株価データと決算短信を用いた 日経平均市況概況記事の自動生成

Automatic generation of articles concerning Nikkei 225 market using stock price and summaries of financial statements

根岸 龍[1]　酒井 浩之[1]　永並 健吾[1]

Ryu Negishi[1], Hiroyuki Sakai[1], Kengo Enami[1]

[1]成蹊大学
[1]SEIKEI University

Abstract: 近年，個人投資家の数は増加している．個人投資家にとって投資をする判断材料として，日経平均株価の動向，株価が上昇下降した銘柄，どのような株価の変動要因があったといった情報が重要になる．これを端的に表しているのが日経平均市況の概況について言及した記事になる．この日経平均市況概況記事の情報は，投資経験の少ない個人投資家はもちろんのこと，投資経験の多い投資家にとっても投資の判断材料として一番身近な情報で需要が高いといえる．しかし，作成にあたって日経平均株価の動向，変動した企業の情報等、多くの情報を参照する必要があるうえに，取引日ごとに掲載するため，人手で作成するには多くの時間と労力が必要になる．そこで，本研究では，株価データと決算短信から株式市場の動向，株価の変動要因となる事業，製品，社会背景等を投資テーマとして抽出することで，日経平均市況概況記事を自動的に生成することを目的とする．具体的には，株価データから大きく変動した企業群を抽出し，各企業の決算短信から抽出したキーワードを基にクラスタリングによる絞り込みを行う．絞り込みを行った企業群の各キーワードをスコア化し投資テーマとして抽出する．株価データと投資テーマを基に日経平均市況概況記事の自動生成を行う．

1. はじめに

近年，個人の投資家の数は増加している．日本取引所グループが行った調査によると，個人投資家は株主数合計（延べ人数）全体の 97.5%を占めており、2020 年度は，前年度比 308 万人増加して 5,981 万人となり，7 年連続で増加している[１]．

個人投資家にとって投資をする判断材料として株価の動向，株価が上昇下降した銘柄，株価の変動要因といった情報が必要になる．個人投資家にとって情報や知識を収集することはとても重要であるが，膨大なデータを人手で収集し，判断することは困難である．

この投資家にとって重要な情報を端的に表しているのが日経平均市況の概況について言及した記事（以降，日経平均市況概況記事とする）になる．この日経平均市況概況記事の情報は，投資経験の少ない個人投資家はもちろんのこと，投資経験の多い投資家にとっても投資の判断材料として一番身近な情報で需要が高いといえる．以下に記事の例を示す．

> 日経平均株価は反発した。前日の米株高や円安が支援材料。好業績が報じられたトヨタや日立など主力の輸出関連株がけん引した。証券や機械、非鉄金属も高い。ただ米国のエボラ出血熱の感染問題などを警戒し、上値では戻り待ちの売り圧力も強い。買い一巡後は伸び悩んでいる。

図１　日経平均市況概況記事の例

一方で，日経平均市況概況記事の作成にあたって日経平均株価の動向，株価が大きく変動した企業を確認するのはもちろんのこと，企業の事業内容，製品，業績，プレスリリースの情報，さらには，政府や日銀の経済対策，社会情勢，景気等の，日々変化する様々な情報を参照しなければならない．さらに，市場取引日ごとに記事を掲載するため，人手で作成するには多くの時間と労力が必要になる．

そこで本研究では，株価データと決算短信から株式市場の動向，株価の変動要因となる事業，製品，社会背景等を投資テーマとして抽出し，日経平均市

況概況記事を自動的に生成することを目的とする．具体的には，株価データから大きく株価が変動した企業群を抽出し，各企業の決算短信から抽出したキーワードを基にクラスタリングによる絞り込みを行う．絞り込みを行った企業群の各キーワードをスコア化し投資テーマとして推定する．そして，株価データと投資テーマを基に，日経平均市況概況記事の自動生成を行う．

2. 関連研究

テキスト情報から市場動向を予測する関連研究として文献[２]を挙げる．文献[２]では「金融経済月報」のテキスト情報から，数年にわたる比較的長期の市場動向分析を支援するテキストマイニング技術の提案をしている．この研究では金融経済月報の記述に大きく依存し，長期の市場動向を対象としているが，本研究では，その日における投資テーマを，決算短信を用いて推定している．

記事データと株価データを用いた関連研究として文献[３][４]を挙げる．文献[３]では過去の記事データと株価データを用いて，記事に含まれる語句の出現と株価変動との関連を計算し，それに基づいて新しい記事内容の株価変動への影響の推測を行っている．しかし，[３]では，銘柄ごとの推測になるため，他銘柄の影響や企業の細かな業績の影響が少なく，投資する上での判断が容易ではない．また，結果より株価データと記事データの相関性を確認することができている．文献[４]では，ニュース記事を単語の係り受け構造が考慮された構文木で表現し，テキストの係り受け構造を捉えるため，再帰的ニューラルネットワーク（ Recursive Neural Network ； RNN）を感情分析へと応用したモデルを用い，株価動向の予測をしている．[3]と同じようニュース記事と株価の相関性が実証されたことと，単語をベクトルで意味的に表現した Word Embedding が株価動向推定に対して有用であることが示された．それらに対して本研究では，株価のデータと企業の決算短信を用いて，記事を自動生成することを目的としており，タスクが異なる．

企業における業績要因を推定する関連研究として文献[５][６]を挙げる．[５][６]では，個人投資家への投資判断支援のために，企業の決算短信より業績要因を抽出する手法を提案している．この手法で抽出した業績要因文を本研究では使用している．

3. 提案手法

3.1. 手法の概要

提案手法の概要を以下に示す．

Step1: 株価データから取引日における日経平均株価の変動率，および，株価が変動率±5%以上変動した企業を取得．

Step2: Step1 で取得した企業ごとに，決算短信から業績要因文[６]を抽出し，抽出された業績要因文から業績要因に含まれる企業キーワードと，そのキーワードの重要度に応じたスコアを取得．

Step3: Step2 で取得した企業ごとの企業キーワードとスコアをもとに階層型クラスタリングを用いて企業のクラスタリングを行い，クラスタの企業集合における企業キーワードから投資テーマを取得．

Step4: 日経新聞記事から日経平均市況概況記事を抽出し，その記事に含まれる，変動要因を表す重要語，変動の大きさを示す表現（「続伸」など）を取得．

Step5: Step3 で取得した投資テーマを示すキーワード，Step1 で取得した企業，Step4 で取得した表現を基に，日経平均市況概況記事を生成．

3.2. 株価変動率の取得

取引日における日経平均株価の変動率，および，株価が変動率±5%以上変動した企業を取得する．ここで，変動率の計算には以下の式を用いる．

$$\text{変動率} = \frac{\text{当日の終値} - \text{前日の終値}}{\text{前日の終値}} \times 100 \qquad (1)$$

3.3. 業績要因文からの企業キーワード取得

酒井らの手法[5]により，決算短信から抽出した企業ごとの業績要因文と，同じく酒井らの手法[6]により業績要因文から抽出された企業ごとの重要キーワード（以降，企業キーワードと定義）と重要度を表すスコアが付与されたデータを使用する．

ここで，業績要因文とは，企業の決算短信から業績に関わる文を抽出したものである．具体的には，新製品，新サービスや，好調，または不振になっている事業の要因が記述されている．企業キーワードは，業績要因文に含まれているキーワードに重要度に応じたスコアが付与されたものになっている．

具体的な業績要因文の抽出手法としては，少数の手がかり表現（「が好調」等）を人手で与え，それに

係る節を取得する．その中で頻繁に出現する表現を共通頻出表現（「売り上げ」等）として抽出し，それに係る節を取得し，新たな手がかり表現を抽出する．この手順を共通頻出表現が獲得されなくなるまで，もしくは予め定めた回数まで繰り返し，業績要因文の抽出に必要な手がかり表現を抽出する．

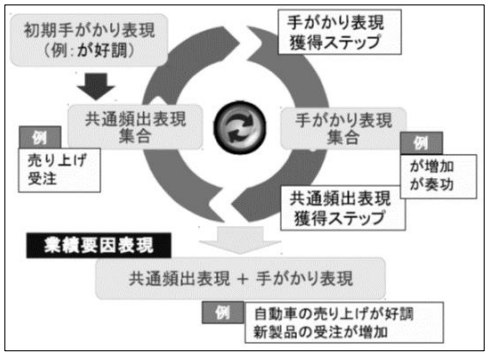

図２ 共通頻出表現・手がかり表現自動獲得手法の概要

さらに，この業績要因文の集合から，TF・IDF 法を基に，決算短信に名詞が出現する確率に基づくエントロピーを加えたスコア[６]により，企業キーワードとスコアを求める．企業キーワードの例を表１に示す．

表１ 東レの企業キーワード例

企業キーワード	スコア
炭素繊維	23.036
フィルム	12.941
複合材料事業	28.211
ケミカル事業	23.594
情報通信材料	29.568

決算短信全体を対象とするよりも，業績要因文のみに出現する企業キーワードに対象を絞ることで，精度の高いキーワード抽出が可能となる．

3.4. 階層クラスタリングによる企業の絞り込み

企業のクラスタリングを行うことで，変動率±５％を基準としただけの企業集合の中で，テーマ性に関係なく株価が上昇している企業を除去する．実際に，変動した企業の中でテーマ性が異なる企業をクラスタリングにより除去する例を図 3 に示す．図 3 では，自動車，鉄鋼業，運送業，化学系に関連する企業が大きく変動した企業として取得されている．この企業群の中でクラスタリングによって分類することで変動した企業の中で多く関連している自動車以外の企業を除去することが可能になる．

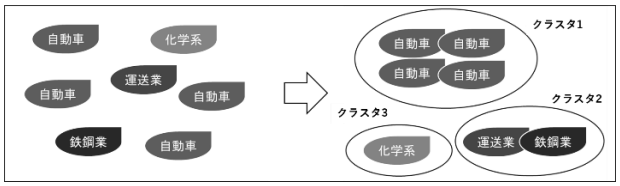

図３ クラスタリングによる除去

前節で得られた企業ごとの企業キーワードとスコアを用いて、階層型クラスタリングの単連結法を用いて企業のクラスタリングを行う．そして，最大クラスタが企業数の 1/3 に絞られるまで行う．企業数を限定することが目的なため，クラスタ中の最大の類似度をクラスタの距離とし，クラスタ同士を融合するため一つのクラスタに企業が集中しクラスタが大きくなりやすい単連結法を採用した．企業 d と企業 t との類似度 $sim(V_d, V_t)$ は，以下のコサイン距離で定義する．

$$sim(V_d, V_t) = \frac{V_d \cdot V_t}{\| V_d \| \cdot \| V_t \|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (2)$$

$$V_d = \{x_1, x_2, \ldots, x_n\},\ V_t = \{y_1, y_2, \ldots, y_n\}$$

V_x：企業 x の企業キーワードを要素，スコアを要素値とするベクトル

２つのクラスタ間(C_i, C_j)の類似度 $sim(C_i, C_j)$は以下の式で定義される．

$$sim(C_i, C_j) = max_{x_k \in C_i, x_l \in C_j}\ sim(x_k, x_l) \quad (3)$$

$sim(x_k, x_l)$ ： 企業 x_kと 企業 x_lとの類似度

3.5. 投資テーマを表す語の推定

企業キーワードから取引日の投資テーマを推定する．まず，企業ごとの企業キーワードスコアと出現回数から 3.4 節で取得された企業数が最大のクラスタに対応する単語スコアを生成する．ここで，クラスタに対応する語 n のスコア $WS(n)$ は以下の式で求める．

$$WS(n) = \log(K(n)) \sum_{i \in k} w(n, i) \quad (K(n) \geq 3) \quad (4)$$

$w(n, i)$：語 n の企業 i における企業キーワードのスコア

$K(n)$：語 n を企業キーワードとしてもつ企業の数
K ：クラスタに含まれる企業の集合

この式は，多くの企業に出現している語ほどスコアが高くなるが，1 つの企業にしか出現しない単語はスコアが低くなる．取得されたクラスタにおける企業集合の中で多くの企業に共通して出現すれば，その日において重要な単語，すなわちテーマになり得る可能性が高くなる．そして，$WS(n)$のスコアが上位に出現する語を，その日の投資テーマとして推定する．本手法によって推定された投資テーマの例を表２に示す

表２　本手法によって推定された投資テーマの例

日付	投資テーマ
2010/1/21	スマートフォン, 産業機器, フォン, 産業用　デジタルカメラ, 家電
2010/2/12	鉄鋼, 化学, 製造装置, 台湾, 建材, 電池, 鉱業　航空機, プラント
2010/2/26	自動車業界, 自動車関連, 金属, 貿易, 北米市場, 自動車メーカー
2010/3/2	衣料, 婦人, 衣料品, 百貨店, カジュアル, ファッション, 土地再
2010/3/5	観光, オフィス, 不動, 賃貸, 宿泊, 開発事業, 都心, 都市

3.6. 日経平均市況概況記事からの情報抽出

日経平均市況概況記事の自動生成にあたり，実際の日経新聞記事における日経平均市況概況記事でよく使われる表現（「続伸」等）を抽出して，記事の生成に使用する．日経新聞記事からの日経平均市況概況記事の抽出については，酒井らの手法[７]にて行う．具体的には，記事の１文目に「日経平均株価は」というフレーズが出現している記事を正例として学習データを自動生成し，自動生成された学習データによる深層学習にて，日経平均市況概況記事の抽出を行う．この日経新聞における日経平均市況概況記事から，記事生成の固有情報となる，株価変化の表現，記事における重要語を取得する．

3.6.1. 日経平均株価の変動表現の取得

株価の連続した変化，変化率を表す表現（株価変動表現と定義）を，当日，前日の変動率から取得した．表現の種類は，日経新聞の概況記事から変動率が＋の日付からポジティブ表現の種類を取得し、－の日付からネガティブ表現の取得を行い，頻度が高いものを変動表現として設定した．また，変動の大きさを表す表現を変動率によって変動表現の前に設定し，学習を行った．取得した株価変動表現の例を以下に示す．

ネガティブ表現： 下落，続落，反落，
ポジティブ表現： 上昇，続伸，反発
大きさ表現：大幅に，小幅に

図３　取得した株価変動表現の例

3.6.2. 重要語の取得

日経新聞における日経平均市況概況記事において，株価が大きく変化した業種や投資テーマ等の言及がある。この情報を重要語として取得を行った．具体的には，「関連」の記述の直前に現れている名詞を抽出した．また，句点「、」「と」「や」といった並列で表現されている単語も取得対象とした．抽出した重要語の例を以下に示す．

輸出，内需，自動車，電機，資源，ハイテク
精密，半導体，中国，不動産，情報通信，金融
機械，素材，建設，石油，小売り，通信，銀行
電気機器

図４　取得した重要語の例

3.7. 日経平均市況概況記事の生成

3.7.1. Word2Vec モデル

投資テーマは，決算短信から抽出した企業キーワードを基に抽出しているため，地名や会計用語等のノイズが完全に除去できておらず，記事の内容として使用するためには不十分になっている．そこで，推定された投資テーマと 3.6.2 節で取得した重要語との類似度を求め，投資テーマと類似度が高い重要語を記事の生成に使用する．ここで，投資テーマと重要語との類似度の計算には Word2Vec モデルの類似度を用いて行う．ここで，本手法では，日経新聞記事 10 年分の全記事を用いて学習した 300 次元の Word2Vec モデルを使用する．なお，学習において，複合語を単語に分割せず，１つの語としてあつかったデータにて Word2Vec モデルの学習を行った．

3.7.2. 投資テーマと関連する重要語の推定

3.5.で取得した投資テーマを基に 3.6.2.で取得した重要語からその日における重要語の推定を行う．3.7.1.の Word2Vec モデルによる類似度計算により，取得した投資テーマから類似度が高い重要語を求め，日経平均市況概況記事の生成に使用する重要語とし

て推定する．

3.7.3. 投資テーマによる変動企業の推定

3.5.で取得した投資テーマを基に 3.2.で取得した企業から日経平均市況概況記事に記載する企業の推定を行う．3.2.で取得した企業（株価が変動率±5%以上，変動した企業）から 3.3.と同様の手法で企業キーワードを取得する．ある日における投資テーマと企業の企業キーワードを 3.7.1.の Word2Vec モデルによって類似度計算を行い，類似度が高い企業キーワードをもつ企業を日経平均市況概況記事に記載する企業として推定する．

3.7.4. 日経平均市況概況記事の生成

株価変動表現，投資テーマに関連する重要語，投資テーマによる変動企業の情報を用いて，日経平均市況概況記事を生成する．本手法にて生成した記事の例を以下に示す．

> ５日の日経平均株価は続伸。半導体、デジタル家電、電機、精密機器関連が上昇。銘柄ごとにみると、東京計器、ミタチ産業、日本カーバイド工業、大日本スクリーン製造等が上昇。

図５　本手法にて生成された日経平均市況概況記事の例

ここで，記事例の「続伸」が株価変動表現，「半導体、デジタル家電、電機、精密機器」が投資テーマに関連する重要語である．

4. 評価

提案手法から生成された記事は，日経平均株価について言及する第一文，上昇下降した投資テーマについて言及した第二文，上昇下降した企業について言及した第三文といった構成になっている．また，第一文は日経平均株価の変動率により生成されるため，市場取引日ごとに生成されるが，第一文，第二文に関しては，企業の株価が大きく変動しない日も存在するため生成されない場合がある．そのため，生成した記事は，市場取引日ごとに記事の文構成の長さは変化している．

株価データは[８]の 2010 年～2014 年の 5 年分を用いた．企業に関しては，東証一部上場企業のみを取得対象とした．また，日経新聞における日経平均市況概況記事の重要語の取得には，2000 年から 2009 年の 10 年分の日経新聞の記事を使用した．

評価方法には取引日ごとに本手法で生成した記事と，実際の日経新聞における日経平均市況概況記事との類似度を，SentenceBert[９] によって求めることで行う．すなわち，本手法によって生成された日経平均市況概況記事と，実際の日経平均市況概況記事との類似度が高ければ，本手法によって生成された記事は適切であるという仮定に基づく評価手法である．具体的には，ある取引日に本手法によって第三文まで生成する場合（本手法），比較手法として，本手法による第一文のみ，第一文から第二文までの記事の 3 つのパターンで，日経新聞の日経平均市況概況記事との文書間類似度の計算を行う．

評価結果を以下の表 1 に示す．表 1 では，記事生成の対象年ごとに，本手法による第一文のみ，第一文から第二文まで，第一文から第三文まで（本手法）の記事の 3 つのパターンと日経新聞の日経平均市況概況記事との文書間類似度の平均を示している．

表１　生成記事の類似度結果

年	1 文	1-2 文	1-3 文（本手法）
2010 年	0.393	0.459	0.488
2011 年	0.383	0.454	0.488
2012 年	0.392	0.475	0.514
2013 年	0.403	0.504	0.524
2014 年	0.403	0.483	0.522

表 1 より，本手法である 1-3 文の記事（本手法）がすべての年において最も類似度が高い結果となっており，本手法によって生成された記事が適切であることを示している．

5. 考察

今回の評価方法では，生成された記事の有効性を示す結果となった．生成した記事は日経新聞の日経平均市況概況記事に比べて短い文章で構成されているため，最長のパターンで類似する可能性は高い．また，抽出した重要語の抽出元は日経新聞の日経平均市況概況記事になっているが，生成する年と異なることで適した評価ができていると考えている．

ここで，記事生成がうまくいった 2014 年 10 月 16 日の例を挙げる．実際に本手法によって生成された記事を以下に示す．

１６日の日経平均株価は大幅に反落。半導体、情報通信、デジタル家電、精密機器関連が上昇。石油、鉄鋼、運輸、建設関連が下降。

図６　生成された日経平均市況概況記事
（2014 年 10 月 16 日）

次に 2014 年 10 月 16 日の株式市場について言及した日経新聞における日経平均市況概況記事を以下に示す.

日経平均株価は大幅に反落し、下げ幅は一時４００円を超えた。１５日の欧米株安や為替市場での円高・ドル安で売りが先行した。東証１部上場銘柄の約９割が下げた。業種別では海運や鉄鋼、鉱業、石油などの下げが目立つ。日経ジャスダック平均株価も反落した。

図７　日経新聞の株概況記事
（2014 年 10 月 16 日）

図７において，株価が下降した業種について「海運や鉄鋼、鉱業、石油など」と記載されており，図６より生成した記事においても「石油、鉄鋼、運輸、建設関連」と記載されている．石油，鉄鋼，輸送と高い精度で株価が大きく変動した業種を推定できている．また，日経平均株価の変動も同じ表現となっている．本手法では東証一部上場企業を対象としているため，東証一部上場企業が目立って大きく変動した日に関しては記事生成の精度が上がっている．

6. まとめ

本研究では，株価データと決算短信から株式市場の動向，株価の変動要因となる事業，製品，社会背景等を投資テーマとして抽出し，日経平均市況概況記事の自動生成を行った．評価結果として，生成した日経平均市況概況記事の有効性を示した．

今後は，投資テーマ推定の精度を改善し，記事生成の向上を目指す. また, 文書生成モデルと比較し，日経平均市況概況記事に適切なモデルの評価を行い，生成した記事内容のより適切な評価を目指す.

参考文献

[１] https://www.jpx.co.jp/markets/statistics-equities/examination/nlsgeu000005nt0v-att/j-bunpu2020.pdf　2020 年度株式分布状況調査の調査結果について　<要約版>　(2021 年 7 月 7 日)

[２] 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”, 情報処理学会誌, Vol25, No.3，pp383-387，2010.

[３] 張へい，松原茂樹：“株価データに基づく新聞記事の評価”，人工知能学会，vol.22，1E2-4, 2008.

[４] 秋田諒，吉原輝，関和広，上原邦昭:“再帰的ニューラルネットワークによる感情分析モデルを用いた株価動向予測”，人工知能学会，vol.29，1J4-OS-13a-5, 2015.

[５] 酒井浩之，松下和暉，北島良三：“学習データの自動生成による決算短信からの業績要因文の抽出”，日本知能情報ファジィ学会誌, vol. 31, no. 2, pp. 653 661, 2019.

[６] 酒井浩之，西沢裕子，松並祥吾，坂地泰紀：“企業の決算短信 PDF からの業績要因の抽出”，人工知能学会論文誌, vol.30, no.1, pp.172 182, 2015.

[７] 酒井浩之，坂地泰紀，和泉潔，松井藤五郎，入江圭太郎, “学習データ自動生成による市況分析コメント作成のための要因文と補完情報の抽出”，第 34 回人工知能学会全国大会, 1D3GS1303-1D3GS1303, 2020.

[８] http://mujinzou.com/　汲めども尽きない 無尽蔵

[９] Reimers Nils and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp.3982–3992, 2019.