新聞記事内容と株価変動との関連性の分析

張　へい†　　松原　茂樹‡

†名古屋大学大学院情報科学研究科
‡名古屋大学情報連携基盤センター

Analysis of Relevance between News Articles and Stock Price Change

He Zhang†　Shigeki Matsubara‡

†Graduate School of Information Science, Nagoya University, Japan
‡Information Technology Center, Nagoya University, Japan

1　はじめに

近年，オンライン株取引の普及に伴い，個人で株取引を行う投資家が増えてきた．しかし，このような個人投資家の多くは機関投資家と異なり，投資判断に必要な情報を十分に収集することが困難である．株取引に有益な情報を適時選択的に提供することが出来れば，個人投資家はより的確に投資判断を下すことが可能となる．例えば，株価の上昇を見込んで株式を買い増したり，逆に下降を予期して株式を売却し，損失を回避したりするといった投資行動がこれまでより容易となる．

投資家が株式投資をする際，最も用いられる情報源の 1 つに新聞が挙げられる．新聞は速報性・網羅性を兼ね備えた点で優れたメディアである．近年，電子化された新聞記事がインターネットから利用可能となり，速報性はさらに向上している．このように随時提供される新聞記事の中から，特定銘柄の株価変動に関連する情報を選択的に抽出出来れば，個人投資家にとって有用な情報となることが期待される．しかし，特定銘柄の株価変動に関連する新聞記事を機械的に抽出する手法は確立されていない．

そこで本稿では，新聞記事と特定銘柄の株価変動との関連性を評価することを試みた．本研究では，過去の新聞記事データと株価データを用いて，両者の関連性を分析する．銘柄名が出現する新聞記事と新聞発売日当日における銘柄の株価変動を対応付け，記事の内容と株価変動との関連性を分析する．

また，分析結果を基にして，比較的最近の記事を評価し，評価結果を過去のデータによる評価とする．この時，この記事と実際に起こった株価変動との関連度も評価し，実際の市場による評価とする．両者の評価間の相関の有無から，記事の内容と株価変動の関連性を明らかにする．

実験では，2003 年から 2006 年までの過去 4 年分の日本経済新聞の記事データと株価データを用いた．まず 2003 年から 2005 年の 3 年分の記事に含まれる単語と株価変動の関連度を計算した．この計算値を用いて 2006 年 1 年分のデータを評価し，実際に起こった株価変動との相関関係を求めた．結果，記事内容と株価変動の間に関連性が認められた．また，特定銘柄の株価変動と関連する新聞記事を抽出することが可能であることが明らかとなった．

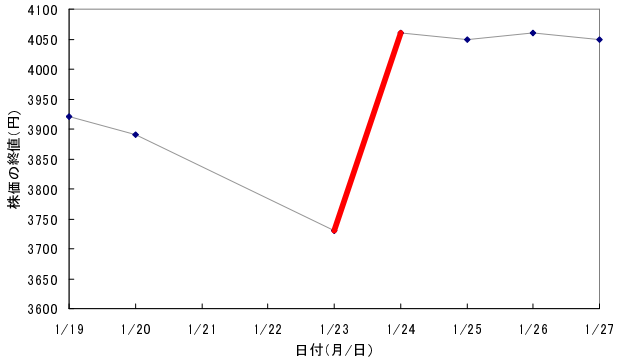


図 1: 2006 年 1 月下旬の小林製薬株の株価変動

2　記事内容と株価変動

株価は投資家による株式の売買によって変動するものである．投資家が投資行動を行う場合，様々なメディアから得た情報を参考にすることが多い．中でも新聞は速報性と網羅性に優れるメディアとして投資家から重要視されている．このことから，新聞記事の内容と投資家の投資行動の間には何らかの関連性が示唆される．これは新聞記事と株価変動の間における関連性と言い換え可能である．

また別の見方をすれば，株価は企業価値を示す指標である．企業価値に影響する要因としては，財務状況や，事業活動の動向，社会情勢などが挙げられる．新聞はこれらの情報を網羅的に提供するメディアである．このような理由からも新聞記事の内容と株価変動の間には関連性の存在が考えられる．

例として，東証 1 部上場の小林製薬株の 2006 年 1 月下旬の株価変動を図 1 に示す．縦軸は各日における株価の終値を表す．1 月 23 日の終値と比べ，1 月 24 日の終値が急上昇したことがわかる．一方，1 月 24 日の日本経済新聞の朝刊に掲載されていた小林製薬の関連記事を図 2 に示す．記事には小林製薬株の株価を上昇させる内容が書かれている．そのため，この記事と株価の上昇の間には関連性の存在が示唆される．

2006年1月24日　本紙朝刊
小林製薬4―12月、**経常益**9%**増**

小林製薬が二十三日発表した二〇〇五年四―十二月期の連結業績は経常**利益**が前年同期比九%**増**の百三十二億円だった。四月に実施した卸事業での事業譲り受けや、厳冬の影響によるカイロの販売**増加**などで**売上高**が**伸びた**。国内工場で消臭芳香剤など家庭用品の製造コスト削減が**進んだ**ことも寄与した。
売上高は一七%増の千九百二十三億円。昨秋発売した点眼薬やマッサージソックス「ムクミキュア」などの新商品が**好調**だった。主力の消臭芳香剤も**伸びた**。**純利益**は一一%増の六十九億円だった。

（日本経済新聞）

図 2: 2006 年 1 月 24 日の小林製薬の関連記事

記事には”利益”，”増加”，”売上高”，”進む”，”好調”などといった株価上昇と関連する単語 (下線を引いた単語) が多く出現している．記事に含まれる全ての単語が株価上昇と関連する単語ではないが，株価上昇と関連する単語が多く出現しているため，この記事が株価を上昇させる内容であると判断される．このように記事内容の分析では，記事に含まれる単語に着目することが 1 つの方法であると考えられる．

そこで本研究では，記事の内容と株価変動との関連性を分析する際，記事に出現している単語に着目し，これらの単語と株価変動との関連度を計算した．

3　関連研究

新聞記事と株価変動の関連性についての研究はすでに幾つか行われている．Koppel らは，企業に関する記事について，株価を上昇させる内容か，下降させる内容かを分類する手法を提案している [1]．小川らは，分類システムにより新聞記事にテーマ情報を付与し，株価変動と関連のあるテーマを抽出することを提案している [2]．酒井らは，企業業績に影響を与えるインパクトの大きい記事を判定する手法を提案している [3]．Lavrenko らは，企業に関する記事が，その後の株価に与える影響を推定する手法を提案している [4]．

文献 [1],[2],[3] では，株価変動や企業業績と関連する記事を判定しているが，特定の記事と銘柄の株価変動に関連する記事を見つける仕組みについては明らかにしていない．一方，本研究では，特定の銘柄の株価変動に対しても関連する記事を見つけ出せることを目指している．また文献 [4] では，与えられる記事が株価に変化を与えるほどのインパクトのある記事であることを前提としている．それに対し，本研究は発信される膨大な量の記事から株価変動と関連する記事を見つけ出すことを目指しているため，対象とする新聞記事は企業名が出現している全ての記事とする．

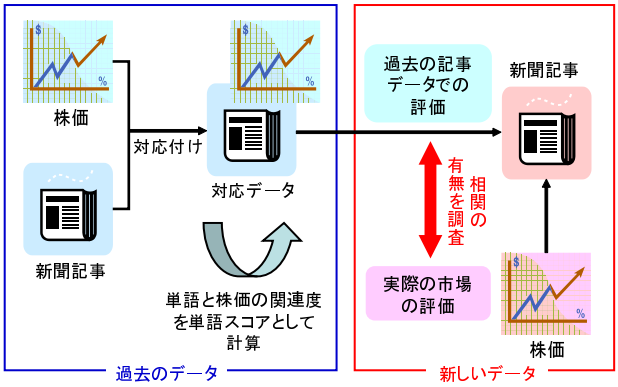


図 3: 新聞記事の評価手法の概要

4　新聞記事の評価手法

我々は新聞記事の内容を分析し，記事と特定銘柄の株価変動との関連性を評価する．記事の内容を分析する際，記事に含まれる単語に着目する．評価手法を図 3 に示す．

まず，過去の新聞記事データと当時実際に起こった株価変動を対応付けて，対応データとする．対応データの記事に含まれる各単語に着目し，単語と株価変動の関連度を単語スコアとして求める．次に，単語スコアを用いて，新しい新聞記事を評価する．この評価は過去の記事データでの評価とする．また，実際の市場の評価として，実際に起こった株価変動との関連度も評価する．最後に，過去の記事データでの評価と実際の市場の評価の両者間の相関の有無から，記事に含まれる単語と株価変動の関連性を明らかにする．

4.1　新聞記事の評価値

ある銘柄の株価変動を表すものとして，株価変動率を用いる．各銘柄の株価データに対して，株価変動率を計算する．株価変動率の計算では単位期間あたりの株価変動の比 r_i を用いる (式 (1) 参照)．

$$r_i = \frac{p_2 - p_1}{p_1} \tag{1}$$

p_2 はある日の株価，p_1 はその日の単位期間前の株価である．

また，ある銘柄の株価変動は同業他社の銘柄あるいは市場全体と同じような動きを表す場合があるため，その銘柄だけに注目するだけでは必ずしも十分ではない．そこで，複数の銘柄の比較による相対的株価変動率を用いる．銘柄 B に対する銘柄 A の株価変動率を計算するには，銘柄 A の株価変動率と銘柄 B の株価変動率との相対的な値を用いる．相対的な値としては，2 つの銘柄の株価変動の差 $r_{A/B}$ を用いる (式 (2) 参照)．

$$r_{A/B} = r_A - r_B \tag{2}$$

ここで，r_A は銘柄 A の株価変動率，r_B は銘柄 B の株価変動率である．

株価データ

銘柄名	日付	終値(円)	前日の終値(円)	変動率
カネボウ	2004年1月7日	114	112	1.8%
小林製薬	2006年1月24日	4,060	3,730	8.8%

記事1

2004年1月7日　本紙朝刊
スーパー系カード会社、6社共同で販促――まず化粧品、ポイント倍増など

ーーーーー、カネボウ、ーーーーーーーーーーサポートになるとともに手数料増加につながるため、ーーーーーーーーーー

記事2

2006年1月24日　本紙朝刊
小林製薬、経常利益9%増

ーーー受けや、厳冬の影響によるカイロの販売増加などで売上高が伸びた。国内工場で消臭芳香剤など家庭用品の製造コスト削減が進んだこともーー

対応付けの結果

記事	記事1	記事2
記事の評価値	0.018	0.088

図 4: 新聞記事と株価変動率の対応付け

新聞記事と株価変動率との対応付けでは，株価データに含まれる全ての銘柄に対する関連記事をピックアップし，新聞記事の発売日当日における銘柄の株価変動率と対応付ける．関連記事は，タイトルもしくは記事本文に銘柄名が出現した全ての記事とする．銘柄の株価変動率と対応する記事の評価値を t_i とする．ここで，記事の評価値 t_i とは，記事と実際の株価変動との関連度である．

カネボウ株及び，小林製薬株における新聞記事と株価変動率の対応付けの例を図 4 に示した．記事 1 と記事 2 の中にそれぞれカネボウと小林製薬の銘柄名が出現しているため，新聞発売日当日の株価変動率と対応付ける．カネボウ株と小林製薬株の株価変動率をそれぞれ記事 1 と記事 2 の評価値とする．

ただし，1 つの記事に複数の銘柄名が出現した場合は，記事の評価値 t_i をそれらの銘柄の株価変動率の平均値で計算する．図 5 に 1 つの記事に複数の銘柄名が出現した場合の記事の評価値の計算例を示した．記事 1 の中に，資生堂，カネボウ，花王やコーセー化粧品の 4 つの銘柄名が出現しているため，記事 1 の評価値を計算する際，4 つの銘柄における変動率の平均値を記事 1 の評価値とする．

4.2 単語スコア

新聞記事と株価変動の対応データを用いて単語スコアを求める．日本語形態素解析システム $ChaSen$[5] を用いて，対応データの新聞記事に対して形態素解析を行う．記事に含まれる全ての単語に品詞情報を付与する．語彙的に意味を持つ単語として，品詞が名詞，形容詞，副詞，動詞，未知語の単語を対象とする．これらの単語に対して，単語のスコアを求める．

w_i の単語スコアとしては，対応データの新聞記事において，w_i が出現していた全ての記事における記事の評価値の平均値 b_j を用いる (式 (3) 参照)．

株価データ

銘柄名	日付	変動率
資生堂	2004年1月7日	-0.8%
カネボウ	2004年1月7日	1.8%
花王	2004年1月7日	3.8%
コーセー化粧品	2004年1月7日	-1.2%

記事1

2004年1月7日　本紙朝刊
スーパー系カード会社、6社共同で販促―まず化粧品、ポイント倍増など

ーーーー資生堂、カネボウ、花王、コーセー化粧品、ーーーサポートになるとともに手数料増加につながるため、ーーーーーーーーーー

$$\text{記事 1 の評価値} = \frac{(-0.008\%)+0.018\%+0.038\%+(-0.012\%)}{4}\times 10^{-2} = 0.009$$

図 5: 複数の銘柄名が出現した記事の評価値

$$b_j = \sum_{i=1}^{n} \frac{t_i}{n} \tag{3}$$

t_i は w_i が出現した各記事における記事の評価値，n は w_i が出現した記事の数である．図 6 に”増加”の単語スコアの計算をする例を示す．

4.3 新聞記事スコア

単語スコアの結果を用いて，新しい記事に対し，記事スコアを求める．記事スコアは新しい記事に対する過去のデータでの評価である．

a_q の記事スコアは，a_q に出現する全ての単語における単語スコア平均値 B_q で求める (式 (4) 参照)．

$$B_q = \sum_{j=1}^{m} \frac{b_j}{m} \tag{4}$$

b_j は a_q に出現した各単語の単語スコア，m は単語の数である．

図 7 に記事 3 の記事スコアを求める例を示す．

5 実験

5.1 実験の概要

新聞記事と株価変動の間に関連性の有無を確認するため，実験を行った．実験には，2003 年から 2006 年までの過去 4 年分の記事データと株価データを用いた．

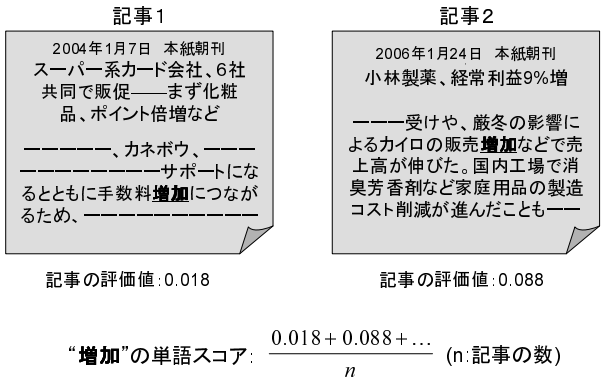
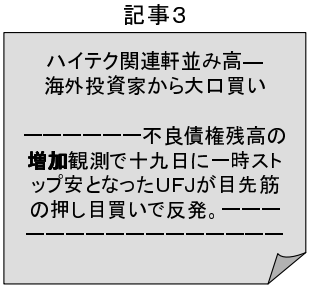


図 6: 過去の記事データを用いた単語スコアの計算



記事3の記事スコア: $\frac{\text{単語1のスコア}+\text{単語2のスコア}+\ldots}{\text{単語の数}}$

図 7: 新しい記事に対する記事スコアの計算

まず，2003 年から 2005 年の 3 年分の記事に含まれる単語の単語スコアを計算した．次に，単語スコアの結果を用いて 2006 年 1 年分の記事に対する記事スコアを求めた．これと同時に，2006 年の株価データから実際に起こった株価変動を算出し，記事の評価値を求めた．最後に，記事スコアと記事の評価値との相関関係を分析した．

結果，記事内容と株価変動との間に関連性が認められた．また，特定銘柄の株価変動と関連する新聞記事を抽出することが可能であることが明らかとなった．

5.1.1 実験データ

実験で用いた記事データは，日本経済新聞社が発行する 3 誌「日本経済新聞」,「日経金融新聞」及び「日経産業新聞」の 2003 年から 2006 年までの 4 年分の新聞記事である．記事数は 282,886 である．年別の記事数を表 1 に示す．

株価データとは，2003 年から 2006 年までの 4 年間における東京証券取引所 (東証 1 部，東証 2 部，東証マザーズ) に上場している全銘柄の株価データベースである．株価データには各銘柄における株価の情報 (1 日の始値，高値，低値や終値など) が記載されている．年別の銘柄数を表 2 に示す．

表 1: 各年の記事数

年	2003 年	2004 年	2005 年	2006 年
記事数	66,986	71,305	71,262	73,333

表 2: 東京証券取引所における各年の銘柄数

年	2003 年	2004 年	2005 年	2006 年
銘柄数	2,206	2,306	2,351	2,416

5.1.2 株価変動率と記事の対応付け

株価変動率を計算する際，単位期間を 1 日とし，株価は終値を用いる．すなわち，式 (1) では，p_1 は銘柄の前日の終値，p_2 は銘柄の当日の終値となる．また，式 (2) で相対的株価変動率を計算する際には，市場の動きを表すものとして，日経平均株価に対する各銘柄の相対的株価変動率を用いる．

株価変動率の計算結果として，株価変動率 $r_i > 0$ のデータの数は 215,628(38.59%)，$r_i < 0$ のデータの数は 343,029(61.39%)，$r_i = 0$ のデータの数は 106(0.02%) だった．株価変動率の計算結果が負になったデータの割合が高いことから，それに基づいて求めた記事の評価値についても，負である記事の割合が高くなった．

株価変動率と記事の対応を行う際， 記事が掲載された新聞の発売時刻による場合分けを行った．すなわち,「日本経済新聞」の朝刊,「日経金融新聞」,「日経産業新聞」は，朝に発売されるため，これらの記事は発売日の株価変動率と対応付けた．一方,「日本経済新聞」の夕刊は夕方に発売されるため，当日の株式取引は既に終わっている．そこで夕刊の記事は当日の株価変動とは関連がないと考え，翌日の株価変動率と対応付けた．記事と対応付け出来た銘柄の数を表 3 に示す．

5.2 実験結果

5.2.1 単語スコアの評価

3 年分の記事データを用いて，抽出した単語 (品詞が名詞，動詞，形容詞，副詞や未知語である単語) の数は 150,153 だった．これらの単語の単語スコアを計算した．単語スコア $b_j \geq 0$ になった単語の数は $68,025(45.4\%)$，$b_j < 0$ になった単語の数は $82,128(54.6\%)$ だった．単語スコアが負になった単語のほうが多い結果となったが，これは前述の通り，記事の評価値が負である記事の割合が高いためである．

ここで，単語スコアを用いて記事を評価可能であることを示すため，簡単な検証を行った．まず，株価変動に影響を与えそうな単語例を実験の協力者に 60 個挙げてもらった．(「株価を上昇させる影響を与えそう」;「株価を下降させる影響を与えそう」;「どちらでもない」という 3 種類の単語を 20 個ずつ) この単語例のリストを表 4 に示す．以降，この 3 種類の単語をそれぞれ，上昇語，下降語，及び，中立語と呼ぶ．

次にこれらの単語例に対して，単語スコアが正と負になった単語の数をそれぞれ求めた．その結果を表 4 に示す．上昇語の

表 3: 記事と対応付けできた銘柄の数

年	2003 年	2004 年	2005 年	2006 年
銘柄数	1,853	1,975	1,984	2,060

表 4: 単語例のリスト

単語 id	上昇語	下降語	中立語
1	ヒット	下方	昨日
2	完了	破棄	先週
3	黒字	台風	明日
4	新設	下がり	今週
5	上がり	下り坂	人事
6	積極	被害	ＣＥＯ
7	好況	地震	経済
8	利益	不振	金利
9	設立	少子化	労働
10	技術	破産	収支
11	上場	危機	国際
12	好調	赤字	世界
13	合併	災害	発表
14	上昇	延期	来週
15	合意	不安	個人
16	堅調	伸び悩む	前期
17	好転	リスク	株価
18	旺盛	軟調	中国
19	回復	テロ	気配
20	上方	出遅れ	ＩＴ

場合は 75% の単語のスコアが正になり，下降語の場合は 90% の単語のスコアが負になった．また，中立語の場合は単語スコアが正になったのは 40%，負になったのは 60% だった．全体的に見ると，60 個の単語の中で単語のスコアが負になったものは全体の約 58.3% であった．これは前述の通り，単語スコアが負になった単語が，全体の 54.6% を占めたためと考えられる．

以上から，単語例の中の上昇語及び下降語は，それぞれ，株価の上昇及び下降と関連する記事に出現し易いと言える．

更に，単語例の単語スコアの大きさを比較するため，3 種類の単語における平均値をそれぞれ計算した．その結果，上昇語は 0.55×10^{-3}，中立語は -0.33×10^{-3}，下降語は -1.78×10^{-3} となった．単語スコアの分布を図 8 に示す．見やすくするため，3 種類の単語における単語スコアをそれぞれ昇順にソートした．単語スコアの分布は，上昇語のスコアが平均的に高く，下降語のスコアが平均的に低いことを示している．このことから，単語例の中の単語と株価変動との間に関連が存在することが言える．

以上から，単語スコアを用いて記事を評価可能であることを検証できた．

5.2.2 記事スコアの評価

単語スコアを用いて，1 年分の 73,333 記事に対して記事スコアを求めた．記事スコア $B_q > 0$ になったのは 34,651 記事，

表 5: 単語サンプルにおける単語スコアの正負

単語の種類	単語スコア b_j		
	$b_j \geq 0$	$b_j < 0$	合計
上昇語	15	5	20
下降語	2	18	20
中立語	8	12	20
合計	25	35	60

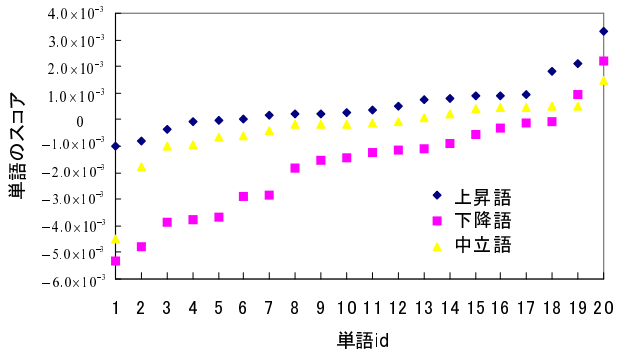


図 8: 単語リストにおける単語スコアの分布

$B_q < 0$ になったのは 38,682 記事だった．記事スコアが負になった記事の割合が高いが，これは前述の通り，単語スコアが負となった単語が全体の約 54.6%を占めているためと考えられる．

ここで，73,333 記事を記事スコアに対して昇順でソートし，ソートしたデータを 10 等分した．そして 10 等分したデータを記事スコアの低い順で 1 から 10 までのグループとした．

すなわち，1 グループは記事スコアが最も低い 10%(7,333 記事) のデータで，10 グループは記事スコアが最も高い 10%(7,333 記事) のデータである．記事スコアと記事の評価値の間の相関関係を示すために，以下の 2 つの評価を行った．

1. 記事スコアの上昇に伴う，記事の評価値の正負の変化
2. 記事スコアの上昇に伴う，記事の評価値の大きさの変化

評価 1 では，1 グループから 10 グループにおける記事の評価値の正負を求めた．各グループにおいて記事の評価値が正となった記事数と，負となった記事数を表 6 にまとめた．結果，記事スコアの上昇に従い，記事の評価値が正となる記事数が増加傾向にあることが認められた．

評価 2 では，1 グループから 10 グループまでの記事における記事の評価値の平均値について求めた．グループ番号に対して記事の評価値の平均値をプロットした結果を図 8 に示した．プロットから，記事スコアの上昇に従って記事の評価値の平均値も増加傾向にあることがわかる．スピアマンの順位相関係数は 0.88 であり，強い相関関係が認められた．

以上 2 点から，記事スコアと記事の評価値の間の相関関係が示された．よって，記事内容と特定銘柄の株価変動の間には関連性が存在していると言える．従って，特定銘柄の株価変動と関連する新聞記事を抽出できる可能性が明らかとなった．

表 6: 記事スコアと記事の評価値の関係の比較

記事スコアの	記事の評価値 t_i	
グループ番号	$t_i \geq 0$	$t_i < 0$
1	3,331(45.42%)	4,002(54.58%)
2	3,423(46.68%)	3,910(53.32%)
3	3,418(46.61%)	3,915(53.39%)
4	3,413(46.54%)	3,920(53.46%)
5	3,503(47.77%)	3,830(52.23%)
6	3,491(47.61%)	3,862(52.39%)
7	3,516(47.95%)	3,817(52.05%)
8	3,578(48.79%)	3,755(51.21%)
9	3,562(48.57%)	3,771(51.43%)
10	3,420(46.64%)	3,913(53.36%)

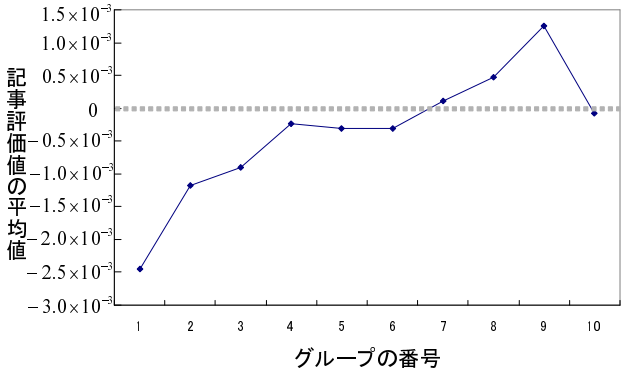


図 9: 記事スコアと記事の評価値の比較

6 おわりに

本稿では，新聞記事と特定銘柄の株価変動との関連性を評価した．実験には，2003 年から 2006 年までの 4 年分の記事データと株価データを用いた．新聞記事データは日本経済新聞社の記事を用い，株価データは東京証券取引所上場の全銘柄の株価データを用いた．はじめに 2003 年から 2005 年の 3 年分の記事に含まれる単語と特定銘柄の株価変動の関連度を求め，単語スコアとした．次に単語スコアを用いて 2006 年 1 年分のデータを記事スコアとして評価した．最後に，記事スコアと実際に起こった株価変動との相関関係を求めた．結果，記事内容と特定銘柄の株価変動の間に関連性が認められた．以上から，特定銘柄の株価変動と関連する新聞記事を抽出出来る可能性が明らかとなった．

今回の評価手法では，記事に含まれる単語に注目し，株価変動との関連度を求めた．しかし，実際の投資判断へ応用するためには，単語間の相互関係を考慮するなど，評価精度に改善の余地がある．今後さらに厳密な記事評価手法を開発し，新聞記事の評価精度を向上させる予定である．

参考文献

[1] 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.

[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，NL 142-19, 2001.

[3] 酒井 浩之，増山 繁：経済新聞記事内容の個々の企業におけるインパクトの判定，情報処理学会研究報告，NL 175-7, 2006.

[4] Lavernko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J: Mining of Concurrent Text and Time Series, In Proceedings of the KDD 2000 Conference Text Mining Workshop, 2001.

[5] 松本 裕治，北内 啓，山下 達雄，平野 善隆，松田 寛，高岡 一馬，浅原 正幸：形態素解析システム『ChaSen』，version2.2.9，使用説明書, 2002.