表

書き込み群	リスト上位ユーザ	急騰急落前	1位	2位	3位	4位
A			5	1	0	4
B			1	2	4	3
C			0	5	4	1
D			4	2	2	2

表 7 はアンケートの集計結果である。

「リスト」上位ユーザは今回の手法によって取得できたユーザの書き込みを示している。

「1 位」~「4 位」は回答者が答えた回答である。最も風説の流布を取得できていると思われる書き込み群を 1 位に選んでもらっている。

「急騰急落前」は株価の急激な変動前の書き込みかどうかをしめしている。

この表からわかることを以下にまとめる。

- 提案手法で取得できた D が安定的に上位を獲得できている。
- B, C はそれぞれ中程度の順位を多く取得できている。
- A は 1 位取得回数は 1 位だが、4 位取得回数も 1 位である。

この結果から、D が最も上位の結果を示しているため、基本的には本提案手法で意図した結果が取得できていることがわかった。本手法は発言の異常度からユーザをランキング化するものであるが、その上位のユーザのコメントのみを他人に見せたことでこのような結果を得たことは、本手法の有効性を示していると言える。

7 考察

7.1 手法について

元々急騰急落という株価の動きを追うような研究は今までなかったが、今回このアルゴリズムを使用することで、適切に株価の「急騰急落」というパターンを使用できたと思われる。

また、パブリックデータの導入についても、適時開示情報については、上手く取得できているように思える。しかし一方で、オンラインニュースの取得については、取得できたものが殆ど経営に関わるようなニュースではなく、単なる銘柄の株価情報であったため、上手く取得できたとは言えないと思われる。

原因としては、まず企業の検索方法が上手くいっていないことが挙げられる。例えば同じ株式会社でも (株) や、株式会社、KK など、その表記は媒体によって異なる。これらの精度を上げることは今後の課題であると思われる。

また、他のニュースにおいて企業名が直接言及されなくても、影響を与えているケースを考えることも課題である。

7.2 実験結果について

評価、返信、書き込みに関しては、それぞれ上手く取得できていると思われる。今後の課題があるとすれば、書き込み内容の異常検知が考えられる。別書き込みを一つ一つ分析するような場合はなおさらである。

複数アカウントの行動については、今回通常の Jaccard 係数ではなく、通常時と急騰急落時を比較した Jaccard 係数の異常度を使用した。目的として「普段同時に書き込まないユーザ群が一緒に書き込む」という現象を取る上では理にかなっていると言える。実際のネット

ワークについては、単なるペアの集まりではなく、実際に3人異常のユーザによるネットワークが形成され、それを取得できた点に関して、手法は効果的だったと言える。より深い分析のため、リプライ関係や書き込み内容の類似度などを分析するとより精度が高くなると思われる。

7.3 評価について

アンケートによって本手法の一定の有効性を示すことはできたが、アンケート回答者をもっと多くすることも課題であると思われる。また、実際の現場で使用してみて、その有効性を評価されることも今後の展望として望まれる。

8 まとめ

本研究では株式掲示板において、ユーザの行動異常度を測定することにより、相場操縦行為を発見する手法について提案した。提案手法は、株価の急騰急落を前における掲示板でのユーザの行動を、様々な側面から分析することにより、ユーザの異常行動具合をランキング化するものである。

その結果、実際に風説の流布を試みた可能性のあるユーザのリストを作成することができた。

さらに、そこで得たランキング上位のユーザの急騰急落前の書き込みに対し、アンケートによって評価することで、本提案手法の有効性が評価できた。

本研究で提案された手法が実際の相場操縦発見の現場で使用されることになれば、著者の望むところである。

参考文献

- [1] Thomas C Chiang and Dazhi Zheng. An empirical analysis of herd behavior in global stock markets. *Journal of Banking & Finance*, Vol. 34, No. 8, pp. 1911–1921, 2010.
- [2] Félix Villatoro. The delegated portfolio management problem: reputation and herding. *Journal of Banking & Finance*, Vol. 33, No. 11, pp. 2062–2069, 2009.
- [3] Jean-Pierre Zigrand. Endogenous market integration, manipulation and limits to arbitrage. *Journal of Mathematical Economics*, Vol. 42, No. 3, pp. 301–314, 2006.
- [4] David J Leinweber and Ananth N Madhavan. Three hundred years of stock market manipulations. *The Journal of Investing*, Vol. 10, No. 2, pp. 7–16, 2001.
- [5] Jean-Yves Delort, Bavani Arunasalam, Maria Milosavljevic, and Henry Leung. The impact of manipulation in internet stock message boards. *International Journal of Banking and Finance*, Forthcoming, 2009.
- [6] Javad Hosseinkhani, Mohammad Koochakzaei, Solmaz Keikhaee, and Javid Hosseinkhani Naniz. Detecting suspicion information on the web using crime data mining techniques. *International Journal of Advanced Computer Science and Information Technology*, Vol. 3, No. 1, pp. 32–41, 2014.
- [7] Kenji Yamanishi and Jun-ichi Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 676–681. ACM, 2002.
- [8] 中原圭介. 仕手株でしっかり設ける投資術. 日本実業家出版, 2005.