人工知能学会研究会資料
SIG-FIN-014-07

複数テキストを用いた個別銘柄の影響要因と関連銘柄の抽出

Extraction of Factors and Related Stocks of Individual Stocks Using Multiple Textual Data

敷地 琢也 [1]　和泉 潔 [1,2]　山田 健太 [1,3]

Takuya Shikichi[1] Kiyoshi Izumi[1,2] Kenta Yamada[1,3]

[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, Japan Science and Technology Agency
[3] 科学技術振興機構 さきがけ
PRESTO, Japan Science and Technology Agency

Abstract: In this study, we proposed a new method for extracting factors and related stocks which affect individual stocks. We combined two text-mining methods which are CPR method for news articles and TF-IDF for summary of financial statements. We showed how stocks are connected through factors in each terms.

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。特に，近年はウェブ上に株式投資に直接または間接的にも関連性があるテキスト情報が常に溢れている．そこで、機械学習を用いたテキストマイニング手法によって、テキスト情報と市場変動の関係性を発見し市場分析に応用する研究が増えてきた[1]。経済指標やマーケットのテクニカル指標等の数値情報には指標化されていないような情報をテキスト情報から素早く自動的に抽出することが期待されている。

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。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、変動要因の詳細な分析は行っていない。和泉ら[3]は、CPR 法を日銀金融経済月報に対して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、市場変動の要因を分析することができるというメリットがあるが、企業の取引先企業や事業内容といった基本情報が含まれていないため、異なった期間には適用されない要因ばかりが抽出されてしまうというデメリットが存在する。個人投資家の投資支援には、ある事象が起こった際の影響を受ける対象とともに、次の投資に活かすためにも、その影響の要因を知ることが必要である。

1.3 研究の目的

本研究では、個人投資家の投資判断支援のために、個別銘柄に影響を与える要因と関連銘柄を抽出する。CPR 法の市場変動の要因を分析できるというメリットを活かしつつ、基本的な情報を含む別のテキストで得られる情報を補うことによって、周辺の期間でも適用可能な要因の抽出を目指す。

それにより投資家はある事象が起こった際の影響を受ける銘柄、そしてその影響の要因を知ることができる。つまり、膨大なニュースなどのテキストデータから効率的に情報を集めることができ、また、企業間の未知の関連性を知ることが可能になるという利点がある。

2. 分析手法

2.1 手法概要

図 1 に手法概要を示す。用いるテキストデータとして、新聞記事データと企業の決算短信データを用いた。新聞記事は、毎日内容が変わるテキストであるのに対して、決算短信のテキストデータは、取引企業や事業内容など時間で変動しない単語を含む文章が多い。

新聞記事データと株価データを基に、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要因単語リストを作成する。次に、個別銘柄の決算短信データに対して、TFI-DF 法を適用し、銘柄を構成する基本単語を持ったリストを作成する。それら二つのリストを基に、単語のフィルタリングを行う。残った単語から関連銘柄を抽出し、再探索を行う。

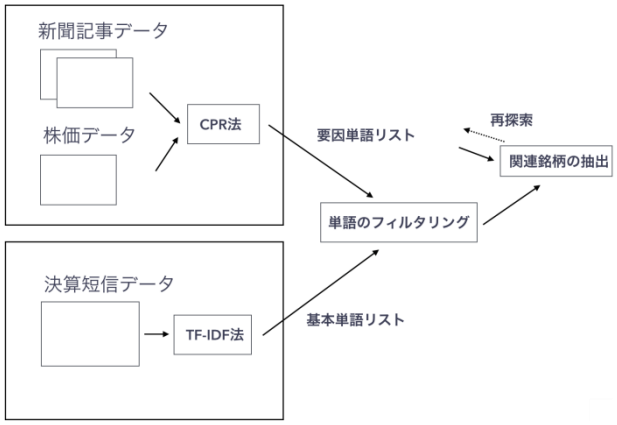

図 1 手法概要図

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。

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

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

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。

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

$a_{i,j}$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した。

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。個別銘柄の決算短信の文書に対して、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。

2.4 単語のフィルタリングと関連銘柄

CPR 法で得られた要因単語リストと TF-IDF 法で得られた基本単語リストを用いて、単語のフィルタリングを行う。フィルタリングには、一方のリストに含まれる単語が他方のリストの単語の文字列の一部に一致する場合、その単語を保持することによって行う。保持した単語を、要因単語リストを基に上位の影響度に持つ銘柄を抽出することによって、関連する銘柄を抽出する。

3. 実験

CPR 法の予測精度の検証と、個別銘柄に影響を与える要因の抽出、銘柄の基本単語の抽出、そして関連銘柄の抽出を目的に実験を行う。

3.1 使用データとパラメータ設定

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 TOPIX100 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や決算短信データを取得できなかった銘柄を除いた表 1 に示す 89 銘柄を実験に用いた。

表 1 実験に用いた 89 銘柄(略称含む)

セブン&アイ HD	信越化学	武田薬品	アステラス製薬
新日鐵住金	コマツ	日立製作所	パナソニック
ソニー	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
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	オリエンタルランド
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	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	MS&AD	第一生命	T'&D
住友不動産	JR 西日本	ヤマト 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			

期間は、2012 年の 1 月 4 日から 2012 年 12 月 28 日のうち、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、1 日ずつずらした 217 期間を用いた。新聞記事データには、日本経済新聞朝刊のデータを用いて、決算短信データは[5]で使用された企業の Web ページに掲載されている決算短信 PDF から抽出したテキストデータを用いた。

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、CPR 法の予測精度の検証には、一日後に相対的株価変動率が上がるか下がるかの予測を期間と銘柄ごとに行いその平均正答率を用いた。

3.2 実験結果と考察

予測が実際の CPR 法における 1 日後の全 89 銘柄 217 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。

次に、特定の期間に着目し複数テキストの分析結果の統合により，個別銘柄に影響を与える要因と関連銘柄を抽出した。2012 年 9 月 14 日に「2030 年代に原子力発電稼働ゼロを目指す」という新聞記事が発表され、市場に大きな影響を与えたと思われるため、その周辺の期間に着目した。また、原子力関連銘柄の一つである東芝に着目して実験結果の例を示す。

表 2 に、東芝に影響を与えた要因単語リストのうち影響度が高い単語の例を示す。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 4 には、東芝の単語フィルタリング結果の例を示す。表 2 の要因単語リストを見て分かるように、東芝に関係のないと思われる単語も多く含まれることが分かる。表 3 の基本単語リスト生成結果からは、東芝に関連のあると思われる単語が多い一方で、決算短信に多く見られるような単語も多い。表 4 の単語フィルタリング結果からは、「火力」、「原子力」、「電力」などのその時期に影響を与えた可能性の高い原子力発電に関わる単語を抽出できている。一方で、「継続」、「変更」、「組織」など影響の要因となっていないと思われる抽象的な単語も抽出されている。

表 2 2012 年 9 月 6 日～2012 年 10 月 19 日における東芝の要因単語リスト生成結果の例

教典	争い	夕刊	下水道
住所	やりとり	調味	実績
追加	新幹線	西宮	通話
京橋	沼田	包装	戒告
国技	氏名	デベロッパー	運航

表 3 東芝の基本単語リスト生成結果の例

電子デバイス	プロダクツ	デジタル	東芝テック
ウェスチングハウス	事業	医用	四半期
フラッシュ	年度	通算	部門
セミコンダクター	メモリ	ストレージ	LSI
半導体	損益	３月	システム
代表執行役	東芝キヤリア	ランディス・ギア	フラッシュメモリ
当社	家庭	ハードディスク装置	照明
NAND	ディスプレイ	パソコン	インフラ
株	映像	増収	Toshiba
芝浦メカトロニクス	白物家電	メディア	売上

表 4 2012 年 9 月 6 日～2012 年 10 月 19 日における東芝の単語フィルタリング結果の例

半導体	ファ	製品	デバイス
システム	カンパニー	室長	アメリカ
リン	火力	テレビ局	決算
単独	グローバル	リート	継続
照明	家電	ディスプレイ	システムズ
変更	大使館	発注	報酬
家庭	予想	映像	エレクトロニクス
組織	影響	電力	営業
執行	電子	広報	ローン

表 5 には、東芝関連銘柄であるフィルタリングされた単語を上位に含む銘柄とその単語を示す。

表 5 2012 年 9 月 6 日～2012 年 10 月 19 日における東芝関連銘柄の例

半導体	火力	製品
ブリジストン	HOYA	富士通
本田技研		三井住友
東京エレクトロン		横浜銀行
東レ		東レ
		ブリジストン

表 5 の結果からは、要因を基に銘柄が抽出できていることが分かる。「半導体」という要因に対しては、ブリジストンや本田技研といったタイヤメーカーや自動車メーカー、そして半導体メーカーの東京エレクトロンが上位に来ている。このように、原子力発電の関連銘柄の一つだと知っていた場合、東芝という銘柄から、半導体という影響を与える要因、さらにはブリジストンや本田技研といった関連銘柄を知ることができる。また、関連銘柄に対して同様に単語のフィルタリングと関連銘柄を抽出することによって、表 6 に示すように関連銘柄を抽出することができた。このように、個別銘柄に影響を与える要因とその要因を基に探索的に関連銘柄を取得することができた。

表 6 2012 年 9 月 6 日～2012 年 10 月 19 日におけるブリヂストン関連銘柄の例

タイヤ	処理	トラック
国際石油開発帝石	大日本印刷	スズキ
	富士フィルム	

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、個別銘柄に影響を与える要因の抽出、銘柄を構成する基本単語の抽出、そしてそれらを基に関連銘柄を抽出することができた。また、得られた関連銘柄に対して、同様の操作を行うことによって、関連銘柄を再探索することができた。これにより個人投資家は、その時期に気づきにくい関連銘柄へ投資する機会の増加や、銘柄に影響する要因を知ることによって次の投資の機会に活かすことができると考えられる。今後の課題としては、以下の 2 つが挙げられる。

1 つ目に、投資支援システムの開発が挙げられる。最終的に投資支援を行うためには、GUI で投資支援システムを開発することが必要でありその中で検証も進める必要がある。

2 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 , 松井 藤五郎 : 金融テキストマイニングの紹介, 石田 基広, 金 明哲 編,「コーパスとテキストマイニング」, 共立出版, 2012.

[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] 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)
[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
[５] 坂地泰紀，酒井浩之，増山繁，決算短信 PDF からの因果関係抽出に基づく過去事象間の関連表示，第 28 回人工知能学会全国大会, 3L4-OS-26b-2, 2014