人工知能学会研究会資料
SIG-FIN-013-09

複数のテキスト情報を用いた株式市場動向の分析

Analysis of Stock Market Trend Using Multiple Text Data

敷地 琢也 [1]　和泉 潔 [1,2]

Takuya Shikichi[1], Kiyoshi Izumi[1,2]

[1] 東京大学大学院工学系研究科
[1]School of Engineering, The University of Tokyo
[2] 科学技術振興機構 CREST
CREST, JST

Abstract: 本研究では、個人投資家に株式の投資判断に必要な情報を提示する投資支援システムを作ることを目的とする。そのために、複数のテキスト情報を用いて、銘柄に変動的に影響する要因の分析と、企業の取引企業や事業内容といった基本情報を取得する。

1. 諸言

1.1 背景

近年、多くの個人投資家が株式投資に参加するようになってきている。しかし、株式に影響する要因は様々であり、投資判断を瞬時に行うのは非常に難しい。これは、情報の膨大さや関連性の複雑さから生じており、投資家を支援する技術の必要性が高まってきている。

1.2 既存研究

テキストデータを用いて金融市場の変動を分析する研究はいくつか存在する。和泉ら[1]は、CPR 法という手法を用いて金融市場の長期の変動の要因を分析している。CPR 法では、市場変動の要因が分析することができるというメリットがあるが、取り扱うテキストが個別銘柄に特化していないため取引先企業や事業内容といった企業の基本的な情報を取得することは困難である。張ら[2]は、新聞記事を評価し、その評価値と個別銘柄の株価変動率に相関があることを示しているが、どういう要因で株価が変動したかまでは分析できていない。また、よって、市場に影響を与える要因を分析し、さらに銘柄の情報を付加して提示する技術はいまだに確立されていない。

テキストデータを用いて金融市場の変動を分析する研究はいくつか存在する。和泉ら[1]は、CPR 法という手法を用いて金融市場の長期の変動の要因を分析している。CPR 法では、市場変動の要因が分析することができるというメリットがあるが、取り扱うテキストが個別銘柄に特化していないため取引先企業や事業内容といった企業の基本的な情報を取得することは困難である。張ら[2]は、新聞記事を評価し、その評価値と個別銘柄の株価変動率に相関があることを示しているが、どういう要因で株価が変動したかまでは分析できていない。また、よって、市場に影響を与える要因を分析し、さらに銘柄の情報を付加して提示する技術はいまだに確立されていない。

1.3 研究の目的

本研究では、個人投資家に株式の投資判断に必要な情報を提示する投資支援システムを作ることを目的とする。具体的には、図 1 のように投資家が読むオンラインニュース中の単語に反応して、その単語に関連する銘柄、そしてさらに銘柄に関連する単語を結びつける。

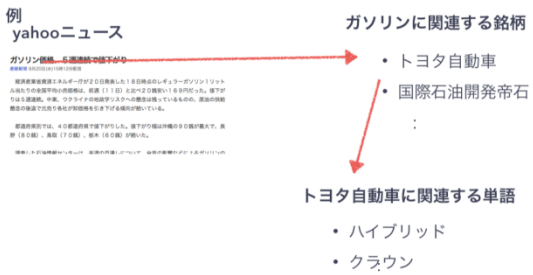

図 1 投資支援システムイメージ[3]

そのために、複数のテキスト情報を用いて、変動的に銘柄に影響する要因の分析と、企業の静的な基本情報を取得する。

2. 分析手法

2.1 手法概要

図 2 に手法概要を示す。新聞記事データと株データを基に、CPR 法でニュース記事中の銘柄に影響を与える重要単語に対して、どの程度影響があるかを紐づけたリストである単語関連銘柄リストを作成する。次に、各銘柄の wikipedia の記事データに対して、TFI-DF 法を適用し、銘柄を構成する単語群を持ったリストを作成する。それらの二つのリストを基に、単語が影響を与える銘柄、そしてその銘柄を構成する基本的な情報を示す。

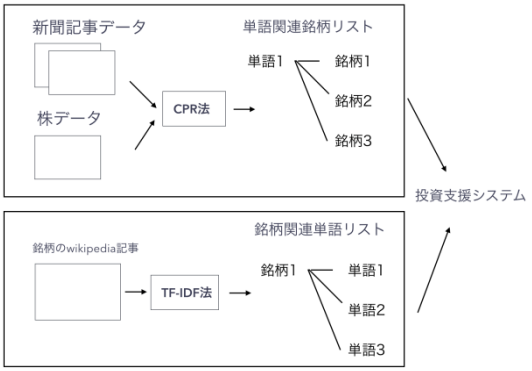

図 2 手法概要図

2.2 CPR 法

CPR 法は、共起解析、主成分分析、回帰分析の 3 つのステップで構成される分析手法である。まず、共起解析では、形態素解析を行い、ニュース記事を単語ごとに区切る。そのうち、名詞、動詞、形容詞のみを抽出する。そして、同一の文中に隣接して出現する組み合わせのうち、少なくとも一方が日経シソーラス[4]に含まれる組みあわせを数え上げる。これは、文章中の経済に関係ない単語を排除するためである。その際、出現回数が閾値以上のみを出現とし、出現パターンの行列を作成する。その行列に対して、主成分分析を行い情報の削減を行う。得られた主成分を説明変数として、予測する株データに対して重回帰分析を行う。

$$r_{i,t} = a_0 + \sum_{j=1}^{n} a_{i,j} x_{j,t} \tag{1}$$

重回帰分析の目的変数には、相対的株価変動率と出来高変動率を用いた。相対的株価変動率は、株価変動率$r_{i,t}$から市場全体の変動率$R_{i,t}$を引いたものである。用いる新聞記事データは日経新聞朝刊のデータであるため、当日の朝に発行された後、その内容に対して市場がその日の始値から反映されていくと考えられる。よって、t 日の相対的株価変動率に、$p_{i,t}$に t 日の始値、$p_{i,t-1}$に t-1 日の終値を用いた。また、前日と比較してどの程度取引量が増えるかを示す出来高変動率に対しては、$p_{i,t}$に t 日の出来高、$p_{i,t-1}$に(t-1)日の出来高を用いた。

$$r_{i,t} = \frac{p_{i,t} - p_{i,t-1}}{p_{i,t-1}} \tag{2}$$

$$r'_{i,t} = r_{i,t} - R_{i,t} \tag{3}$$

CPR 法で得られた主成分の寄与率と重回帰分析の係数を基に、単語 i が銘柄に与える影響度E_iを以下の式で定義する。

$$E_i = \sum_{j=1}^{n} a_{i,j} \mu_{i,j} \tag{4}$$

$a_{i,j}$は、重回帰分析の際の係数を指し、$\mu_{i,j}$は第 j 主成分に単語 i が与える寄与率を示す。これらの一連の流れにより、単語が各銘柄にどの程度影響を与えるかを定量化することができる。

2.3 TF-IDF 法

TF－IDF 法は情報検索の分野で一般的に用いられる単語の重みづけの手法である。文書 d 中のキーワード i の重みは、(1)式で定義される。$tf_{i,d}$が、対象テキスト j 中のキーワード i の出現頻度を表し、N は総文書数、df_iは、キーワード i が出現する文書の数を示す。

$$w_{i,d} = tf_{i,d} \times \log\left(\frac{N}{df_i}\right) \tag{5}$$

それぞれのキーワードの文章内の出現頻度が高く、特定の文章にのみ出現している単語は重要とみなされる。個別銘柄の wikipedia の文書に対して、形態素解析を行い、名詞のみを抽出する。各キーワードに対して、各銘柄の文書の集合を総文書として TF-IDF を行う。各銘柄に対する単語とその重みをリストとして保持する。

3. 実験

3.1 使用データとパラメータ設定

TOPIX１００という東証一部上場銘柄の中で時価総額、流動性の高い大型銘柄で構成される指標が存在する。その構成銘柄 100 銘柄の内 2012 年に株式分

割を行っている銘柄を除いた表1に示す95銘柄を実験に用いた。

表 1 実験に用いた 95 銘柄(略称含む)

セブン&アイ HD	信越化学	武田薬品	新日鐵住金
コマツ	日立	パナソニック	ソニー
ファナック	日産自動車	トヨタ自動車	ホンダ
キャノン	三井物産	三菱商事	三菱 UFJ FG
三井住友 FG	みずほ FG	野村 HD	東京海上
三井不動産	三菱地所	JR 東日本	NTT
NTT ドコモ	ソフトバンク	国際石開帝石	大和ハウス
積水ハウス	日揮	アサヒ GHD	キリン HD
味の素	東レ	旭化成	三菱ケミカル HD
花王	エーザイ	第一三共	大塚 HD
OLC	富士フィルム	資生堂	JX
ブリヂストン	旭硝子	JFE	住友金属鉱山
住友電工	SMC	クボタ	ダイキン
東芝	三菱電機	日本電産	富士通
京セラ	村田製作所	日東電工	三菱重工業
いすゞ自動車	スズキ	富士重工業	ニコン
HOYA	リコー	大日本印刷	任天堂
伊藤忠商事	丸紅	東京エレクトロン	住友商事
ユニ・チャーム	イオン	りそな HD	横浜銀行
三井住友トラスト	オリックス	大和証券	NKSJHD
MS&AD	第一生命	T'&D	住友不動産
JR 西日本	ヤマト HD	ANA HD	中部電力
関西電力	東京ガス	大阪ガス	セコム
ファーストリテイリング	アステラス	大東建託	

期間は、2012 年の 1 月 4 日から 2012 年 12 月 28 日のうち 30 期間をランダムで選択し、学習期間を 30 日間、予測期間を学習期間の次の日、1 日後とした。CPR 法における単語の出現パターン行列を作る際の出現頻度の閾値は 2 とし、主成分分析の際の主成分の数の上限は 15 とした。また、TF-IDF 法で用いる wikipedia の各銘柄の記事は、ノイズを除去するために一段落が 20 文字以上の文章のみ用いた。

3.2 実験結果

予測が実際の CPR 法における 1 日後の全 95 銘柄、30 期間の相対的株価変動率の平均予測精度は、55.6%という結果になった。また、出来高変動率の 1 日後の平均予測精度は、60.8%という結果になった。

表 2 に相対的価格変動率を用いた際の単語に対する影響度が高い銘柄を示す。表 3 に相対的価格変動率を用いた際の単語に対する影響度が高い銘柄を示す。

表 2 2012 年 3 月 22 日～2012 年 5 月 7 日の期間で相対的価格変動率を用いたプラスの影響度が高い 5 銘柄の例

自動車	大震災	原子力
日東電工	いすゞ自動車	ユニ・チャーム
MS&AD インシュアランス	SMC	旭硝子
横浜銀行	ファナック	セブン&アイ・ホールディングス
ソニー	富士通	丸紅
ファーストリテイリング	日産	-

表 3 2012 年 3 月 22 日～2012 年 5 月 7 日の期間で出来高変動率を用いたプラスの影響度が高い 5 銘柄の例

自動車	大震災	原子力
村田製作所	ユニ・チャーム	澤藤電機
日本電産	ANA	ファーストリテイリング
任天堂	オリックス	京セラ
ニコン	NTT	ファナック
ダイキン	日本電産	ヤマト

表 4 日東電工、村田製作所の TF－IDF 法結果の上位重要後 5 つ

日東電工	村田製作所
液晶	検知
工業	コンデンサ
日立	ジャイロセンサ
中継	一輪車
マラソン	部品

表 4 に、表 2、表 3 で「自動車」が影響する銘柄 1 位であった日東電工と村田製作所の TFI-DF 法による上位単語を示す。日東電工は、自動車と関係のある単語が少ないが、村田製作所は、コンデンサや部品といった単語が上位単語に並んでいる。村田製作所の wikipedia 文章から「コンデンサ」が含まれる文章を全て抜粋したところ、「主力商品はセラミックスコンデンサで世界随一のシェアを誇る。」、「村田製作所は積層セラミックコンデンサでトップの地位を走る」、「コンデンサ - スマートフォンやカーエレクトロニクス等に使用されるチップ積層セラミックコンデンサなど。」と 3 つの文章で書かれていた。これらの文章から村田製作所とコンデンサは強い結びつきがあり、カーエレクトロニクスという記述から、「自動車」に関連があることが分かる。

4. 課題と展望

本研究により、CPR 法を用いた個別銘柄の相対的株価変動率と、出来高変動率の予測が短期的な予測にも有用であることが示せた。また、TF-IDF 法を用いて wikipedia から、個別銘柄の基本的な構成要素を抽出し、CPR 法と結びつけて、考えることができた。今後の課題としては、以下の 3 つが挙げられる。

1 つ目に、銘柄の種類や検討期間をより広範囲に広げることが挙げられる。どういった期間や銘柄に本手法が有効なのかを検証する必要がある。

２つ目に、投資支援システムの開発が挙げられる。最終的な投資支援システムを評価するためには、個人投資家にシステムを使ってもらいアンケートを取るという手法が考えられる。そのために、個人投資家が実際に使えるように GUI で投資支援システムを開発し、使用してもらって検証する必要がある。

３つ目に、各パラメータの妥当性の検証が挙げられる。主成分の数や、出現パターンの行列にする際の閾値などで結果が大きく変わることも考えられる。そのため各パラメータの最適化をする必要がある。

参考文献

[1] 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, 人工知能学会論文誌, Vol. 25,No. 3, pp. 383–387 (2010)

[2] 張 へい, 松原 茂樹株価データに基づく新聞記事の評価 JSAI2008-1E2-4(2008)

[3] Yahoo 経済ニュース http://headlines.yahoo.co.jp/

[4] 日本経済新聞デジタルメディア.日経シソーラス http://telecom21.nikkei.co.jp/help/contract/price/23/help_KIJI_thes.html