株価予測のための新聞データでの表整理技術を用いた記載事項の内容分析

村田　真樹[a]

要約

Bollen らは，ツィッターでの感情分析を利用して，株価の予測がある程度可能であることを示した．われわれは，過去に，経済と感情に関わる表現を，新聞のテキストデータで単語ネットワークにより分析し，金利が上昇すると株価は下落することを示唆する結果を得た．本研究では，より論理的に株価を予測するために，新聞には，株価に関わるどのような内容が記載されているかを調査した．表整理技術を利用して調査した結果，(a)ダウ平均，(b)為替，(c)市場の情報 1，(d)市場の情報 2，(e)世界情勢，(f)海外株価指数，(g)以来に類するもの，(h)下げ幅，(i) 東証，(j)国際的な情報，(k)指摘事項，(l)利回り，(m)金融緩和，(n)業績，(o)取引開始，(p)関連，(q)貿易，(r)欧州，(s)警戒，(t)トランプ，(u)新型コロナ，(v)海外投資家の情報が新聞に書かれることがわかった．「トランプ政権」「新型コロナ」などの特殊な事象に関わる内容も得られた．

JEL 分類番号： C82

キーワード：株価，新聞データ，内容分析

*本研究は，公益財団法人石井記念証券研究振興財団の助成金を受けて実施された．

[a] 鳥取大学大学院工学研究科情報エレクトロニクス専攻，および，鳥取大学工学部附属クロス情報科学研究センター murata@tottori-u.ac.jp

1. はじめに

Bollen らは，ツィッターでの感情分析を利用して，株価の予測がある程度可能であることを示した(Bollen et. el., 2011)．われわれは，過去に，経済と感情に関わる表現を，新聞のテキストデータで単語ネットワーク(Kamihigashi et. el., 2017)において分析した(村田・金子・上東・馬, 2018)．実際にその分析を，「経済」「景気」「財政」「感情」に関わる単語を対象に行った．「景気」に関わる分析では，景気がよいと消費が増える，金利が上昇すると株価は下落すること，さらに国債に影響することを示唆する結果が得られた．新聞では，株価の値動きとともに，その時の経済環境を文章で説明している．本研究では，より論理的に株価を予測するために，新聞での株価に関わる文章を分析し，新聞には，株価に関わるどのような内容が記載されているかを調査した．

2. 表整理技術を用いた分析事例 1

2007 年から 2020 年の毎日新聞において，「前日終値比」「前日比」のいずれかを含み，「東京株式市場」「日経平均株価」を含む記事を抜き出した．2,084 件の記事が抜き出された．

次に，この記事群をクラスタリングした．K-means 法でクラスタ数 を 100 と指定してクラスタリングを行い，その中から密集度が高く件数も適度に多いクラスタを選択して実験データとする．密集度 0.989，21 件の記事を含むクラスタを抜き出し，実験データとする．ここで，密集度とはクラスタ内の情報の関連具合を表したものであり，似たような情報が詰まったクラスタは密集度が高くなる．fasttext で単語ベクトルを作りそれに基づき文書ベクトルを作り，文書間の類似度を算出し K-means でクラスタリングしている．密集度も同様に文書ベクトルを利用して算出する．

この 13 件の記事に対して，表整理技術(Murata et. al., 2021)で表に整理する．この表整理技術では，複数の記事を入力すると，記事中の文をクラスタリングして，表の行に，入力した記事，表の列に，クラスタリングした文を配置するような表を生成できる．よく似た内容の文が同じ列に配置された表が生成できる．実際にこの技術を利用したところ，13 列が生成された．表では左の方の列が重要度が高いものが配置される．重要度は密集度と表の充填率（空白の欄が少ないほどよい）の積でもとまる．手法の詳細は文献にある．生成された表のうちの一部を表 1 に示す．

13 列のうちの最後の 3 列は断片的なものや記事の著者の情報であり不要なものであった．残る 10 列の内容は，重要度順に以下のものがあった．その列にあった文の例も付す．

表１　生成された表の一部

記事タイトル	ダウ平均	為替	市場の情報
外為・株式：東証　８００円割れ寸前　一時６５８円安，年初来安値.	23 日の東京株式市場の日経平均株価は 2 日続落し，一時，前日終値比 658 円 08 銭安の 8016 円 61 銭まで値を下げた. 前日のニューヨーク市場でダウ工業株 30 種平均が 500 ドル超下落した流れを引き継いだ.	午後 0 時 45 分現在は同 531 円 27 銭安の 8143 円 42 銭.	23 日の東京株式市場は，前日の欧米市場の急落や急激な円高を受けて全面安の展開となった. 世界同時株安に歯止めがかからない状態に，市場では，03 年 4 月 28 日につけたバブル崩壊後の最安値を割り込むとの観測も強まっている.
ギリシャ:財政危機　日米欧で株価急落　支援決定，来月１０日で調整.	27 日のニューヨーク株式市場のダウ工業株 30 種平均は 7 営業日ぶりに急反落し，前日終値比 213・04 ドル安の 1 万 991･99 ドルと 2 月 4 日以来の大幅な下げを記録した. 日経平均株価は大幅反落し，取引時間中としては 3 営業日ぶりに 1 万 900 円台を割り込んだ.	円を買う動きも強まり，円相場は一時 1 ドル＝92・81 円と 3 営業日ぶりに 92 円台をつけた.	前日の欧米株も急落しており，市場では欧州発の世界的な株安が進むとの警戒感も出ている 28 日の東京株式市場でも，取引開始直後から売りが先行した.

(a)ダウ平均

9 日の東京株式市場は，ニューヨーク株式市場で 600 ドル以上株価が急落した流れを受けて，取引開始直後から売り一色となり，日経平均株価は一時，前日終値比 400 円超安い 8600 円台に下落.

(b)為替

同日の東京外国為替市場は，人民元切り下げを受けてドル高が進むとの思惑から円売り・ドル買いが優勢となり，円相場は一時 1 ドル＝125 円 24 銭まで下落し，6 月 8 日以来約 2 カ月ぶりの円安水準になった.

(c)市場の情報 1

香港，台湾市場などでも年初来安値を更新した.

(d)市場の情報 2

鉄鋼，海運など景気の先行きに敏感に反応する銘柄や，自動車，電機などの輸出関連株を中心に幅広い銘柄が売られている．大和証券ＳＭＢＣの宮沢一輝マーケットアナリストは「市場が不安感に支配され，過去の経験や企業体力では説明できない値動きをしており，

日経平均が 7600 円を割る可能性は十分ある」と指摘している.

(e)世界情勢

世界が同時に景気後退に陥るとの懸念が急浮上し，アジアや欧州でも株価を押し下げている．外国為替市場で欧米の景気不安からドルとユーロが売られ，急激に円高が進行したことで，輸出依存度が高い日本企業の業績悪化懸念が強まった.

(f)海外株価指数

アジア市場でも韓国総合指数が一時 9%超下落するなど主要な株価指数の下落が相次いでいる.

(g)以来に類するもの

取引時間中に 8000 円台をつけるのは，東日本大震災直後の 3 月 17 日以来，約 4 カ月半ぶり.

(h)下げ幅

1 日の下げ幅としては過去 9 番目の下げ幅.

(i) 東証

ＴＯＰＩＸも 2 日続落し，同 47・50 ポイント安の 841・73 で取引されている．東証 1 部の午前の出来高は 12 億 2800 万株.

(j)国際的な情報

アイスランド，ハンガリーが相次いで債務不履行の瀬戸際に追い詰められ，パキスタンが国際通貨基金に支援要請するなど，米国発の金融危機は債務危機に発展

自動で処理したものなので，綺麗にデータが分けられていないところもあるが，新聞で日経平均について記載する際，上記のような情報を同時に書くことが多いということがわかる.

3. 表整理技術を用いた分析事例 2

2 節で抜き出した 2,084 件の記事群を，K-means 法でクラスタ数 を 20 と指定してクラスタリングを行い，その中から密集度が高く件数も適度に多いクラスタを選択して実験データとする. 密集度 0.983, 79 件の記事を含むクラスタを抜き出し, 実験データとする.

文献と類似する技術で表を生成する. 列数を 20 個に固定して K-means 法で文をクラスタリングして表を整理する．表の行は，79 件の記事で，列は，文が 20 個のクラスタにクラスタリングされたものとなる．この 20 個の列の情報のうち，2 節で得られた(a)-(j)以外の情報でそれなりにまとまったものは一つであり，それを以下に示す．例文も付す.

(k)指摘事項

永見和彦・岡三証券投資情報部門理事は「世界同時株安といっても昨年と今年は要因が異

なり，市場の投資意欲にも大きな違いがある」と指摘する．

さらに，同じ 79 件の記事群で，列数を 50 個に固定して K-means 法で文をクラスタリングして表を整理する．表の行は，79 件の記事で，列は，文が 50 個のクラスタにクラスタリングされたものとなる．この 50 個の列の情報のうち，2,3 節で得られた(a)-(k)以外の情報でそれなりにまとまったものは一つであり，それを以下に示す．例文も付す．

(l)利回り

株や商品の資金が国債に流れ込む「質への逃避」が進み，日本の国債市場では，長期金利の指標となる新発 10 年物国債の利回りが前日比 0・055%低い 1・19%まで低下．

(m)金融緩和

とはいえ，取引時間中の 9000 円割れは，日銀が追加の金融緩和と事実上の「インフレ目標」導入を決めた 2 月 14 日以来，約 3 カ月ぶり．

(n)業績

今後の株価の見通しについて「日本企業の好調な業績が買い支えとなり，一方的に下落するとは考えにくい」との見方もある．

(o)取引開始

東京市場もこの流れを引き継ぎ，取引開始から全面安の展開になった．

(p)関連

株式市場では，自動車，電機などの輸出関連株の売りが加速した．

(q)貿易

米国を震源とする貿易戦争への懸念が，世界の金融市場を揺さぶっている．

(r)欧州

前日の海外市場では，欧州金融機関の資金調達が苦しくなるとの見方から，ユーロを売る動きが加速．

(s)警戒

ただ，このところの急上昇への警戒感も根強い．

(t)トランプ

トランプ政権は今年に入り，中国を主な標的に太陽光パネルの緊急輸入制限発動を決めるなど保護主義的な政策を実行．

(u)新型コロナ

米株価が 25 日に続落したのは，米保健福祉省のアザー長官が「米国で新型肺炎が広がる可能性が強い」との見方を示すなど「対岸の火事」とみられていた新型コロナウイルスが米国でも感染拡大することへの懸念が強まったため．

(v)海外投資家

関係者からは「低迷脱出には，結局は海外投資家が戻ってくるのを待つしかない」との声も出ている．

新聞で日経平均について記載する際，上記のような情報を同時に書くことが多いということがわかる．表で列数を増やすとより多くの内容を取得できる．「トランプ政権」「新型コロナ」などの特殊な事象に関わる内容も得られた．

4. おわりに

本稿では，毎日新聞のデータから，日経平均株価に関係する記事のうち，類似する記事群を抜き出し表整理技術を利用して，表に整理した．表の行に記事群，表の列に，類似する文が集まるように文をクラスタリングしたクラスタが配置される．この表で，列の情報を調べることで，どのような情報が新聞に書かれるかを調査した．その結果，(a)ダウ平均，(b)為替，(c)市場の情報 1，(d)市場の情報 2，(e)世界情勢，(f)海外株価指数，(g)以来に類するもの，(h)下げ幅，(i) 東証，(j)国際的な情報，(k)指摘事項，(l)利回り，(m)金融緩和，(n)業績，(o)取引開始，(p)関連，(q)貿易，(r)欧州，(s)警戒，(t)トランプ，(u)新型コロナ，(v)海外投資家の情報が新聞に書かれることがわかった．「トランプ政権」「新型コロナ」などの特殊な事象に関わる内容も得られた．今後は得られた情報をもとに論理的に株価予測をする手法を検討する．また，記載情報を分析し，人が非合理的な行動をする例がないかの調査も行いたい．

引用文献

Bollen, J., H. Mao, X. Zeng, 2011. Twitter Mood Predicts the Stock Market, Journal of Computational Science, Volume 2, Issue 1, pp. 1-8.

Kamihigashi T., M. Murata, Q. Ma, 2017. Use of Web Search Engines in TF-IDF based Word Network Construction for Extracting Useful Information, Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS 2017), pp. 1-4.

村田真樹，金子徹，上東嵩，馬青, 2018. 単語ネットワークを用いた経済と感情に関わる表現の分析，行動経済学会第 12 回大会, pp.1-6.

Masaki Murata, Kensuke Okazaki, and Qing Ma, 2021. Improved Method for Organizing Information Contained in Multiple Documents into a Table, Journal of Natural Language Processing, Vol. 28, No. 3, pp.802-823.