1P-1

Web情報に基づく株価変動の予測支援のための知識構築

竹内秀太† 奥村紀之† 奥村学‡

香川高等専門学校 情報工学科†
東京工業大学 精密工学研究所‡

1. はじめに

近年，インターネットの爆発的な普及に伴い，株取引が身近なものとなり，専門家以外の個人投資家が急増している．また，株式投資で利益を得る方法として，株の売買の差額により利益を得る方法がある．この方法では，後に株価が上昇すると考えられる株を購入し，購入時よりも高い株価で売却することで，利益を得ることができる．そのため，この方法により利益を得る場合，株価変動の予測は重要なものとなる．

上昇傾向のみに限らず，株価の変動を予測する際には，その株を発行している企業に関する様々な情報が株価変動予測の重要な判断材料となる．また，インターネットが普及しWeb上から様々な情報を入手することが可能となったことから，このような企業の情報はWeb上から取得することができると考えられる．

そこで，本研究では，Web 上から企業の情報を抽出し，テキストマイニングシステムを用いて分析を行うことで得られるデータを基に株価変動を予測することを最終目標としている．また，現在は，株価変動を予測するためのキーワードを知識ベースとして構築することを目標としており，本稿ではこれについて述べていく．

2. 関連研究

株価に関する研究は広く行われており，その中には，張へいら[1]が行っている新聞の記事データと株価データを用いた研究がある．

この研究では，過去の新聞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を推測している．

張へいらは，株価は企業価値を表す指標の一つで，企業の業績や財務等の状況，社会情勢の影響により変動するものであると述べた．一方，新聞はそのような株価変動の要因となる事柄を広範に提供するメディアで，記事内容と株価変動が関連すると予想した．また，ニュース報道を投資判断の材料として利用する投資家が多いことからも，客観的にみて記事内容が株価変動に影響を与える可能性が高いという考えのもと，研究を行った．その結果，実際の株価変動データと比較したところ，両者の間の相関を確認することができている．

そこで本研究では，テキストマイニングのソフトウェアである IBM ® Content Analytics を用いて，過去の新聞の記事データから，企業名と単語との相関が高い記事を抽出し，株価変動データと比較する．これにより，単語と株価変動の関連性を確認し，それを基に知識ベースを構築する．

3. 新聞コーパスによる分析

IBM ® Content Analytics を用いて，過去の日本経済新聞の記事を構造化し，新聞コーパスを作成する．この新聞コーパスを利用して任意の単語を含む記事を絞り込み，そこからその単語と相関の高い形容詞を含む記事に絞り込む．その後，さらにその形容詞と相関の高い企業名を含む記事に絞り込む．これにより抽出された記事の日付付近の株価データを企業ごとに確認することで，任意の単語がその形容詞と高い相関を示した際の株価変動の傾向を分析し，株価変動の予測につなげる(図 1)．

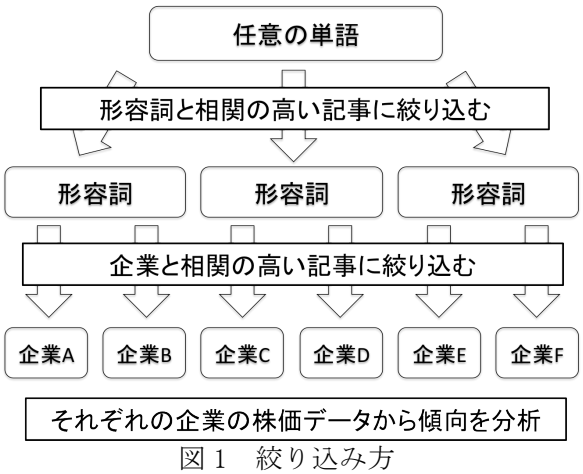

図 1 絞り込み方

「A Knowledge Construction for Supporting Prediction of Stock Price Fluctuation Based on Web Information」
† Shuta Takeuchi
† Noriyuki Okumura
‡ Manabu Okumura
† Kagawa National College of Technology Department of Information Engineering
‡ Tokyo Institute of Technology Precision and Intelligence Laboratory

4. 分析結果

前項に記した方法で分析を行った結果の一部を表 1 に示す．表 1 を見ると，上方修正という単語に着目して分析を行った場合，その株価の多くは上昇しており，下方修正という単語に着目して分析を行った場合，その株価の多くは下落していることが分かる．また，相関の高い形容詞で区分することで，変動の時期や程度などを含めた傾向が得られている．上方修正と相関の高い形容詞と相関の高い企業であるトモニホールディングス株の株価変動を図 2 に示す．

しかし，高い相関を示した様々な企業において高い確率で共通する株価変動であったことから共通する株価の傾向としているが，一部の記事において例外的傾向が出ているものも存在する．この例外的傾向を示した記事の特徴を把握し，絞り込みの際に除外することができれば予測精度が向上すると考えられるため，例外的傾向を示す記事の除外方法については今後検討していく予定である．

また，予測精度については，相関の高い単語による区分の回数や株価変動の傾向を確認する企業数をより多くすることでも，向上が見込める．

表 1　株価の傾向

任意の単語	相関の高い形容詞	共通する株価の傾向
上方修正	穏やかだ	その月は始値より終値が高い
	堅調	その月は始値より終値が高い
	小幅だ	その月の始値と終値は大きく変化しない
下方修正	弱気だ	２ヶ月後の月の終値がその月の終値より低い
	神経質だ	３ヶ月後の月の終値がその月の終値より低い
	軟調だ	その月は始値より終値が低い

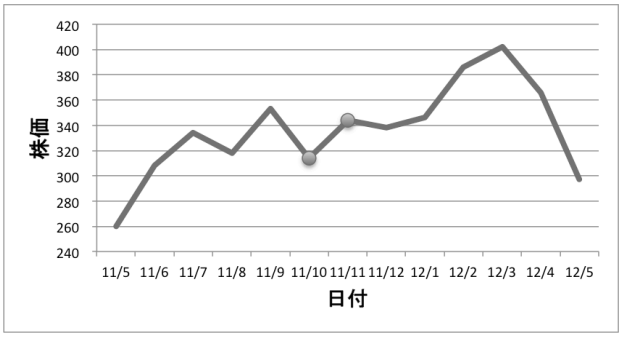

図 2　トモニホールディングス株の株価変動

5. 問題点と解決策

機械的に解析を行うため，企業名が企業としての意味合い以外を持っている場合，その企業と関係のない記事を抽出してしまう可能性があり，株価変動予測が困難となる．また，単語と相関が高い企業名であっても，その記事の内容と企業が直接関係していない場合が存在し，これを検出してしまうことで予測精度が低下する．

これらの問題点の解決方法として，多義性のある名称の企業については別途，絞り込み方法を検討していく予定である．また，相関が高いにも関わらず記事内容と企業が直接関係していないものが検出される問題については，企業名が括弧で括られているなど，特定の条件下において検出されるという限定的なものであるため，絞り込みの際にこの条件に含まれる記事を除外することで解決できると考えている．

6. おわりに

本研究では，IBM ® Content Analytics を用いて，日本経済新聞の過去の記事データから任意の単語とその単語と相関の高い形容詞が含まれる記事を絞り込み，そこからその形容詞と相関の高い企業名が含まれる記事を抽出した．その後，抽出された記事の日付付近の株価データを企業ごとに確認することで，任意の単語がその形容詞と高い相関を示した際の株価変動の傾向を分析し，株価予測につながる知識ベースの構築を行った．

今後は，例外的傾向を示す記事の除外方法および多義性のある名称の企業における記事の絞り込み方法を検討していくとともに，より多くの単語に対して分析を行い知識ベースの拡張を進める．また，相関の高い単語による区分の回数や株価変動の傾向を確認する企業数をより多くすることで，予測精度の向上を目指す．

参考文献

[1] 張へい，松原茂樹：株価データに基づく新聞記事の評価，第 22 回人工知能学会全国大会論文集，2008

[2] 小川知也，渡部勇：株価データと新聞記事からのマイニング，情報処理学会研究報告，2001