

新聞記事内容と株価変動の関連性の定量的分析

張 へい†

† 名古屋大学大学院情報科学研究科

松原 茂樹‡

‡ 名古屋大学情報連携基盤センター

1 はじめに

近年、オンライン株取引の普及に伴い、個人投資家が急増している。機関投資家と異なり、個人投資家が利用できる情報の種類には限界があり、そのような状況下では新聞記事は投資判断における重要な情報源となる。しかし、個人投資家が関連するすべての記事を参照することは容易ではない。関心のある銘柄について、その株価変動に関連する記事を選択的に抽出し提供できれば、個人投資家にとって有用性が高い。

本稿では、新聞記事内容と株価変動との関連性の分析について述べる。記事を構成する言語表現に基づいて算出した新聞記事の評価値と、実際の株価データとを比較した。209,553 記事 (3 年分) に出現した表現を用いて 73,333 記事 (1 年分) を評価した。株価変動データと比較した結果、両者の間の相関性を確認することができた。

2 関連研究

株価データと新聞記事の関連性については、実データを用いた分析及び実験がいくつか行われている。小川らは、新聞記事をテーマ別に分類し、株価変動との関連性の高いテーマを判定する実験を実施している [1]。また、Koppel ら、ならびに、Mittermayer は、ある企業に関連する記事内容が、その株価の上昇要因または下降要因いずれであるかで、記事を分類する手法を提案している [2, 3]。さらに、Lavernko らは、Web 上のニュース記事を用いて株価変動を予測するシステム AAnalyst を開発している [4]。

しかしながら、これらはいずれも、記事内容の特徴付けを単語に基づいて実行しており、記事の内容を十分に捉えられていない。また、実験については、限定された銘柄に関連する少数の記事を対象としているに過ぎず、結果の汎用性を示すには不十分である。一方、本研究では、構文的関係に基づいて記事内容の特徴付けを実施している。また、2400 以上の銘柄を対象に、73,333 記事の評価を行うものであり、規模の点でも優位性が高い。

3 株価データを用いた新聞記事の評価

新聞記事内容を分析し、株価変動への影響を評価する。評価の流れを図 1 に示す。本研究では、記事テキストを構成単位に分割し、各単位の評価値の総計から新聞記事の評価値を算出する。

3.1 過去の記事の評価

まず、ある銘柄 s の株価変動率を定義する。銘柄 s の日付 i における株価を $p_i(s)$ とする。このとき、日付 i から単位期

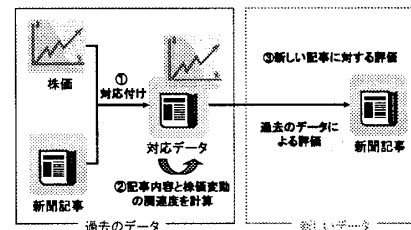


図 1: 新聞記事の評価手法の概要

間 $k(\geq 1)$ 日後の株価変動率 $r_{i,k}(s)$ を、

$$r_{i,k}(s) = \frac{p_{i+k}(s) - p_i(s)}{p_i(s)} - R_{i,k} \quad (1)$$

で計算する。ただし、 $R_{i,k}$ は基準株価変動率であり、市場全体の変動状況を表す指標が用いられる。

各銘柄の株価変動率から、過去の新聞記事の評価値 (以下、記事評価値) を算出する。記事評価値には、株式市場における投資家による評価、すなわち、記事に出現する銘柄の株価変動率を使用する。日付 i の大引け後に発行された新聞記事 A_i に、 n 個の銘柄名 $s_1, \dots, s_n$ が出現するとき、記事 A_i の評価値 $value(A_i)$ を、

$$value(A_i) = \frac{1}{n} \sum_{j=1}^n p(s_j) \quad (2)$$

で計算する。

3.2 新しい記事の評価

過去の新聞記事に対する記事評価値を用いて、新しい記事の評価値を推定する。推定には、記事を構成する言語単位の評価値 (以下、単位評価値) を使用する。

ある言語単位 u の評価値 $value(u)$ を、 u が出現した過去の全ての記事の評価値の平均とする。言語単位 u が n 個の記事 $A_1, \dots, A_n$ に出現するとき、 u の単位評価値を

$$value(u) = \frac{1}{n} \sum_{j=1}^n value(A_j) \quad (3)$$

で算出する。

新しい記事 A の評価値 $value'(A)$ は、記事 A を構成する言語単位の評価値の平均で求める。すなわち、 n 個の言語単位 $u_1, \dots, u_n$ から構成される新しい記事 A の評価値を、

$$value'(A) = \frac{1}{n} \sum_{j=1}^n value(u_j) \quad (4)$$

により推定する。

3.3 新聞記事の構成要素単位

記事を構成する言語単位としていくつか考えられるが、本研究では依存関係 (dependency) を使用する。依存関係とは、



図 2: 依存関係の例

修飾・被修飾など、文構成素間の意味的關係を表現する単位の一つであり、日本語の場合には、係り受け関係と呼ばれることがある。本研究では、依存関係として以下の2種類を採用した。

- 文節間依存関係：日本語係り受け構造における係り文節と受け文節の対。
- 単語間依存関係：日本語係り受け構造における係り文節と受け文節における主辞 (head) となる単語対、あるいは、各文節内で依存関係にある単語対。

ただし、文節内の依存関係については、隣接する単語対を抽出した。

言語表現「ベビーフード最大手の和光堂を」から抽出される文節間、及び、単語間依存関係を図2に示す。矢印の始端と終端の対が依存関係を表している。ただし本研究では、単語間依存関係の対象を、助詞、助動詞、記号以外とし、動詞については原形を用いた。また、依存関係の種類を、係り文節内の助詞の種類 (は、が、の、を、に、へ) により定め、依存関係に導入した。すなわち、依存関係を、「依存元」「依存先」「依存の種類」の3項組とした。

4 実験

4.1 実験の概要

新聞記事と株価変動の関連性を調べるために、実データを用いた分析実験を実施した。実験は、2003年から2005年までの3年間の209,553記事ならびに株価データを用いて、2006年の1年間の73,333記事の評価値を推定し、それを実際の株価変動データと照合することにより行った。

新聞記事として、日本経済新聞社が発行する「日本経済新聞」「日経金融新聞」「日経産業新聞」の3誌を使用した。株価データとして、東京証券取引所に上場している全銘柄 (東証1部、東証2部、東証マザーズ) を対象とした。2006年には、2,416銘柄が上場していた。

銘柄の株価変動率の算出では、単位期間を1日 (すなわち、 $k=1$) とし、株価には各日の終値を用いた。基準株価には、日経平均株価を使用した (式1参照)。新聞記事と株価変動率の対応付けについては、「日本経済新聞」の朝刊、「日経金融新聞」「日経産業新聞」に対しては発刊当日の株価変動率を、「日本経済新聞」の夕刊に対しては発刊翌日の株価変動率を使用した。

実験における評価値推定のための言語単位として、文節間、及び、単語間依存関係に加え、単語を言語単位とする方法も比較のために実行した。なお、形態素解析には chasen を、係り受け解析には cabocha を使用した。

4.2 実験結果

本手法により算出した記事推定値と、記事に出現する銘柄の株価変動率との相関を調べた。すなわち、2006年の全記事 (73,333記事) を、記事推定値の昇順ソートにより、100グループに等分割し (グループ1が最も推定値の低い記事、グ

表 1: スピアマンの順位相関係数

構成単位	単語	文節間依存関係ペア	単語間依存関係ペア
相関係数	0.52	0.59	0.71

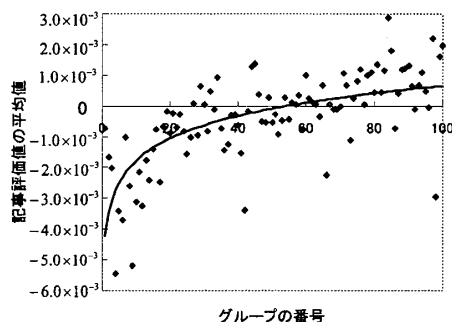


図 3: 記事スコアと記事の評価値の比較

ループ100は最も高い記事から構成)、記事グループとその評価値平均との間の順位相関係数を算出した。

表1に、使用した言語単位ごとの順位相関係数を示す。いずれを使用した場合においても、両者の間に相関が存在することがわかった。また、単語を単独で用いる場合よりも、依存関係を使用した場合の方が相関が高く、特に、単語間依存関係を用いた場合に最も高く、相関係数は0.71となった。この場合における、記事グループ番号とグループごとの記事評価値の平均をプロットしたグラフを図3に示す。このグラフからも両者の相関性を確認することができる。

5 おわりに

本稿では、新聞記事内容と株価変動との関連性の分析について述べた。記事を構成する言語表現に基づいて算出した新聞記事の評価値と、実際の株価データとを比較した。3年分の記事に出現した言語表現を用いて1年分の記事を評価し、株価変動データと比較した結果、両者の間の相関性を確認することができた。今後は、記事内容をより特徴付けるために、記事を構成する言語表現の種類について検討する予定である。

参考文献

- [1] 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, 情報処理研究報告, NL142-19, 137-144, 2001.
- [2] Koppel, M. and Shtrimberg, I.: Good News or Bad News? Let the Market Decide, AAAI Spring Symposium on Exploring Attitude and Affect in Text, 86-88, 2004.
- [3] Mittermayer, M. A. Forecasting Intraday Stock Price Trends with Text Mining Techniques. Proceedings of the 37th Hawaii International Conference on System Sciences, 64-73, 2004.
- [4] Lavernko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J.: Mining of Concurrent Text and Time Series, Proceedings of KDD 2000 Conference Text Mining Workshop, 37-44, 2001.