業種別による新聞記事と株価動向の関係の解析

廣川 敬真 † 吉田 稔 ††　山田 剛一 † 増田 英孝 † 中川 裕志 ††
† 東京電機大学 未来科学部　†† 東京大学 情報基盤センター

1 はじめに

株価動向予測におけるテクニカル分析及びファンダメンタルズ分析は、数学的な予測による分析が多い [1]。新聞記事には、”上昇”、”下落”、”発言”、”発売”、”中止”、”リコール”、”リストラ”、など、株価動向に対して影響を及ぼすと考えられる情報が数多く含まれているが、これらの表現を数学的な処理に直接利用できないため、新聞記事のテキスト情報は、株価動向の予測にはあまり活用されていない。

そこで本研究では、ファンダメンタルズ分析を基に、新聞記事のテキスト情報と株価動向との対応付けを行う。新聞記事が株価動向へ及ぼす影響を調べるため、従来研究 [2] で行われていた手法を用い、株価変動率を用いて記事評価値を算出する。そして、語句が株価に与える影響を調べるために、記事評価値を用いて語句評価値を算出し、解析を行う。また、業種によって同じ単語であっても、株価動向に与える影響が異なることも十分に考えられる。

そのため、本研究では、新聞記事を業種別に分類し解析を行う。業種別に分類することにより、業種によって単語が株価動向に与える影響の違いや新聞記事と株価動向の関係性について解析する。業種別に分類した語句評価値を用いて、記事推定値を算出し、記事評価値と記事推定値について評価を行った。

2 関連研究

現在、新聞記事のテキスト情報を利用した研究として、杉浦らのテキスト情報から統計量名と統計量を抽出するもの [3] や小川らのテキスト情報をマイニングすることで新聞記事を各テーマに分類し、どのテーマが出現したときに株価が上昇・下落するかを解析するもの [4]、張らの株価データを用いて新聞記事と株価変動の相関を評価する方法 [2] がある。

杉浦ら, 小川らの研究はテキスト情報の一部を対象として抽出や解析を行っているため情報量としては不十分である可能性が考えられる。また、張らの研究では、すべての銘柄およびテキスト情報全体を対象としているが、語句評価値を算出する際に、業種で分けずにすべてを対象として行っている。ある単語によっては、業種によって株価動向に与える影響が異なり、ノイズとなってしまう可能性も考えられる。

3 新聞記事と株価動向の関係の解析方法

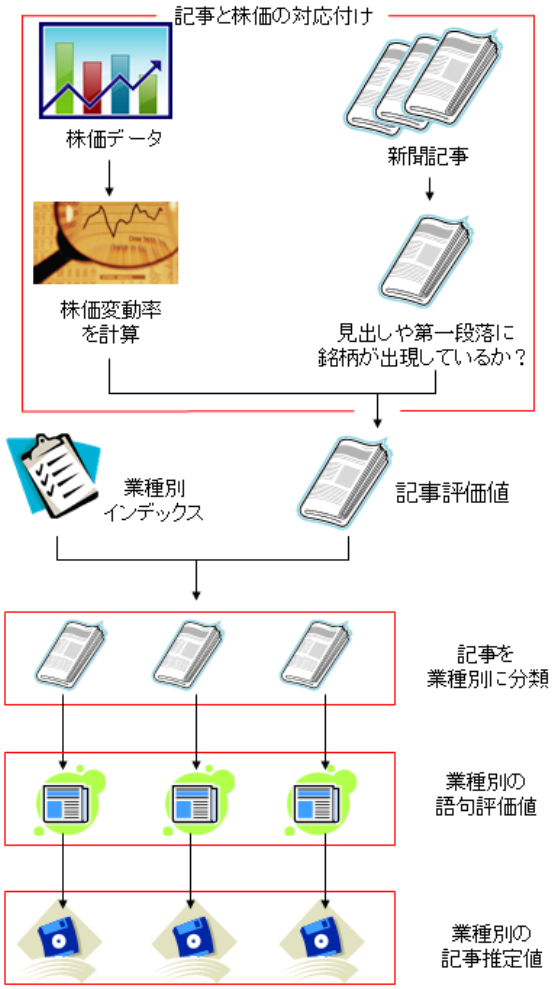

図 1: 研究の全体像

図 1 は、今回の研究の全体像である。株価変動率、記事評価値、語句評価値、記事推定値の算出方法については、張らの研究のアプローチを利用している。張

Analysis of Relationship between Stock Trends and News Articles Categorized by Industry
† Takamasa HIROKAWA(hirokawa@cdl.im.dendai.ac.jp)
†† Minoru YOSHIDA(mino@r.dl.itc.u-tokyo.ac.jp)
† Kouichi YAMADA(yamada@im.dendai.ac.jp)
† Hidetaka MASUDA(masuda@im.dendai.ac.jp)
†† Hiroshi NAKAGAWA(nakagawa@dl.itc.u-tokyo.ac.jp)
Tokyo Denki University (†)
University of Tokyo (††)

らの研究では、記事評価値を算出した後、業種ごとに分類せず、すべての記事を対象として語句評価値を算出し、記事推定値を算出しているが、我々の研究では、業種別に語句評価値を算出し、それぞれの業種において記事推定値を算出する。

3.1 株価変動率

新聞記事と株価動向の関係性を調べるためには、新聞記事と株価動向を対応付けする必要がある。そのため、株価変動率 (式 1) を用いる。

$$r_{ij}(s) = \frac{p_j(s) - p_i(s)}{p_i(s)} \tag{1}$$

式中の記号は、i, j:日付, s:銘柄, $p_i(s)$:日付 i における銘柄 s の株価, $r_{ij}(s)$:日付 i から j における株価変動率を表す。

3.2 市場の影響を考慮した株価変動率

株価変動率だけを用いるのでは、市場全体の動きに影響を受けて上昇または下落してしまうことが考えられ、市場全体の動きの影響を考慮する必要がある。そのため、市場を考慮した株価変動率 (式 2) を用いる。

$$r'_{ij}(s) = r_{ij}(s) - R_{ij} \tag{2}$$

式中の記号は、R_{ij}:日経平均株価の変動率, $r'_{ij}(s)$:市場を考慮した株価変動率を表す。

3.3 記事評価値

市場を考慮した株価変動率を用いて、新聞記事における株価動向への影響力を調べるため、記事評価値を算出する。記事評価値は、新聞記事 a において、見出しと第一段落に出現するすべての銘柄名の市場を考慮した株価変動率 (式 2) の平均値とする。

$$article(a) = \frac{1}{l}\sum_{k=1}^{l} r'_{ij}(s_k) \tag{3}$$

式中の記号は、l:銘柄数, a:新聞記事, $article(a)$:新聞記事 a における記事評価値を表す。

3.4 記事を業種別に分類

新聞記事を業種別に分類するために、業種と銘柄の対応表を用意した。新聞記事に業種情報を付与する方法として、新聞記事の見出しと第一段落にある銘柄が出現した際に、その銘柄が属する業種をその新聞記事が属する業種とする。また、業種の異なる複数の銘柄が出現した際は、一番多く出現した業種をその新聞記事に付与する。

3.5 業種別語句評価値

記事評価値を用いて、語句における株価動向への影響を調べるため、語句評価値を算出する。語句評価値は、その語句が存在する対象業種のすべての記事における記事評価値の平均値とする。また、語句評価値は業種別に算出する。

$$word(d) = \frac{1}{m}\sum_{k=1}^{m} article(a_k) \tag{4}$$

式中の記号は、d:語句, m:語句 d が出現する記事数, $word(d)$:語句 d における語句評価値を表す。

3.6 業種別記事推定値

業種別語句評価値を用いて、新しい記事へ適用し記事推定値を算出する。記事推定値は、新聞記事を構成する複数語句の語句評価値の平均値とする。また、語句評価値と同様に記事推定値も業種別に算出する。

$$newarticle(a) = \frac{1}{n}\sum_{k=1}^{n} word(d_k) \tag{5}$$

式中の記号は、d:語句, n:記事 a に出現する単語数, $newarticle(a)$:記事 a における記事推定値を表す。

3.7 単語対への拡張

以上の計算を、新聞記事を対象に行う。単語の単純な出現だけに着目したのでは、否定表現などを抽出することが出来ず、構文情報を考慮する必要がある。

そのため、張らのアプローチである単語対を利用し、構文解析器である CaboCha[5] を用いて、単語の係り受け情報を抽出し、同様に評価値を算出する。

- 単語対…単語の係り元と係り先を 1 つの情報としてまとめる
- 助詞を含む単語対…単語対において、単語の係り元と係り先の間に助詞が出現した際は、助詞の情報も含めて 1 つの情報としてまとめる

単語対に用いる単語の品詞の種類として、助動詞以外を用い、動詞については原形を用いる。また、助詞については、張らと同様に「は、が、の、を、に、へ」を用いる。

4 評価

4.1 実験環境

本研究では、日本経済新聞 98 年,99 年版 [6] を対象とし、ニュース記事を抽出する。株価データは、Pan-Rolling の相場データ集・国内相場版 [7] を使用する。新聞記事データベースは、記事番号、見出し、日付、第一段落、全文を格納する。また、株価データベースは、銘柄コード、日付、四本値 (始値、高値、安値、終値)、出来高を格納する。

4.2 実験方法

今回の研究の実験として、2 つの方法で実験を行う。(1) 業種別に分類することにより業種間の単語対に違いが見られるか、(2) 業種毎の記事推定値を用いて記事評価値の単純な上昇・下落を予測できるかを評価する。記事推定値は、98 年の新聞記事と株価データを用いて学習データを構成し、99 年の記事へ適用したものを用いる。

株価変動率において、新聞記事が掲載された日にちとその翌日の差分を取った 1 日後差分、新聞記事が掲載された日にちとその 5 日後との差分を取った 5 日後差分の 2 つにおいて評価を行う。(1) の実験では、単語対に対して「円」と「高」、「円」と「安」が含まれるものについて行う。その結果が、表 1 と表 2 である。

(2) の実験では、株価動向の小さな変動はノイズと考えられるため、記事推定値においてしきい値を設け、+0.0005 以上、または、-0.0005 以下のものだけを今回の実験の評価対象とする。

表 3、表 4 は、それぞれの業種についての正解率である。表中の数値は、分母が総記事数を表し、分子が正解した記事数を表す。

表 5 から表 14 は、それぞれの業種における 1 日後差分と 5 日後差分における予測の正解の記事数・不正解の記事数を表したものである。推定値上昇と推定値下落は、記事推定値が上昇・下落と予測したものを表し、評価値上昇と評価値下落は、記事評価値が上昇・下落したものを表す。表中の数値は、正解・不正解した記事数を表す。

4.3 業種別に分類した結果

表 1 より、1 日後差分ではあまり変化がなく参考にならない結果となった。表 2 より、5 日後差分については、「電機・精密」、「自動車」に関して、「円」、「高」を含む単語対の値よりも「円」、「安」を含む単語対の値が大きい結果となった。また、「小売り」、「食品」に関しては、「円」、「安」を含む単語対の値よりも「円」、「高」を含む単語対の値が大きい結果となった。これら結果により、同じ単語であっても業種により株価動向に対する影響が異なることが分かる。

表 1: 単語対について (1 日後差分)

業種	1 日後差分 円高	1 日後差分 円安
電機・精密	0.011	0.010
自動車	0.010	0.012
小売り	0.014	0.016
食品	0.006	0.002

表 2: 単語対について (5 日後差分)

業種	5 日後差分 円高	5 日後差分 円安
電機・精密	-0.010	0.008
自動車	0.013	0.041
小売り	0.020	0.012
食品	0.022	0.005

4.4 予測結果

表 3: 正解率 (1 日差分)

業種	記事数/総記事数	正解率
未分類	12793/26425	0.4841
電機・精密	2083/4060	0.5131
自動車	925/1940	0.4768
小売り	762/1634	0.4663
食品	809/1516	0.5336

表 3 より「未分類」と比較して「電機・精密」と「食品」については、正解率が向上しており、「自動車」と「小売り」はあまり変化が見られない結果となった。また、表 4 より、「未分類」と比較して、「電機・精密」に関しては、正解率が向上する結果となり、「自動車」はあまり変化が見られず、「小売り」と「食品」に関しては下がる結果となった。これら結果より、語句を業種別に分類することにより、1 日後差分では、分類精度が向上する業種や変化が見られない業種が存在することを確認し、5 日後差分では、分類精度が向上する業種・変化が見られない業種・下がる業種が存在することを確認した。

5 おわりに

業種別に分類した新聞記事と株価動向の予測を行った。単語対については、業種間において、株価動向に与える影響が異なることが分かった。また、予測に関しては、業種別に分類しない方法と業種別に分類した

表 4: 正解率 (5 日差分)

業種	記事数/総記事数	正解率
未分類	12437/26056	0.4773
電機・精密	2083/4060	0.5123
自動車	1064/2226	0.4780
小売り	854/1946	0.4388
食品	835/1861	0.4487

表 5: 未分類 (1 日差分)

	推定値上昇	推定値下落
評価値上昇	8115	4013
評価値下落	9219	4678

方法を比較した結果、分類精度が上がる業種・変化が見られない業種・下がる業種があることが分かった。また、今回の実験では出現した単語対について重み付けを行っていない。そのため、今後の展望として、単語対に対して TFIDF などの重み付けを行う方法が考えられる。

参考文献

[1] 月本 洋：実践データマイニング金融・競馬予測の科学、オーム社 (1999)

[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] 杉浦 隆博、吉田 稔、山田 剛一、増田 英孝、中川 裕志：新聞記事の数値による情報検索システムの提案と実装、JSAI2007-2H5-9(2007).

[4] 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 自然言語処理研究会研究報告 Vol.2001 No.20 pp.137-144 (2001).

[5] CaboCha http://chasen.org/ taku/software/cabocha/

[6] 日本経済新聞社、CD-日本経済新聞 98 年,99 年版

[7] PanRolling http://www.panrolling.com/

表 6: 電機・精密 (1 日差分)

	推定値上昇	推定値下落
評価値上昇	1589	465
評価値下落	1512	494

表 7: 自動車 (1 日差分)

	推定値上昇	推定値下落
評価値上昇	590	293
評価値下落	722	335

表 8: 小売り (1 日差分)

	推定値上昇	推定値下落
評価値上昇	417	245
評価値下落	627	345

表 9: 食品 (1 日差分)

	推定値上昇	推定値下落
評価値上昇	396	302
評価値下落	405	413

表 10: 未分類 (5 日差分)

	推定値上昇	推定値下落
評価値上昇	9347	2572
評価値下落	11047	3090

表 11: 電機・精密 (5 日差分)

	推定値上昇	推定値下落
評価値上昇	1552	927
評価値下落	1298	785

表 12: 自動車 (5 日差分)

	推定値上昇	推定値下落
評価値上昇	684	342
評価値下落	820	380

表 13: 小売り (5 日差分)

	推定値上昇	推定値下落
評価値上昇	621	160
評価値下落	932	233

表 14: 食品 (5 日差分)

	推定値上昇	推定値下落
評価値上昇	670	113
評価値下落	913	165